למידה חישובית

אלי דיין[1]

[1] <eliyahud@post.tau.ac.il>

תקציר

מסמך זה יביא את סיכומי השיעורים מהקורס למידה חישובית, שהועבר על ידי פרופ' ישי מנצור בסמסטר א' בשנה"ל תשע"ג.

תוכן עניינים

פרק 1

מה זה למידה חישובית?[1]

התחום צמח מתוך תחום הבינה המלאכותיתת ודומה לסטטיסטיקה, מבחינת השאלות ששואלים.

1.1 סוגי הבעיות

1. סיווג - Classification.

   נניח שלכל קלט יש סיווג נכון (ואולי יחיד). נרצה לדעת מהו. לדוגמה:

   (א) spam - סינון דואר זבל. כל הודעה יכולה להיות הודעה אמיתית או הודעת דואר זבל. נרצה לסווג את ההודעה בהתאם.

   (ב) סרטן (או מציאת מחלות באופן כללי).

   (ג) כרטיס אשראי - האם טרנסאקציה היא לגיטימית? לפסול אותה או לאשר אותה?

   (ד) דוגמה למקרה בו לקלט מספר סיווגים שונים: נושא המאמר (למשל פוליטיקה, ספורט ועוד).

2. בעיות בקרה.

   צריך לקבל החלטה שמשפיעה על מה שיהיה הקלט בפעם הבאה. דוגמאות:

   (א) לשחק משחקים (למשל ארבע בשורה, או שח־מט).

   (ב) רובוטים - למשל שליטה על מסוק.

3. מערכות המלצה - Collaborative Filter (כמו המלצות לחברים ב־Facebook).

בקורס נתמקד בבעיות סיווג.

1.2 סוגים של ML[2]

1.2.1 Supervised vs. Unsupervised

Supervised מקבלים דוגמאות עם סיווג $\langle a_1, a_2, \ldots, a_n \rangle, y$.

Unsupervised אין סיווג. המטרה היא לחלק את הנתונים לקבוצות בעלות דמיון.

בקורס נתמקד בבעיות Supervised.

[1] שיעור שהתקיים בתאריך 21.10.2012.

[2] קיצור של Machine Learning.

1.2.2 Active vs. Passive

Passive יש מאגר מידע שמשתמשים בו בצורה פאסיבית. למשל: data base -> השערה -> חיזוי

Active האלגוריתם בוחר את הקלטים שיסווגו.

בקורס נתמקד באלגוריתמים שהם Passive.

1.2.3 Teacher

Teacher למידה על ידי מומחה.

בקורס לא נתעסק בזה.

1.2.4 Batch vs. Online

Batch כל המידע נתון מראש.

Online המידע אינו נתון מראש. האלגוריתם רואה דוגמה, מנחש סיווג, מקבל את הסיווג הנכון, וממשיך לדוגמה הבאה.

בקורס נדבר על שתי הגישות.

1.3 בניית מודל ML

בכל מודל, נרצה להגדיר את המאפיינים הבאים:

1. איך הדוגמאות מיוצרות? למעשה, אנו רוצים כאן הנחה לגבי הסביבה.
2. להשוות סיווגים של אלגוריתמים שונים. הרבה שגיאות קטנות לעומת מעט שגיאות גדולות.
3. מיפוי מדוגמה לסיווג (השערה).

בהינתן שתי השערות, $+$ ו־$-$, וההתפלגויות D_+ ו־D_- של $+$ ו־$-$ בהתאמה, נסמן:

$$\lambda = \Pr[+] = 1 - \Pr[-]$$

כמו כן:

$$D(x) = \lambda \cdot D_+(x) + (1-\lambda) \cdot D_-(x)$$

בהינתן נקודה (x,y), היינו רוצים לענות על השאלה: מה ההסתברות שההשערה $+$ נכונה בהינתן (x,y)? כלומר, היינו רוצים למצוא את $\Pr[+ \mid (x,y)]$. נפתח את הנוסחה:

$$\begin{aligned} p = \Pr[+ \mid (x,y)] &= \frac{\Pr[(x,y) \mid +] \cdot \Pr[+]}{\Pr[(x,y)]} \\ &= \frac{D_+(x,y) \cdot \lambda}{D(x,y)} \end{aligned}$$

כאן, אנחנו צריכים לשאול את עצמנו מה המטרה של הסיווג. למשל, אם אנחנו מאתרים סרטן על פי הנתונים של החולה, אז יש חשיבות רבה לסיווג שניתן. אם נחליט שהאיש חולה, נפנה אותו לעוד בדיקות. לעומת זאת, אם נחליט שהאיש בריא, נשחרר אותו. לכן, היינו מעדיפים שבמקרים של ספק נשלח את האיש לבדיקות נוספות.

Loss Model 1.3.1

0-1Loss 1.3.1.1

במקרה של סיווג דו־ערכי:

$$\ell(0,1) = 1 = \ell(1,0)$$

$$\ell(0,0) = 0 = \ell(1,1)$$

במקרה הכללי:

$$\ell(y_1, y_2) = \begin{cases} 0 & y_1 = y_2 \\ 1 & y_1 \neq y_2 \end{cases}$$

כלומר, במידה וצדקנו, ההפסד (Loss) שלנו הוא 0. בכל מקרה אחר, 1.
בדוגמה לעיל (עם הסרטן):

$$\begin{aligned} \Pr[+ \mid (x,y)] &\geq \Pr[- \mid (x,y)] \\ p &\geq 1-p \\ p &\geq 1/2 \end{aligned}$$

Quadratic Loss 1.3.1.2

המודל אמור להביא קירוב טוב ביותר לכל המקרים, מבחינת מרחק.
אם ההסתברות הנכונה היא p, ההפסד עבור ההסתברות q הוא:

$$\begin{aligned} \ell(q) &= p(1-q)^2 + (1-p)q^2 \\ \frac{d}{dq}\ell &= -2p(1-q) + 2(1-p)q = 0 \\ -p + pq + q - pq &= 0 \\ p &= q \end{aligned}$$

כלומר, התשובה הטובה ביותר היא p (כמצופה).

Logarithmic Loss 1.3.1.3

אם הסיווג הוא $+$, אז השגיאה תהיה $-\log q$. אם הסיווג הוא $-$, אז השגיאה תהיה $-\log(1-q)$.
אם ההסתברות הנכונה היא p, ההפסד עבור ההסתברות q הוא:

$$\begin{aligned} \ell(q) &= -p\log q - (1-p)\log(1-q) \\ \frac{d}{dq}\ell &= -\frac{p}{q} + \frac{1-p}{1-q} = 0 \\ p(1-q) - q(1-p) &= 0 \\ p &= q \end{aligned}$$

כלומר, התשובה הטובה ביותר היא p (כמצופה).

הערה בכל השגיאות קיבלנו כי התשובה האופטימאלית היא p. אז למה יש סוגים שונים של שגיאות? מכיוון שהשגיאות עוסקות במקרים התת־אופטימליים, והם אילו שמעניינים אותנו.

1.3.2 הנחות על השערות

הנחה מפורשת:

$$f(x) = \sum_{i=1}^{d} \alpha_i x_i$$

או באופן מדוייק יותר, $f(x) + \text{Noise}$.
מחלקות השערות:

$$H = \left\{ \sum_{i=1}^{d} \alpha_i x_i \geq \theta \right\}$$

כאשר $\sum \alpha_i = 1$, $\alpha_i \geq 0$ ו־$\theta \in [0,1]$.

1.3.3 שיטות הסקה

1.3.3.1 הסקה בייסיאנית (Bayes)

ניעזר בהתפלגות prior (המפלה את ה"אמונות" שלנו).
למשל, אם הקלט הוא מדגם $S = \{\langle x_i, b_i \rangle\}$, $b_i \in \{0,1\}$, $x_i \in \mathbb{R}^d$, ואנחנו מחפשים את $f(x)$. אזי:

$$\Pr[f(x) = 1 \mid S, x] = \sum_{h \in H} h(x) \cdot \Pr[f = h \mid S]$$

$$\Pr[f = h \mid S] = \frac{\Pr[S \mid f = h] \cdot \Pr[f = h]}{\Pr[S]}$$

אותנו $\Pr[S]$ לא מעניין, כי זה הנתון שלנו, ולכן אנחנו יכולים להניח כי $\Pr[S] = 1$. לכן:

$$\Pr[f = h \mid S] = \Pr[S \mid f = h] \cdot \Pr[f = h]$$

כעת, ישנן שתי שיטות לבחירת h:

1. Maximum Likelihood:

$$h_{\text{ML}} = \arg\max_{h \in H} \Pr[S \mid f = h]$$

2. Maximum Aposteriori:

$$h_{\text{MAP}} = \arg\max_{h \in H} \Pr[S \mid f = h] \cdot \Pr[f = h]$$

1.3.3.2 מודל PAC

ההתפלגות D לא ידועה. כמוה גם פונקציית המטרה f אינה ידועה. H היא מחלקת ההשערות.
המטרה שלנו היא למצא $h \in H$ שתביא למינימום את $\varepsilon(h) = \Pr_D[h(x) \neq f(x)]$.

בעיה אין לנו את D, אלא רק מדגם S מתוך D.
נגדיר:

$$\hat{\varepsilon}(h) = \frac{1}{|S|} \cdot \sum_{x \in S} I[h(x) \neq f(x)]$$

כאשר $I[\delta]$ הוא האינדיקטור של δ, כלומר:

$$I[\delta] = \begin{cases} 1 & \delta \\ 0 & \neg\delta \end{cases}$$

כעת, אנחנו יכולים למצוא $h \in H$ שיביא למינימום את $\hat{\varepsilon}(h)$.

שאלה מה גודל $|\varepsilon(h) - \hat{\varepsilon}(h)|$? בהמשך נבטיח גם חסם הכללה:

$$\forall h \in H. |\varepsilon(h) - \hat{\varepsilon}(h)| \leq \delta$$

המתודולוגיה הכללית שלנו תהיה:

1. נבחר h שיביא למינימום את $\hat{\varepsilon}(h)$ (ERM).
2. נוכיח חסם הכללה.

דוגמאות למחלקות השערות

1. קו ישר על מישור (או $\mathbb{R}^{n-1}$ שמפריד בתוך $\mathbb{R}^n$).
2. עץ החלטה.

דוגמה[3] נניח כי $X = \{0,1\}^d$, $H = \{f_{x_1}, f_{x_2}, \ldots, f_{x_d}\}$ כאשר $f_{x_i}(\langle x_1, x_2, \ldots, x_n \rangle) = x_i$, ו־$D \sim \text{Unif}(H)$. נגריל פונקציה $f(x)$ כלשהי.
מה יקרה אם $m < 1/2 \log d$ דוגמאות? ההסתברות ש־x_1 מסווג m דוגמאות נכון: $\frac{1}{2^m}$.
ההסתברות שקיים x_i שמסווג את כל הדוגמאות נכון: $1 - \left(1 - \frac{1}{2^m}\right)^d$. אם נשאיף $m \to d$, נקבל $\frac{1}{e}$. בהסתברות קבועה, קיים x_i שמסווג את כל הדוגמאות נכון.

למעשה, היינו רוצים לחשב את $\min_h \hat{\varepsilon}(h) + \text{Complexity}(h)$, מכיוון שכאשר הסיבוכיות גדולה מדי, אנחנו נקבל Overfitting למדגם, והשגיאה בעולם שמעבר למדגם תהיה גדולה מדי.

1.3.3.3 מודלים מכוונים (Online)

בזמן t רואים קלט x_t, מוציאים סיווג y_t, רואים את הסיווג האמיתי $f(x_t)$, צוברים הפסד $\ell(y_t, f(x_t))$ וממשיכים.

[3] כל הקטע הזה לא ברור בכלל.

המטרה להביא למינימום את $\mathcal{L} = \sum_t \ell\left(y_t, f\left(x_t\right)\right)$.

פתרון נניח כי $h^* \in H$, שעבורה $\mathcal{L}$ קטן. כלומר:

$$\mathcal{L}\left(h^*\right) = \sum_t \ell\left(h^*\left(x_t\right), f\left(x_t\right)\right)$$

נרצה:

$$\mathcal{L}_{\min} \leq \mathcal{L}\left(h^*\right) + R$$

1.4 מבנה הקורס

- מודלים בסיסיים:
  - Bayes.
  - PAC.
  - Regression.
- אלגוריתמי Online:
  - Perception.
  - Regret.
  - Boosting.
- Generalization Bounds:
  - VC-dim.
  - Radamacher.
- אלגוריתמים:
  - Decission Trees.
  - SVM.
  - Fourier Transform.

פרק 2

[1]Bayesian Inference

2.1 מבוא

בשיעור זה נציג את מודל ההסקה הבייסיאני (Bayes Inference).
המודל משתמש בהתפלגות prior, שמשקפת את האמונות שלנו לגבי ההשערה הנכונה או הסיווג הנכון, ובסוף מקבלים התפלגות posterior, שמשקפת את מה שהתהליך למד.
יש 3 דרכים להסקה בייסיאנית:

- ML (או Maximum Likelihood).
- MAP (או Maximum A Posteriori).
- Bayes Posterior Rule.

2.2 כלל Bayes

$$\Pr[A \mid B] = \frac{\Pr[B \mid A] \cdot \Pr[A]}{\Pr[B]} \tag{2.2.1}$$

איך אנחנו נשתמש בזה? אנחנו נרצה לדעת את ההסתברות שההשערה h היא נכונה בהינתן המדגם data, כלומר את $\Pr[h \mid \text{data}]$.
לפי חוק Bayes:

$$\Pr[h \mid \text{data}] = \frac{\Pr[\text{data} \mid h] \cdot \Pr[h]}{\Pr[\text{data}]}$$

נשים לב ש־$\Pr[\text{data}]$ הינו בחזקת קבוע, מכיוון שלרוב אין לנו מידע על בחירת המדגמים, ולכן בהרבה מקומות ניתן להשמיט אותו.

2.3 דוגמה: זיהוי סרטן

נתונה ערכה לזיהוי של מחלת הסרטן. בהינתן מטופל, הערכה עשוייה להחזיר אחת משתי תשובות: 0 או 1, המייצגות מטופל בריא או חולה בהתאמה.
כמו כן, נתון כי:

[1] שיעור שהתקיים בתאריך 28.10.2012.

- אם המטופל חולה בסרטן, בסיכוי של 98% הערכה תחזיר עבורו 1.
- אם המטופל בריא, בסיכוי של 97% הערכה תחזיר עבורו 0.
- ההסתברות הכללית לסרטן בקרב האוכלוסייה היא 1%.

היינו רוצים למצוא את הסיכוי שמטופל חולה בסרטן אם הערכה טוענת שהוא חולה, כלומר את $\Pr[\text{Cancer} \mid 1]$.
נפרמל את הנתונים שלנו:

$$\Pr[1 \mid \text{Cancer}] = 98\% = 0.98$$

$$\Pr[0 \mid \neg\text{Cancer}] = 97\% = 0.97$$

$$\Pr[\text{Cancer}] = 1\% = 0.01$$

כמו כן, נסיק כי:

$$\Pr[\neg\text{Cancer}] = 1 - \Pr[\text{Cancer}] = 1 - 0.01 = 0.99$$

$$\Pr[1 \mid \neg\text{Cancer}] = 1 - \Pr[0 \mid \neg\text{Cancer}] = 1 - 0.97 = 0.03$$

$$\begin{aligned}\Pr[1] &= \Pr[1 \mid \text{Cancer}] \cdot \Pr[\text{Cancer}] + \Pr[1 \mid \neg\text{Cancer}] \cdot \Pr[\neg\text{Cancer}] = \\ &= 0.98 \cdot 0.01 + 0.03 \cdot 0.99 = 0.0395\end{aligned}$$

לכן, ולפי כלל Bayes:

$$\begin{aligned}\Pr[\text{Cancer} \mid 1] &= \frac{\Pr[1 \mid \text{Cancer}] \cdot \Pr[\text{Cancer}]}{\Pr[1]} = \\ &= \frac{0.98 \cdot 0.01}{0.0395} = 0.248 = 24.8\%\end{aligned}$$

באופן מפתיע, אם הערכה מזהה מטופל כחולה בסרטן, בסיכוי של פחות מ־25% הוא אכן חולה, כלומר 3 מתוך 4 מטופלים שהערכה חוזה שהם חולים, הם למעשה בריאים. אם היינו רוצים להיות מדוייקים יותר, יכולנו לבשר לכלל המטופלים שהם בריאים, ואז השגיאה הייתה 1% בלבד.
הסיבה לכך שהשגיאה גדולה מאוד היא ההסתברות לסרטן בקרב כלל האוכלוסייה, שהיא קטנה מאוד – 1% בלבד.

2.4 דוגמה: התפלגות נורמלית

2.4.1 התפלגות נורמלית

הגדרה נאמר שמשתנה מקרי (מ"מ) Z מתפלג נורמלית $N\left(\mu,\sigma^2\right)$, ונסמן $Z\sim N\left(\mu,\sigma^2\right)$ כאשר התוחלת שלו היא μ, והשונות שלו היא σ^2.
תכונות המ"מ $Z\sim N\left(\mu,\sigma^2\right)$:

$$\mathrm{E}\left[Z\right]=\mu$$

$$\begin{aligned}\mathrm{Var}\left[Z\right] &= \mathrm{E}\left[\left(Z-\mathrm{E}\left[Z\right]\right)^2\right]= \\ &= \mathrm{E}\left[Z^2\right]-\mathrm{E}^2\left[Z\right]= \\ &= \sigma^2\end{aligned}$$

$$\Pr\left[a\leq Z\leq b\right]=\int_a^b \frac{1}{\sqrt{2\pi\sigma^2}}\cdot e^{-\frac{1}{2}\cdot\left(\frac{\mu-x}{\sigma}\right)^2}dx$$

2.4.2 תיאור הבעיה

נניח כי $\mu,\sigma\sim N\left(0,1\right)$, כלומר μ ו־σ הוגרלו מבעוד מועד על ידי מ"מ נורמלי סטנדרטי. נתון מ"מ $Z\sim N\left(\mu,\sigma^2\right)$. ננסה בעזרת למידה חישובית למצוא את μ ו־σ.
נניח כי נתונות לנו n דגימות של המ"מ Z – $z_1,z_2,\ldots,z_n$.
לפי חוק Bayes:

$$\Pr\left[(\mu,\sigma)\mid z_1,z_2,\ldots,z_n\right]=\frac{\Pr\left[z_1,z_2,\ldots,z_n\mid(\mu,\sigma)\right]\cdot\Pr\left[(\mu,\sigma)\right]}{\Pr\left[z_1,z_2,\ldots,z_n\right]}$$

כידוע:

$$\Pr\left[z_1,z_2,\ldots,z_n\mid(\mu,\sigma)\right]=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\cdot e^{-\frac{1}{2}\cdot\left(\frac{\mu-z_i}{\sigma}\right)^2}$$

$$\Pr\left[(\mu,\sigma)\right]=\frac{1}{\sqrt{2\pi}}\cdot e^{-\mu^2/2}\cdot\frac{1}{\sqrt{2\pi}}\cdot e^{-\sigma^2/2}$$

כמו כן, נשים לב כי מבחינתנו $\Pr\left[z_1,z_2,\ldots,z_n\right]$ הוא קבוע שרק מנרמל את הנוסחה, ולכן נשמיט אותו בחישובים שלנו.

2.4.3 שיטת ML

אנחנו מחפשים את ההשערה h_{ML}, שמוגדרת לפי:

$$h_{\mathrm{ML}}=\arg\max_{h\in H}\Pr\left[\mathrm{data}\mid h\right]$$

במקרה שלנו, אנחנו מחפשים את μ ו־σ שיביאו למקסימום את $L(\mu,\sigma)$, שמוגדר על ידי:

$$\begin{aligned} L(\mu,\sigma) &= \Pr[z_1, z_2, \ldots, z_n \mid (\mu,\sigma)] = \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\cdot\left(\frac{\mu - z_i}{\sigma}\right)^2} \end{aligned}$$

נוציא log:

$$\log L(\mu,\sigma) = \sum_{i=1}^{n} -\frac{1}{2} \cdot \left(\frac{\mu - z_i}{\sigma}\right)^2 - \frac{n}{2}\log 2\pi - n\log\sigma$$

נגזור לפי μ:

$$\begin{aligned} \frac{\partial}{\partial\mu}\log L(\mu,\sigma) = \sum_{i=1}^{n} \frac{z_i - \mu}{\sigma} \cdot \frac{1}{\sigma} &= 0 \\ \sum_{i=1}^{n} z_i &= n\cdot\mu \\ \mu &= \frac{1}{n}\cdot\sum_{i=1}^{n} z_i \end{aligned}$$

נגזור לפי σ:

$$\begin{aligned} \frac{\partial}{\partial\sigma}\log L(\mu,\sigma) = \sum_{i=1}^{n} \frac{1}{\sigma} \cdot \left(\frac{z_i - \mu}{\sigma}\right)^2 - \frac{n}{\sigma} &= 0 \\ \sum_{i=1}^{n} (z_i - \mu)^2 &= n\cdot\sigma^2 \\ \sigma^2 &= \frac{1}{m}\cdot\sum_{i=1}^{n} (z_i - \mu)^2 \end{aligned}$$

2.4.4 שיטת MAP

אנחנו מחפשים את ההשערה h_{MAP}, שמוגרת לפי:

$$\begin{aligned} h_{\mathrm{MAP}} &= \arg\max_{h\in H} \Pr[h \mid D] = \\ &= \arg\max_{h\in H} \frac{\Pr[D \mid h]\cdot\Pr[h]}{\Pr[D]} \end{aligned}$$

הערה מעתה, $D = \mathrm{data}$.

במקרה שלנו, אנחנו מחפשים את μ ו־σ שיביאו למקסימום את $L_{\mathrm{MAP}}(\mu,\sigma)$, שמוגדר על ידי:

$$L_{\mathrm{MAP}}(\mu,\sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\cdot\left(\frac{z_i - \mu}{\sigma}\right)^2} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\mu^2/2} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\sigma^2/2}$$

נוציא log:

$$\log L_{\text{MAP}}(\mu,\sigma) = \sum_{i=1}^{n} -\frac{1}{2}\cdot\left(\frac{z_i-\mu}{\sigma}\right)^2 - \frac{n}{2}\log 2\pi - n\log\sigma - \frac{\mu^2}{2} - \frac{\sigma^2}{2} - \log 2\pi$$

נגזור לפי μ ו־σ:

$$\begin{aligned}\frac{\partial}{\partial\mu}\log L_{\text{MAP}}(\mu,\sigma) &= \sum_{i=1}^{n}\frac{z_i-\mu}{\sigma^2} - \mu = 0\\ \frac{\partial}{\partial\sigma}\log L_{\text{MAP}}(\mu,\sigma) &= \sum_{i=1}^{n}\frac{(z_i-\mu)^2}{\sigma^3} - \frac{n}{\sigma} - \sigma = 0\end{aligned}$$

קיבלנו מערכת משוואות על μ ו־σ. ננסה לפתור:

$$\begin{aligned}\frac{1}{n}\cdot\sum_{i=1}^{n} z_i &= \mu\cdot\left(1+\frac{\sigma^2}{n}\right)\\ \frac{1}{n}\cdot\sum_{i=1}^{n}(z_i-\mu)^2 &= \sigma^2\cdot\left(1+\frac{\sigma^2}{n}\right)\end{aligned}$$

2.4.5 שיטת Posterior Bayes

נניח[2] ש־$\mu\sim N(\mu,1)$ ו־$Z\sim N(\mu,1)$ (הנחנו כי $\sigma=1$). אזי:

$$\begin{aligned}\Pr[\mu] &= \frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{1}{2}\cdot(\mu-\eta)^2}\\ \Pr[z\mid\mu] &= \frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{1}{2}\cdot(z-\mu)^2}\end{aligned}$$

ולפי חוק Bayes:

$$\begin{aligned}\Pr[\mu\mid z] &= \frac{\Pr[z\mid\mu]\cdot\Pr[\mu]}{\Pr[z]} =\\ &= \alpha\cdot e^{-\frac{1}{2}\cdot\left(\mu^2-2\mu\eta+\eta^2\right)-\frac{1}{2}\cdot\left(z^{2-2\mu z+\mu^2}\right)} =\\ &= \alpha\cdot\exp\left\{-\frac{1}{2}\cdot\left(2\mu^2-2\mu(\eta+z)+\eta^2+z^2\right)\right\} =\\ &= \alpha\cdot\exp\left\{-\left(\mu-\frac{\eta+z}{2}\right)^2+\underbrace{\left(\frac{\eta+z^2}{2}\right)^2-\eta^2-z^2}_{\text{Normalization}}\right\}\end{aligned}$$

כאשר α הוא קבוע כלשהו (ניתן לחילוץ מתוך החישוב לעיל).
נגדיר כעת:

$$\begin{aligned}\hat{\mu} &= \frac{\eta+z}{2}\\ \hat{\sigma}^2 &= \frac{1}{2}\end{aligned}$$

[2] או למעשה האמונה המוקדמת שלנו.

כעת, נוכל להניח כי $Z \sim N\left(\hat{\mu}, \hat{\sigma}^2\right)$, ונוכל להמשיך את התהליך.

לחלופין, נניח שהגרלנו $\mu \sim N\left(\eta, s^2\right)$ וש־$Z \sim N\left(\mu, \sigma^2\right)$. נוכל להפעיל את התהליך על הנתונים $z_1, z_2, \ldots, z_n$ שהוגרלו לפי Z. נסמן: $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$.

כלל ההסקה שלנו יהיה:

$$\begin{aligned}\hat{\mu} &= \frac{\frac{1}{s^2}\cdot\eta + \frac{n}{\sigma^2}\cdot\bar{z}}{\frac{1}{s^2}+\frac{n}{\sigma^2}} \\ \hat{\sigma}^2 &= \left(\frac{1}{s^2}+\frac{n}{\sigma^2}\right)\end{aligned}$$

נשים לב שעבור ההשערה : $s = \sigma$:

$$\begin{aligned}\hat{\mu} &= \frac{\eta + n\bar{z}}{1+n} \\ \hat{\sigma}^2 &= \frac{\sigma^2}{1+n}\end{aligned}$$

מבחינה איכותית, $\hat{\mu}$ מתקרב יותר ויותר לממוצע, ו־$\hat{\sigma}^2$ הולך ויורד שזה דבר טבעי ככל שיש יותר נתונים.

הערה ראינו כאן שתי אפשרויות: איך עושים את התהליך עבור נקודה אחת, ואיך עושים אותו עבור n נקודות.

אולם, אם נעשה n פעמים את התהליך מנקודה לנקודה, נקבל בדיוק את אותה התוצאה שהיינו מקבלים מביצוע התהליך עבור n נקודות ישירות.

Learning a Concept Class 2.5

נניח כי H מחלקת השערות, ו־$f \in H$ היא פונקציית המטרה. נניח כי $\langle x, f(x)\rangle$ היא דוגמה.

נשים לב ש־$\Pr[h(x)=1] \in \{0,1\}$ עבור $h \in H$ מסויימת, כי h ידועה לנו מראש, ולכן אנחנו יכולים לבדוק אם $h(x)=1$ או $h(x)=0$.

נסמן: $S = \{\langle x_i, b_i\rangle\}$, כאשר $b_i = f(x_i) \in \{0,1\}$. אזי:

$$\Pr[S \mid h] = \begin{cases} 0 & \exists i.b_i \neq h(x_i) \\ \Pr[x_1, x_2, \ldots, x_n] & \forall i.b_i = h(x_i) \end{cases}$$

הגדרה נאמר ש־$h \in H$ עקבית אם היא מסווגת את כל הנקודות ב־S בצורה נכונה, כלומר $\forall i.b_i = h(x_i)$.

נגדיר את $H' \subseteq H$ לפי $H' = \{h \in H \mid h \text{ is consistent}\}$.

נסתכל על כל כלל הסקה שיש לנו:

ML יחזיר השערה עקבית כלשהי.

MAP יחזיר השערה עקבית עם המשקל הפוסטריורי הגבוה ביותר, כלומר את ההשערה העקבית שהכי האמנו בה.

Bayes ה־support יהיה H' ומשקול מחדש[3]. המשקול המחודש יבוצע לפי הנוסחה:

$$B(y) = \sum_{h \in H'} \frac{h(y) \cdot \Pr[h]}{\Pr[H']}$$

2.6 דוגמה: Biased Coins

נניח כי מטילים מטבע m פעמים, ורואים k הצלחות. רוצים להעריך את ההסתברות p של המטבע.

נחשב את p בעזרת ML:

$$\Pr[(k,m) \mid p] = \binom{m}{k} \cdot p^k \cdot (1-p)^{m-k}$$

לאחר חישוב מקבלים כי $p = k/m$. איך עושים את זה? מוציאים log ומחלקים ב־m, ואז אפשר לגזור. לא נראה כאן את החישוב.

2.6.1 חוק Laplace

נניח התפלגות prior אחידה (Uniform). כלומר, התפלגות לכל המטבעות האפריות היא אחידה:

$$\Pr[p \le \theta] = \int_0^\theta dp = \theta$$

נחשב את ההסתברות ל־k הצלחות מ־m הטלות:

$$\begin{aligned}
\Pr[(k,m)] &= \int_0^1 \Pr[k \mid p] \cdot \Pr[p]\, dp = \\
&= \int_0^1 \binom{m}{k} x^k (1-x)^{m-k}\, dx = \\
&= \cancel{\left[\binom{m}{k} \cdot \frac{x^{k+1}}{k+1} \cdot (1-x)^{m-k}\right]_0^1} \\
&\quad + \int_0^1 \binom{m}{k} \cdot \frac{x^{k+1}}{k+1} \cdot (1-x)^{m-k-1} \cdot (m-k)\, dx = \\
&= \int_0^1 \binom{m}{k+1} \cdot x^{k+1} \cdot (1-x)^{m-k+1}\, dx = \\
&= \int_0^1 \Pr[k+1 \mid p] \cdot \Pr[p]\, dp = \Pr[(k+1,m)]
\end{aligned}$$

נשים לב שהשתמשנו כאן במעברים בזהות הבאה:

$$\binom{n}{k} \cdot \frac{n-k}{k+1} = \binom{n}{k+1}$$

[3] למשל, אם $H = \{h_1, h_2, h_3, h_4\}$ וכן: $\forall 1 \le i \le 4.\ \Pr[h_i] = i/10$. נניח כי $H' = \{h_2, h_3\}$. נצטרך למשקל מחדש את h_2 ו־h_3, כך ש־$\Pr[h_2] + \Pr[h_3] = 1$.

נשים לב:

$$\begin{aligned}\Pr\left[(k,m)\right] &= \int_0^1 p^k \cdot (1-p)^{m-k}\, dp = \\ &= \frac{1}{\binom{m}{k}} \cdot \frac{1}{m+1}\end{aligned}$$

נחשב:

$$\begin{aligned}\mathrm{E}\left[p \mid (k,m)\right] &= \int_0^1 p \cdot \frac{\Pr\left[(k,n) \mid p\right] \cdot \Pr\left[p\right]}{\Pr\left[(k,m)\right]} dp = \\ &= \int_0^1 p \cdot \frac{p^k \cdot (1-p)^{m-k}}{\frac{1}{m+1} \cdot \frac{1}{\binom{m}{k}}} dp = \\ &= (m+1) \cdot \binom{m}{k} \cdot \int_0^1 p^{k+1} \cdot (1-p)^{m-k}\, dp = \\ &= (m+1) \cdot \binom{m}{k} \cdot \frac{1}{m+2} \cdot \frac{1}{\binom{m+1}{k+1}} = \\ &= \frac{\cancel{m+1}}{m+2} \cdot \frac{k+1}{\cancel{m+1}} = \\ &= \frac{k+1}{m+2}\end{aligned}$$

למעשה, התיקון של Laplace מוסיף לנו עוד שתי דגימות, כאשר רק באחת מהן יש הצלחה.

למה זה יותר טוב? אם יש הרבה הטלות, זה עדיין קרוב למה שהיינו חושבים. אולם אם יש מעט הטלות, אז עדיין יש מקום למקרי הקיצון.

2.6.2 פונקציות Loss

למעשה, לא נוכל אף פעם לחזות מה יקרה בהטלת המטבע הבאה. לכן, נכנס לכאן העניין של ההפסד (Loss).

כאשר מדברים על פונקציות Loss, יש לקחת בחשבון שני מרכיבים להפסד:

1. Bayes Risk – זה ההפסד הבלתי־נמנע, גם אם אנחנו מכירים את כל הפרטים במערכת. למשל, בבעית המטבע, לא נוכל אף פעם לחזות במדויק מה תהיה תוצאת ההטלה הבאה, אלא רק בהסתברות מסויימת.
2. Regret – ההפסד בגלל הערכה לא נכונה של המערכת.

תזכורת פונקצית log loss מוגדרת באופן הבא:

$$\ell_{\log}\left(x\right) = \begin{cases}\log \frac{1}{p} & f\left(x\right) = 1 \\ \log \frac{1}{1-p} & f\left(x\right) = 0\end{cases}$$

עבור θ-bias, ה־log loss הצפוי הוא:

$$\ell_{\log} = \theta \cdot \log \frac{1}{p} + (1-\theta) \cdot \log \frac{1}{1-p}$$

כמובן שהערך המינימלי מתקבל עבור $p = \theta$. ואם באמת $p = \theta$, אז נסמן:

$$H(\theta) = \theta \cdot \log \frac{1}{\theta} + (1-\theta) \cdot \log \frac{1}{1-\theta}$$

לערך זה נקרא האנטרופיה של θ. זהו ההפסד שנובע מהמערכת עצמה, ולא מכך שאיננו מכירים את המערכת, ולכן זהו ה־Bayes Risk.
נבדוק כמה אנחנו רחוקים מה־Bayes Risk (כלומר מהו ההפסד הנוסף, ה־Regret) כשאנחנו משתמשים בחוק Laplace: נניח כי יש לנו T דגימות בזמן. אזי:

$$\begin{aligned}
&\mathrm{E}[\log\ \mathrm{loss}] = \\
= &\int_0^1 \sum_{m=1}^{T} \sum_{k=0}^{m} \left[\theta \cdot \log \frac{m+2}{k+1} + (1-\theta) \cdot \log \frac{m+2}{m-k+1} \right] \\
&\binom{m}{k} \cdot \theta^k \cdot (1-\theta)^{m-k}\, d\theta = \\
= &\sum_{m=1}^{T} \sum_{k=0}^{m} \binom{m}{k} \cdot \log \frac{m+2}{k+1} \cdot \int_0^1 \theta^{k+1} \cdot (1-\theta)^{m-k}\, d\theta + \\
&\sum_{m=1}^{T} \sum_{k=0}^{m} \binom{m}{k} \cdot \log \frac{m+2}{m-k+1} \cdot \int_0^1 \theta^{k} \cdot (1-\theta)^{m-k+1}\, d\theta = \\
= &\sum_{m=1}^{T} \sum_{k=0}^{m} \left[\frac{1}{m+1} \cdot \frac{k+1}{m+2} \cdot \log \frac{m+2}{k+1} + \frac{1}{m+1} \cdot \frac{m-k+1}{m+2} \cdot \log \frac{m+2}{m-k+1} \right] = \\
= &\sum_{m=1}^{T} \sum_{k=0}^{m} \frac{1}{m+1} \cdot H\left(\frac{k+1}{m+1} \right) \sim \\
\sim &\ T \cdot \int_0^1 H(\theta)\, d\theta + O\left(\underbrace{\sum_{m=1}^{T} \frac{1}{m}}_{\log T} \right)
\end{aligned}$$

ננסה לחסום את $T \cdot \int_0^1 H(\theta)\, d\theta$:

$$\sum_{i=1}^{m/2} \frac{1}{m} \cdot H\left(\frac{i-1}{m} \right) \le \int_0^1 H(\theta)\, d\theta \le \sum_{i=1}^{m/2} \frac{1}{m} \cdot H\left(\frac{i}{m} \right)$$

$$\begin{aligned}
\sum_{i=1}^{m/2} \frac{1}{m} \cdot H\left(\frac{i}{m} \right) - \sum_{i=1}^{m/2} \frac{1}{m} \cdot H\left(\frac{i-1}{m} \right) &= \frac{1}{m} \cdot \sum_{i=1}^{m/2} \left[H\left(\frac{i}{m} \right) - H\left(\frac{i-1}{m} \right) \right] = \\
&= \frac{1}{m} \cdot (1-0) = \frac{1}{m} \underset{m \to \infty}{\longrightarrow} 0
\end{aligned}$$

x_1	x_2	$\ldots$	x_n	C
0	1		1	$+1$
1	0		0	-1
$\vdots$	$\vdots$		$\vdots$	$\vdots$
0	1		0	$+1$

טבלה 2.1: דוגמה למדגם עבור סיווג בייסיאני למרחב בינארי

כלומר, ככל ש־m גדל, כך $T \cdot \int_0^1 H(\theta)\, d\theta$ הולך ומתקרב ל־$T \sum_{i=1}^{m/2} \frac{1}{m} \cdot H\left(\frac{i}{m}\right)$, שהוא ה־Bayes Risk.

לכן, הראנו שבאמצעות הפעלת חוק Laplace, קיבלנו את ה־loss האופטימלי (ה־Bayes Risk), עם תוספת לוגריתמית במספר הטלות המטבע (T).

Naïve Bayes 2.7

סיווג בייסיאני: מרחב בינארי 2.7.1

נתון המצב הבא: יש לנו שתי מחלקות $+1$ ו־-1, וכל דוגמה מתוארת על ידי n מאפיינים, כאשר x_i ($1 \leq i \leq n$) הוא משתנה בינארי, שערכו 0 או 1. דוגמה לקלט כזה ניתן למצוא בטבלה 2.1.

רוצים לבנות השערה $h\colon \{0,1\}^n \to \{+1,-1\}$. לפי כלל Bayes:

$$\Pr[C = +1 \mid x_1, x_2, \ldots, x_n] = \frac{\Pr[x_1, x_2, \ldots, x_n \mid C = +1] \cdot \Pr[C = +1]}{\Pr[x_1, x_2, \ldots, x_n]}$$

קל להעריך את $\Pr[C = +1]$ מהנתונים.

אבל איך נעריך את $\Pr[x_1, x_2, \ldots, x_n \mid C = +1]$? Naïve Bayes מבוסס על הנחת אי־תלות:

$$\Pr[x_1, x_2, \ldots, x_n \mid C = +1] = \prod_{i=1}^{n} \Pr[x_i \mid C = +1]$$

כל מאפיין x_i הוא ב"ת באחרים ברגע שאנחנו יודעים את הערך של C. לכן, לכל $1 \leq i \leq n$, יש לנו שני מאפיינים:

$$\begin{aligned} \theta_{i,+1} &= \Pr[x_i = 1 \mid C = +1] \\ \theta_{i,-1} &= \Pr[x_i = 1 \mid C = -1] \end{aligned}$$

קיבלנו כאן $2n$ פרמטרים בלתי־תלויים (ב"ת).

תזכורת

אי שוויון Markov: אם $X \geq 0$ מ"מ, אזי $\Pr[X \geq \lambda] \leq \frac{\mathrm{E}[X]}{\lambda}$.

אי שוויון Cheviechev: אם $X \geq 0$ מ"מ, אזי $\Pr[X \geq \lambda] \leq \frac{\mathrm{E}[X^2]}{\lambda^2}$, $\Pr[|X - \mathrm{E}[X]| \geq \lambda] \leq \frac{\mathrm{Var}[X]}{\lambda^2}$.

אי שוויון Chernoff: אם $X_1, X_2, \ldots, X_m$ מ"מ ב"ת כך ש־$\forall i. \mathrm{E}[X_i] = p$, אזי $\Pr\left[\left|\frac{1}{m} \cdot \sum_{i=1}^{m} X_i - p\right| \geq \lambda\right] \leq e^{-2\lambda^2 m}$.

2.7.2 פענוח של Naïve Bayes

לפי Bayes ו־MAP, אנחנו צריכים להעריך את $\Pr[C=+1 \mid x_1, x_2, \ldots, x_n]$ לעומת $\Pr[C=-1 \mid x_1, x_2, \ldots, x_n]$.
קל לעשות זאת בעזרת log וחלוקה (והשוואה ל־0):

$$\begin{aligned}
&\log \frac{\Pr[C=+1 \mid x_1, x_2, \ldots, x_n]}{\Pr[C=-1 \mid x_1, x_2, \ldots, x_n]} = \\
= \quad &\log \frac{\Pr[x_1, x_2, \ldots, x_n \mid C=+1] \cdot \Pr[C=+1]}{\Pr[x_1, x_2, \ldots, x_n \mid C=-1] \cdot \Pr[C=-1]} = \\
= \quad &\log \frac{\Pr[C=+1]}{\Pr[C=-1]} + \log \prod_{i=1}^{n} \frac{\Pr[x_i \mid C=+1]}{\Pr[x_i \mid C=-1]} = \\
= \quad &\log \frac{\Pr[C=+1]}{\Pr[C=-1]} + \sum_{i=1}^{n} \log \frac{\Pr[x_i \mid C=+1]}{\Pr[x_i \mid C=-1]}
\end{aligned}$$

לכן, הסקנו כי

$$\log \frac{\Pr[C=+1 \mid x_1, x_2, \ldots, x_n]}{\Pr[C=-1 \mid x_1, x_2, \ldots, x_n]} = \log \frac{\Pr[C=+1]}{\Pr[C=-1]} + \sum_{i=1}^{n} \log \frac{\Pr[x_i \mid C=+1]}{\Pr[x_i \mid C=-1]}$$

כלומר, כל x_i משפיע על החיזוי:

- אם $\Pr[x_i \mid C=+1] = \Pr[x_i \mid C=-1]$, אז ל־$x_i$ אין השפעה על החיזוי.
- אם $\Pr[x_i \mid C=-1] = 0$, אז x_i משפיע על שאר הקולות (דומה להטלת וטו).
- באופן דומה, אם $\Pr[x_i \mid C=+1] = 0$.

נסמן:

$$\begin{aligned}
w_i &= \log \frac{\Pr[x_i=1 \mid C=+1]}{\Pr[x_i=1 \mid C=-1]} - \log \frac{\Pr[x_i=0 \mid C=+1]}{\Pr[x_i=0 \mid C=-1]} \\
b &= \log \frac{\Pr[C=+1]}{\Pr[C=-1]} + \sum_{i=1}^{n} \log \frac{\Pr[x_i=0 \mid C=+1]}{\Pr[x_i=0 \mid C=-1]}
\end{aligned}$$

כלל ההחלטה שלנו יהיה $\operatorname{sign}\left(b + \sum_{i=1}^{n} x_i w_i\right)$.

2.7.3 התפלגות נורמלית

השלב הבא ב־Naïve Bayes אומר ש־$\Pr[x_i \mid C] \sim N(\mu_{i,C}, \sigma_i)$[4].
נעשה את אותו החישוב:

$$\begin{aligned}
&\log \frac{\Pr[C=+1 \mid x_1, x_2, \ldots, x_n]}{\Pr[C=-1 \mid x_1, x_2, \ldots, x_n]} = \\
= \quad &\log \frac{\Pr[C=+1]}{\Pr[C=-1]} + \sum_{i=1}^{n} \log \frac{\Pr[x_i \mid C=+1]}{\Pr[x_i \mid C=-1]}
\end{aligned}$$

[4] חשוב לשים לב ש־$\mu_{i,C}$ תלוי ב־C, ואילו σ_i אינו תלוי ב־C.

$$
\begin{aligned}
\log \frac{\Pr\left[x_i \mid C=+1\right]}{\Pr\left[x_i \mid C=-1\right]} &= \log \frac{e^{-\frac{1}{2}\cdot\left(\frac{\mu_{i,+1}-x_i}{\sigma_i}\right)^2}}{e^{-\frac{1}{2}\cdot\left(\frac{\mu_{i,-1}-x_i}{\sigma_i}\right)^2}} = \\
&= -\frac{1}{2}\cdot\left(\frac{\mu_{i,+1}-x_i}{\sigma_i}\right)^2 + \frac{1}{2}\cdot\left(\frac{\mu_{i,-1}-x_i}{\sigma_i}\right)^2 = \\
&= \frac{1}{2\cdot\sigma^2}\cdot\left(\mu_{i,+1}+\mu_{i,-1}-2x_i\right)\cdot\left(\mu_{i,-1}-\mu_{i,+1}\right) = \\
&= \frac{1}{2}\cdot\frac{\mu_{i,-1}-\mu_{i,+1}}{\sigma_i}\cdot\frac{\frac{\mu_{i,+1}+\mu_{i,-1}}{2}-x_i}{\sigma_i}
\end{aligned}
$$

פרק 3

מודל ה־PAC[1]

3.1 מודל ה־PAC

ראשי התיבות PAC פירושן *Probably Approximately Correct*.
המטרה של המודל היא למצוא השערה h, שבהסתברות גבוהה (Probably) היא מדוייקת (Approximately Correct).
מבנה השיעור:

1. דוגמה ללימוד PAC.
2. מודל פורמלי ומחלקת השערות סופית.
3. Occam Razor ודוגמאות.

3.2 דוגמה אינטואיטיבית

נניח כי בן־אדם טיפוסי מקיים את החיתוך של שתי התכונות הבאות:

$$\begin{aligned} 60 \leq \quad \text{Weight} \quad &\leq 90 \\ 1.60 \leq \quad \text{Height} \quad &\leq 1.90 \end{aligned}$$

אזי R היא מחלקת בני האדם. נשים לב לכך ש־R הוא מלבן על המישור. אנחנו נמצא מחלקה R' שהיא קירוב של R.
במודל PAC אין הנחה על התפלגות הדוגמאות. ההנחה היחידה שלנו היא שקיימת התפלגות, והיא נדגמת באופן i.i.d.
נתאר את הבעיה:

- קלט: אוסף S של דוגמאות מסווגות.
- פלט: R' (מלבן).
- מטרה: משקל קטן עבור $D\left(R \triangle R'\right)$ (חיסור סימטרי).

[1] שיעור שהתקיים בתאריך 04.11.2012.

3.2.1 מציאת השערה טובה

על השגיאה ניתן להסתכל באופן הבא:

$$R \triangle R' = \underbrace{(R \setminus R')}_{\text{False-Positive}} \cup \underbrace{(R' \setminus R)}_{\text{False-Negative}}$$

המטרה שלנו היא למצוא השערה R', כך שבהסתברות $1-\delta$:

$$\Pr\left[\text{error}\right] = D\left(R \triangle R'\right) \leq \varepsilon$$

כאשר R היא פונקציית המטרה.

3.2.2 אופן הלימוד

אם נסתכל על המישור, יש שתי בחירות אינטואיטיביות:

- $R_{\min}$ – המלבן החוסם הקטן ביותר.
- $R_{\max}$ – המלבן החוסם הגדול ביותר.

איך נחשב את $R_{\min}$ (לדוגמה) בהינתן קבוצת דגימות $S = \{\langle (x_1, y_1), b_1 \rangle, \cdots, \langle (x_m, y_m), b_m \rangle\}$? נסמן:

$$m_x = x_1 = M_x$$

$$m_y = y_1 = M_y$$

נעבור על $\langle (x_i, y_i), b_i \rangle$. אם $b_i = +$:

- אם $m_x > x_i$, אז נגדיר $m_x = x_i$.
- אם $M_x < x_i$, אז נגדיר $M_x = x_i$.
- אם $m_y > y_i$, אז נגדיר $m_y = y_i$.
- אם $M_y < y_i$, אז נגדיר $M_y = y_i$.

כעת, $R_{\min} = \{(x, y) \mid m_x \leq x \leq M_x \wedge m_y \leq y \leq M_y\}$.
נשים לב שזהו אלגוריתם יעיל (פולינומיאלי) ב־m, מספר הדגימות שלנו.

3.2.3 מספר הדגימות

ננתח את הבעיה באופן פורמלי. נתונים לנו:

- ε – דיוק.
- δ – ביטחון.

אנחנו מחפשים אלגוריתם A, כך שעבור מדגם בגודל $m(\varepsilon, \delta)$, A מחזיר מלבן R', כך שבהסתברות $1-\delta$:

$$\Pr\left[\text{error}\right] \leq \varepsilon$$

ננתח את $m(\varepsilon, \delta)$ עבור האלגוריתם A שמחזיר את $R_{\min}$.
נשים לב ש־$R_{\min} \subseteq R$, ולכן:

$$\Pr[\text{error}] = D(R \triangle R_{\min}) = D(R \setminus R_{\min})$$

לכן, נשים לב שהשגיאה האפשרית שלנו נמצאת כולה ב־R.
נחלק את השגיאה שלנו לארבעה מלבנים: T_1' (החלק שמעל $R_{\min}$), T_2' (החלק שמשמאל ל־$R_{\min}$), T_3' (החלק שמתחת ל־$R_{\min}$) ו־T_4' (החלק שמימין ל־$R_{\min}$).
אזי:

$$R \triangle R_{\min} = \bigcup_{i=1}^{4} T_i'$$

$$D(R \triangle R_{\min}) \leq \sum_{i=1}^{4} D(T_i')$$

לכן, אם נראה כי $\forall i. D(T_i') \leq \varepsilon/4$, אז סיימנו, כי:

$$D(R \triangle R_{\min}) \leq 4 \cdot \varepsilon/4 = \varepsilon$$

הבעיה הקונספטואלית שלנו במצב הזה היא שהגדרנו את T_i' רק אחרי שכבר ראינו את המדגם ובנינו את $R_{\min}$. החוכמה היא להגדיר מאורעות שאינם תלויים במדגם. רק אז נוכל לחשב מה המשקל שלהם.
נשים לב לכך שמכיוון שבהכרח $R_{\min} \subseteq R$, כל השגיאה תהיה תמיד ב־$R \setminus R_{\min}$. לכן, נחפש T_1, T_2, T_3, T_4 באופן הבא:

- T_1 הוא תת־מלבן של R הצמוד לדופן העליונה של R.
- T_2 הוא תת־מלבן של R הצמוד לדופן השמאלית של R.
- T_3 הוא תת־מלבן של R הצמוד לדופן התחתונה של R.
- T_4 הוא תת־מלבן של R הצמוד לדופן הימנית של R.
- לכל $1 \leq i \leq 4$ מתקיים $D(T_i) = \varepsilon/4$.

נשים לב לכך שהבחירה של $\{T_i\}_{i=1}^4$ אינה תלוייה במדגם. נרצה לחשב את ההסתברות ש־$T_i' \subseteq T_i$. כלומר, שבהסתברות $1 - \delta$, $\forall i. T_i' \subseteq T_i$.
נסתכל על T_1: מתי $T_1' \subseteq T_1$? אם קיימת ב־S דוגמה $\langle (x,y), + \rangle$ כך ש־$(x,y) \in T_1$, אזי $T_1' \subseteq T_1$.
לפי הבנייה:

$$\Pr[(x,y) \notin T_1] = 1 - \frac{\varepsilon}{4}$$

מכיוון שהדגימות של S הן i.i.d, נסיק כי:

$$\Pr[\forall (x_i, y_i) \in S. (x_i, y_i) \notin T_1] = \left(1 - \frac{\varepsilon}{4}\right)^m$$

את אותם החישובים ניתן לעשות גם עבור $\{T_i\}_{i=2}^4$. לכן:

$$\begin{aligned}\Pr[\text{error}] &= \Pr[\exists i.\forall (x,y) \in S.\,(x,y) \notin T_i] \\ &\leq \sum_{i=1}^{4} \Pr[\forall (x,y) \in S.\,(x,y) \notin T_i] \\ &\leq 4 \cdot \left(1 - \frac{\varepsilon}{4}\right)^m \leq 4e^{-\frac{\varepsilon}{4}m} < \delta\end{aligned}$$

המעבר האחרון באי־השוויון בוצע לפי הזהות $1 - x \leq e^{-x}$. לסיום, נסיק כי:

$$m > \frac{4}{\varepsilon} \cdot \ln \frac{4}{\delta}$$

3.3 הצגה פורמלית של מודל ה־PAC

3.3.1 הקדמה

- המטרה שלנו היא ללמוד השערה מתוך קבוצה ידועה מראש של השערות.
- הסביבה סטוכסטית.
- הדגימות נדגמות מהתפלגות i.i.d.
- ההתפלגות על הלמידה ($train$) זהה להתפלגות על הבדיקה ($test$).
- הפתרון צריך להיות יעיל: גודל המדגם ביחס לזמן החישוב צריך להיות פולינומיאלי ב־$\frac{1}{\varepsilon}$ ו־$\ln \frac{1}{\delta}$.

3.3.2 הגדרת מודל ה־PAC

יהי X מרחב הדוגמאות, ותהי D התפלגות מעל X. נאמר כי $C \subseteq \{c \mid c : X \to \{0,1\}\}$ היא מחלקת פונקציות המטרה.

תהי $c_t \in C$ פונקציית המטרה שלנו. תהי $H \subseteq C$ מחלקת ההשערות שלנו. תהי $h \in H$ השערה שנלמדה.

נגדיר את השגיאה של ההשערה h:

$$\text{error}(h) = \Pr_D[h(x) \neq c_t(x)] = D(h \triangle c_t)$$

כמו כן, נגדיר את $EX(c_t, D)$ להיות אורקל ($Oracle$), המחזיר דגימה $x \in X$ שנדגמה לפי D וסיווג $c_t(x)$: $\langle x, c_t(x) \rangle$.

הגדרה נאמר כי C נלמדת PAC על ידי H אם קיים אלגוריתם A כך שלכל פונקציית מטרה $c_t \in C$, לכל התפלגות D מעל המרחב X, ולכל פרמטרים $\varepsilon, \delta > 0$, אם נותנים ל־A גישה ל־$EX(c_t, D)$, אזי בהסתברות $1 - \delta$, A יחזיר השערה $h \in H$ כך ש:

- אם $c_t \in H$ ($realizable$), אזי:

$$\text{error}(h) \leq \varepsilon$$

- אם $c_t \notin H$ (*non-realizable*), אזי:

$$\text{error}(h) \leq \varepsilon + \min_{\bar{h} \in H} \text{error}(h)$$

נאמר כי C נלמדת ביעילות אם A רץ בזמן פולינומיאלי ב־$\frac{1}{\varepsilon}$, $\ln \frac{1}{\delta}$, וגם ב־n ו־ℓ, כאשר n הוא גודל הדוגמה (בביטים) ו־ℓ הוא גודל פונקציית המטרה c_t (גם כן בביטים).

3.4 מחלקות השערות סופיות

נדון במקרה בו H סופית.

3.4.1 המקרה $c_t \in H$

הגדרה נאמר כי השערה $h \in H$ עקבית אם $\forall x \in S. h(x) = c_t(x)$.
במקרה שלנו, מובטח לנו שקיימת h עקבית.
נניח כי קיימת $h \in H$ עקבית כך ש־$\text{error}(h) > \varepsilon$. אזי h נקראת מאורע רע.

$$\Pr[\forall i. h(x_i) = c_t(x_i) \mid \text{error}(h) > \varepsilon] \leq (1-\varepsilon)^m \leq e^{-\varepsilon m}$$

$$\begin{aligned} \Pr[A \text{ returns } h \wedge \text{error}(h) > \varepsilon] &\leq \\ \Pr[\exists h \in H. \text{error}(h) > \varepsilon \wedge \forall i. h(x_i) = c_t(x_i)] \leq |H| \cdot e^{-\varepsilon m} &\leq \delta \end{aligned}$$

נחלץ את m:

$$\begin{aligned} |H| \cdot e^{-\varepsilon m} &\leq \delta \\ e^{-\varepsilon m} &\leq \frac{\delta}{|H|} \\ -\varepsilon m &\leq \ln \frac{\delta}{|H|} \\ m &\geq \frac{1}{\varepsilon} \ln \frac{|H|}{\delta} \end{aligned}$$

3.4.2 המקרה $c_t \notin H$

נסמן:

$$h^* = \arg\min_{h \in H} \text{error}(h)$$

אזי, לכל $h \in H$:

$$\text{error}(h) \geq \text{error}(h^*) > 0$$

נסמן:

$$\beta = \text{error}\left(h^*\right)$$

המטרה שלנו תהיה:

$$\text{error}\left(h\right) \leq \beta + \varepsilon$$

נגדיר:

$$\widehat{\text{error}}\left(h\right) = \frac{1}{m}\sum_{i=1}^{m} I\left(h\left(x_i\right) \neq c_t\left(x_i\right)\right)$$

הבחירה הטבעית שלנו תהיה:

$$\bar{h} = \arg\min_{h\in H} \widehat{\text{error}}\left(h\right)$$

בחירה זו נקראת גם ERM[2].
נבחר מדגם מספיק גדול בגודל m כך שבהסתברות $1-\delta$ מתקיים:

$$\forall h \in H.\, \left|\widehat{\text{error}}\left(h\right) - \text{error}\left(h\right)\right| \leq \frac{\varepsilon}{2}$$

כעת, נקבל כי:

$$\begin{aligned}\text{error}\left(\bar{h}\right) &\leq \widehat{\text{error}}\left(\bar{h}\right) + \frac{\varepsilon}{2} \leq \widehat{\text{error}}\left(h^*\right) + \frac{\varepsilon}{2} \\ &\leq \text{error}\left(h^*\right) + \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \beta + \varepsilon\end{aligned}$$

לסיום, בהינתן $h \in H$:

$$\begin{aligned}\Pr\left[\left|\widehat{\text{error}}\left(h\right) - \text{error}\left(h\right)\right| \geq \varepsilon/2\right] &\leq 2e^{-2(\varepsilon/2)^2 m} \cdot |H| \leq \delta \\ e^{-2(\varepsilon/2)^2 m} &\leq \frac{\delta}{2\left|H\right|} \\ -2\left(\frac{\varepsilon}{2}\right)^2 m &\leq \ln\frac{\delta}{2\left|H\right|} \\ m &\geq \frac{8}{\varepsilon^2}\cdot\ln\frac{2\cdot|H|}{\delta}\end{aligned}$$

הערה למה במקרה הראשון (realizable), m מתנהג ביחס הפוך ל־ε, ואילו במקרה השני (non-realizable), m מתנהג ביחס הפוך ל־ε^2? ההבדל הוא שבמקרה הראשון, דגימה אחת מספיקה כדי לפסול השערה. לעומת זאת, במקרה השני, גם אם h אינה עקבית, היא עלולה להיות ההשערה הטובה ביותר.

[2] ראשי תיבות של *Empirical Risk Management*.

אלגוריתם 1.3 אלגוריתם ELIM ללמידת Boolean Disjunctions

בהינתן דגימה בגודל m, מתחילים עם L.
לכל דגימה שלילית $\langle x, - \rangle$, מוציאים את כל הליטרלים החיוביים.
כעת, כל הדוגמאות השליליות מסווגות נכון. גם כל הדוגמאות החיוביות מסווגות נכון, כי $c_t \subseteq L_{\text{final}}$.

3.4.3 דוגמה – למידת Boolean Disjunctions

נתונים n משתנים בולאנים – $x_1, \ldots, x_n$, וקבוצת ליטרלים:

$$L = \{x_1, \bar{x_1}, \ldots, x_n, \bar{x_n}\}$$

נגדיר *disjunction* על ידי ביטוי OR של ליטרלים. למשל: $x_1 \vee x_3 \vee \bar{x_5}$.
נגדיר את $C = H$ להיות מחלקת ה־disjunctions. אזי: $|C| = |H| = 3^n$ (כי כל משתנה בולאני יכול להופיע בשני ליטרלים או לא להופיע, ששקול להופעת שני הליטרלים המתאימים). $X = \{0,1\}^n$ ופונקציית המטרה שלנו היא:

$$c_t = \bigvee_{\substack{j \in S \\ \ell_j \in L}} \ell_j$$

נשים לב לעובדה שבהינתן דגימה שלילית $\langle x, - \rangle$, ניתן לפסול את כל הליטרלים החיוביים שלה. למשל, $\langle 001, - \rangle$ אומר לנו כי $\bar{x_1}, \bar{x_2}, x_3 \notin c_t$. זה הרעיון שעומד בבסיסו של האלגוריתם ELIM (אלגוריתם 1.3).
לפי חישוב קודם:

$$m \geq \frac{1}{\varepsilon} \ln \frac{|H|}{\delta} = \frac{n \ln 3}{\varepsilon} + \frac{1}{\varepsilon} \ln \frac{1}{\delta}$$

3.4.4 דוגמה – למידת Pairity

במקרה הזה:

$$\begin{aligned} X &= \{0,1\}^n \\ c_t &= \bigoplus_{j \in S} x_j \end{aligned}$$

למשל, $c_t = x_1 \oplus x_3 \oplus x_5$. אזי $|C| = 2^n$.
אפשר להסתכל על הבעיה כאל בעיית פתרון מערכת משוואות. נגדיר את z_i כך ש־$z_i = 1$ פירושו $i \in S$. אזי נוכל להסיק מערכת משוואות בצורה פשוטה. למשל, מ־$\langle 0110, + \rangle$ נסיק את המשוואה $z_2 + z_3 = 1$, ומ־$\langle 1111, - \rangle$ נסיק את המשוואה $z_1 + z_2 + z_3 + z_4 = 0$ (כמובן שהמשוואות הן מעל $\mathbb{Z}_2$).
קיבלנו מערכת משוואות אותה צריך לפתור. נסיק כי:

$$\begin{aligned} m &\geq \frac{1}{\varepsilon} \ln \frac{|H|}{\delta} = \frac{1}{\varepsilon} \ln \frac{2^n}{\delta} \\ &= \frac{n \ln 2}{\varepsilon} + \frac{1}{\varepsilon} \ln \frac{1}{\delta} \end{aligned}$$

3.5 Occam Razor

Enteties should not be multiplied unnecessarily

(מתוך דבריו של William Occam בשנת 1320).
המשמעות שלנו למשפט שלו היא שניתן ל־H לגדול יחד עם גודל המדגם m.

הגדרה (α, β) הוא אלגוריתם *Occam* ללימוד מחלקה C על ידי מחלקה H אם $\alpha \geq 0$ ו־$0 \leq \beta < 1$, ובהינתן דגימה בגודל m, האלגוריתם מוציא השערה $h \in H$ כך ש־h עקבית, וגודל h חסור על ידי $n^\alpha m^\beta$ (כאשר n הוא גודל דוגמה בודדת בביטים).

3.5.1 אלגוריתמי Occam ומודל ה־PAC

משפט יהי A אלגוריתם Occam עבור C המשתמש ב־H. אזי A הוא אלגוריתם PAC, כאשר:

$$m \geq \left(\frac{n^\alpha}{\varepsilon} \ln 2\right)^{\frac{1}{1-\beta}} + \frac{2}{\varepsilon} \ln \frac{1}{\delta}$$

הוכחה נקבע את m ו־n. אזי A מחזיר השערה h עם לכל היותר $n^\alpha m^\beta$ ביטים. לכן, מספר ההשערות האפשריות חסום על ידי $|H| \leq 2^{n^\alpha m^\beta}$.
לכן:

$$\begin{aligned}
m &\geq \frac{1}{\varepsilon} \ln \frac{|H|}{\delta} \geq \frac{n^\alpha m^\beta \ln 2}{\varepsilon} + \frac{1}{\varepsilon} \ln \frac{1}{\delta} \\
m &\geq \max\left\{\frac{2n^\alpha m^\beta \ln 2}{\varepsilon}, \frac{2}{\varepsilon} \ln \frac{1}{\delta}\right\} \\
m &\geq \frac{2n^\alpha m^\beta \ln 2}{\varepsilon} \\
m &\geq \left(\frac{2n^\alpha}{\varepsilon} \ln 2\right)^{\frac{1}{1-\beta}}
\end{aligned}$$

מש"ל. □

3.5.2 דוגמה – למידת OR של k משתנים

המטרה: להוריד את גודל המדגם מ־$O(n)$ ל־$O(k \log n)$.
לצורך הדוגמה, ניזכר בבעיית Set-Cover: הקלט הוא $S_1, \ldots, S_t \subseteq U$, ואנחנו מחפשים $S_{i_1}, \ldots, S_{i_\ell} \subseteq U$, כך ש:

$$\bigcup_j S_{i_j} = U$$

נשתמש באלגוריתם החמדן לפתרון Set-Cover (אלגוריתם 2.3).
ננתח אותו: נניח שיש כיסוי בגודל k. אזי:

$$\forall j. \exists t \in S_{\text{opt}} \cdot |U_j \cap S_t| \geq \frac{|U_j|}{k}$$

אלגוריתם 2.3 האלגוריתם החמדן ל־Set-Cover

1. $S \leftarrow \emptyset$, $j \leftarrow 0$, $U_0 \leftarrow U$
2. כל עוד $U_j \neq \emptyset$:
   (א) בחר $S_i = \arg\max_{S_r} \{|S_i \cap U_j|\}$
   (ב) $S \leftarrow S \cup \{i\}$
   (ג) $U_{j+1} \leftarrow U_j \setminus S_i$
   (ד) $j \leftarrow j + 1$
3. החזר S

לכן:

$$\begin{aligned} |U_{j+1}| &\leq |U_j| - \frac{|U_j|}{k} = \left(1 - \frac{1}{k}\right)|U_j| \\ &= \left(1 - \frac{1}{k}\right)^{j+1}|U_0| \end{aligned}$$

עבור $1 + k \ln |U|$ איטרציות, כיסינו את כל האיברים.
כדי ללמוד OR, נריץ ELIM (אלגוריתם 1.3). נבצע רדוקציה ל־Set-Cover:

$$U = \{x \mid \langle x, + \rangle \in S\}$$

(כל הדגימות החיוביות)

$$S_{\ell_i} = \{x \in U \mid \ell_i \in x\}$$

נריץ את האלגוריתם החמדן ל־Set-Cover (אלגוריתם 2.3), ונקבל $1 + k \ln m$.
גודל הקידוד:

$$\begin{aligned} (1 + k \ln m) \log (2n) &= O\left(l \ln m \ln n\right) \\ m &> \frac{k}{\varepsilon} \ln m \ln n + \frac{1}{\varepsilon} \ln \frac{1}{\delta} \end{aligned}$$

או בהצגה קצת שונה:

$$m > c \cdot \frac{k}{\varepsilon} \ln^2 n + \frac{1}{\varepsilon} \ln \frac{1}{\delta}$$

פרק 4

מודל Online[1]

נחשוב על הדוגמה הבאה: רובוט מסווג תפוזים לאיכות טובה או לא טובה. לאחר כל סיווג הוא מקבל משוב ממומחה. הרובוט יכול לעדכן את הפונקציה שלו, ואז לעבור לתפוז הבא.
תיאור המודל הכללי שלנו:

1. האלגוריתם מקבל דגימה x.
2. האלגוריתם נותן תחזית לסיווג של x – $b(x)$ (מכונה השערה נוכחית, או *current hypothesis*).
3. האלגוריתם מקבל את הסיווג הנכון – $c^{\star}(x)$ (זוהי פונקציית המטרה).
4. האלגוריתם ממשיך לדגימה הבאה.

נתבונן במודל *Adverserial* – כלומר, סדרת הקלט הגרועה ביותר שתיתן לנו את מספר השגיאות $c^{\star}(x) \neq b(x)$ המקסימלי.

4.1 למידה של מפריד לינארי

יש אוסף של נקודות חיוביות ואוסף של נקודות שליליות ב־$\mathbb{R}^n$ (או $\{0,1\}^n$[2]). אנחנו רוצים למצוא וקטור w וחסם w_0 כך ש־$w \cdot x = w_0$ (על מישור).
האלגוריתם שלנו יהיה $w \cdot x > 0$ גורר סיווג חיובי, $w \cdot x < 0$ גורר סיווג שלילי. בלי הגבלת הכלליות, $w_0 = 0$ (נוכל תמיד להוסיף קואורדינטה שתמיד תהיה 1).

4.1.1 אלגוריתם Perceptron

הרעיון המרכזי העומד בבסיס האלגוריתם הוא שאם אין שגיאה, אין צורך לעדכן את ההשערה הנוכחית שלנו, ואם יש שגיאה, אז נעדכן את ההשערה הנוכחית בכיוון השגיאה.
הנחת יסוד שלנו באלגוריתם תהיה ש־$\|x_t\| = 1$, כלומר שכל הקלטים יהיו וקטור יחידה.
אלגוריתם 1.4 מתאר את מהלך האלגוריתם Perceptron.

משפט תהי S סדרת דוגמאות מסווגות שעקבית עם מפריד לינארי $w^{\star}$ ($\|w^{\star}\| = 1$). אזי מספר השגיאות n של אלגוריתם Perceptron על הסדרה S חסום על ידי $1/\gamma^2$, כאשר:

$$\gamma = \min_{x \in S} \frac{|w^{\star} \cdot x|}{\|x\|}$$

[1] שיעור שהתקיים בתאריך 11.11.2012. מבוסס על סיכומיו של אולג.
[2] לרוב זה לא ישפיע על האלגוריתם.

אלגוריתם 1.4 אלגוריתם Perceptron

1. נתחיל עם $w_1 = \vec{0}$ ו־$t = 1$.
2. בהינתן x_t נסווג $+$ אם ורק אם $w_t \cdot x_t > 0$.
3. אם שגינו, נשים $w_{t+1} \leftarrow w_t + \ell(x_t) \cdot x_t$, כאשר $\ell(x_t)$ הוא ה־$label$:

$$\ell(x_t) = \begin{cases} +1 & x_t \text{is positive} \\ -1 & x_t \text{is negative} \end{cases}$$

4. נמשיך ל־$t+1$.

γ מכונה המפריד או ה־$margin$. כאשר $\|x\|$ מנורמל, זהו המרחק המינימלי מהעל־מישור.

הוכחה נתבונן בשני המדדים $w_t \cdot w^\star$ ו־$\|w_t\|$. נניח כי בכל צעד האלגוריתם מבצע שגיאה (שאר המהלכים אינם רלוונטיים).

טענת עזר 1

$$w_{t+1} \cdot w^\star \geq w_t \cdot w^\star + \gamma$$

הוכחה נניח כי x דגימה חיובית (ויש עליה שגיאה). אזי:

$$\begin{aligned} w_{t+1} \cdot w^\star &= (w_t + x) \cdot w^\star = w_t \cdot w^\star + \underbrace{x \cdot w^\star}_{>0} \\ &\geq w_t \cdot w^\star + \gamma \end{aligned}$$

כאשר המעבר האחרון נובע מהגדרת γ.
עבור x שלילי נקבל הוכחה דומה. □

טענת עזר 2

$$\|w_{t+1}\|^2 \leq \|w_t\|^2 + 1$$

הוכחה שוב נניח ש־x דגימה חיובית. אזי:

$$\begin{aligned} \|w_{t+1}\|^2 &= \|w_t + x\|^2 = \|w_t\|^2 + \|x\|^2 + \underbrace{2x \cdot w_t}_{<0} \\ &\leq \|w_t\|^2 + \left\|x^2\right\| = \|w_t\|^2 + 1 \end{aligned}$$

נשים לב כי $2x \cdot w_t < 0$ כי הייתה שגיאה. □

לאחר M שגיאות:

$$\begin{aligned} w_{M+1} \cdot w^\star &\geq \gamma \cdot M \\ \|w_{M+1}\|^2 &\leq \sqrt{M} \end{aligned}$$

לכן:

$$\begin{aligned} \gamma \cdot M &\leq w_{M+1} \cdot w^\star \leq w_{M+1} \cdot \frac{w_{M+1}}{\|w_{m+1}\|} \\ &= \|w_{M+1}\| \leq \sqrt{M} \\ M &\leq \frac{1}{\gamma^2} \end{aligned}$$

וסיימנו. □

מה קורה אם אין מפריד מושלם?
נסמן: TD_γ – המרחק שצריך להזיז את הנקודות כדי לקבל מפריד של γ. טענת עזר 1 תהפוך ל:

$$w_{t+1} \cdot w^\star \geq w_t \cdot w^\star + \gamma - \mathrm{D}_{\gamma,t}$$

כאשר $\mathrm{D}_{\gamma,t}$ הוא המרחק שצריך להזיז את x_t כדי לקבל את המפריד γ.
לכן:

$$w_{M+1} \cdot w^\star \geq \gamma \cdot M - \mathrm{TD}_\gamma$$

טענת עזר 2 תישאר נכונה, ואז נקבל:

$$\sqrt{M} \geq \gamma \cdot M - \mathrm{TD}_\gamma$$

חסם לפתרון:

$$M \leq \frac{1}{\gamma^2} + \frac{2}{\gamma} \cdot \mathrm{TD}_\gamma$$

4.1.2 אלגוריתם Margin Perceptron

אלגוריתם 2.4 מתאר את מהלך האלגוריתם.

משפט לכל סדרת נקודות S עקביות עם מפריד $w^\star$ ($\|w^\star\| = 1$) ו־$\gamma = \min_{x \in S} \frac{\|w^\star \cdot x\|}{\|x\|}$.
אז מספר השגיאות חסום על ידי $\frac{12}{\gamma^2}$.

הערה הייתרון ב־Margin Perceptron הוא שמקביל מפריד יותר טוב עם margin גדול יותר.

אלגוריתם 2.4 אלגוריתם Margin Perceptron

1. נגדיר: $w_1 = \ell(x_1) \cdot x_1$.
2. נחזיר את התחזית שלנו:
   (א) אם $\frac{\gamma}{2} \leq \frac{w_t \cdot x}{\|w_t\|}$ אז נאמר חיובי.
   (ב) אם $-\frac{\gamma}{2} \geq \frac{w_t \cdot x}{\|w_t\|}$ אז נאמר שלילי.
   (ג) אחרת נאמר שגיאה (margin mistake).
3. אם התחזית שלנו שגתה, נעדכן:

$$\begin{aligned} w_{t+1} &\leftarrow w_t + \ell(x_t) \cdot x_t \\ t &\leftarrow t+1 \end{aligned}$$

טענת עזר 1

$$w_{t+1} \cdot w^\star \geq w_t \cdot w^\star + \gamma$$

הוכחה

$$\begin{aligned} w_{t+1} \cdot w^\star &= (w_t + \ell(x_t) \cdot x_t) \cdot w^\star \\ &= w_t \cdot w^\star + \ell(x_t) \cdot x_t \cdot w^\star \\ &\geq w_t \cdot w^\star + \gamma \end{aligned}$$

טענת עזר 2

$$\|w_{t+1}\| \leq \|w_t\| + \frac{1}{2 \cdot \|w_t\|} + \frac{\gamma}{2}$$

4.2 מודל Margin Bound

נניח כי $c^\star \in C$ פונקציית המטרה.
בשלב t:

1. האלגוריתם מקבל x_t.
2. האלגוריתם בוחר סיווג $b_t = h_t(x_t)$.
3. האלגוריתם רואה את הסיווג הנכון $c^\star(x_t)$.

הגדרה למחלקה C יש אלגוריתם $\mathcal{A}$ עם חסם שגיאה M אם לכל $c^\star \in C$ ולכל סדר דוגמאות S מספר השגיאות הוא לכל היותר M.

כמו כן, נניח כי C סופית.

4.2.1 האלגוריתם Consistent (CON)

בשלב t נגדיר את C_t להיות מחלקת כל ההשערות העקביות. נבחר $h_t \in C_t$ ונחזיר $b_t = h_t(x_t)$.
אזי:

- $C_{t+1} \subseteq C_t$.
- אם יש שגיאה בזמן t אז $C_{t+1} \subset C_t$.

לכן, $M \leq |C| - 1$.

4.2.2 אלגוריתם חציה (HAL)

כמו קודם, נגדיר את C_t להיות מחלקת כל ההשערות העקביות בזמן t.
כמו כן, נגדיר:

$$\begin{aligned} \text{one} &= \{c \in C_t \mid c(x_t) = 1\} \\ \text{zero} &= \{c \in C_t \mid c(x_t) = 0\} \end{aligned}$$

התחזית תהיה 1 אם ורק אם $|\text{zero}| < |\text{one}|$. גם כאן:

- $C_{t+1} \subseteq C_t$.
- אם יש שגיאה בשלב t, אז:

$$|C_{t+1}| \leq \frac{|C_t|}{2}$$

לכן נסיק כי $M \leq \log_2 |C|$.

4.3 הקשר בין Mistake Bound ומודל ה־PAC

בהינתן אלגוריתם שמרני $\mathcal{A}$ עם חסם שגיאה M, אפשר להגדיר אלגוריתם PAC לאותה הבעיה, $\mathcal{A}_{\text{PAC}}$. אלגוריתם 3.4 מציג כיצד ניתן לעשות זאת.

משפט $\mathcal{A}_{\text{PAC}}$ לומר PAC את C.

הוכחה בכל בלוק שלא עצרנו, ביצענו שגיאה אחת לפחות. כמו כן, $\mathcal{A}$ מבצע לכל היותר M שגיאות.
לכן, אם הגענו ל־M שגיאות, ההשערה של A מושלמת. אחרת, בבלוק האחרון ראינו $\frac{1}{\varepsilon} \cdot \ln \frac{M}{\delta}$ דגימות, ואת כולן סיווגנו נכון.
נניח של־h_i יש שגיאה גדולה או שווה ל־ε. אזי:

$$\Pr\left[\forall j. h_i(x_j) = c^\star(x_j)\right] \leq (1-\varepsilon)^{1/\varepsilon \cdot \ln M/\delta} \leq \frac{\delta}{M}$$

כלומר, ההסתברות ש־$\mathcal{A}_{\text{PAC}}$ יוציא השערה רעה (עם שגיאה גדולה או שווה ל־ε) היא $M \cdot \frac{\delta}{M} = \delta$. □

אלגוריתם 3.4 מציאת אלגוריתם PAC מ־Mistake Bound

1. ניקח מדגם בגודל $M \cdot \frac{1}{\varepsilon} \cdot \ln \frac{M}{\delta}$.
2. נחלק ל־M קבוצות שוות.
3. נריץ את $\mathcal{A}$ על הקבוצה ה־i.
   (א) אם לא ביצע שגיאה, נחזיר את ההשערה הנוכחית.
   (ב) אם ביצע שגיאה, נמשיך לקבוצה ה־$i+1$.
4. אם סיימנו את M הקבוצות, נחזיר את ההשערה הנוכחית.

אם הגענו לשלב זה, סימן ש־$\mathcal{A}$ ביצע M שגיאות. מכיוון שזהו חסם השגיאה, נקבל השערה מושלמת.

אלגוריתם 4.4 אלגוריתם Online ללמידה של OR

1. אתחול: $L = \{x_1, \bar{x_1}, x_2, \bar{x_2}, \dots\}$.
2. בזמן t:
   (א) מקבלים דגימה $z = z_1 z_2 \cdots z_n$.
   (ב) נותנים תחזית לפי $h_L(z)$.
   (ג) אם הייתה שגיאה:
      i. נגדיר $S_z = \{\ell_i \mid \ell_i \text{is positive in } z\}$.
      ii. נשים $L \leftarrow L \setminus S_z$.

4.4 למידה של OR

היה לנו אלגוריתם אלימינציה[3]. אלגוריתם 4.4 הוא גרסת Online שלו.

משפט מספר השגיאות יהיה לכל היותר $n+1$ שגיאות (כאשר $|L|$ התחלתי הוא בגודל $2n$).

הוכחה בשגיאה הרשונה, נפסול בדיוק n ליטרלים. כל שגיאה נוספת פוסלת לפחות ליטרל אחד נוסף. □

4.5 אלגוריתם Winnow

אלגוריתם 5.4 מתאר את אלגוריתם Winnow לחישוב מפריד לינארי. פה אנחנו פותרים את OR על ידי מפריד לינארי במרחב $\{0,1\}^n$.

משפט אלגוריתם Winnow לומד OR של r משתנים חיוביים עם לכל היותר $O(r \log n)$ שגיאות.

[3] ראה אלגוריתם 1.3.

אלגוריתם 5.4 אלגוריתם Winnow

1. נאתחל $w^1 = (1, 1, \ldots, 1)$, $w_0 = n$ (בניגוד ל־0 שהיה קודם).
2. בהינתן נקודה x, נסווג אותה כחיובית אם ורק אם $w^t \cdot x \geq n$.
3. אם הייתה שגיאה:
   (א) אם $h(x) = 0$ ו־$c^\star(x) = 1$, אז $w_i^{t+1} \leftarrow 2w_i^t$: $\forall i. x_i = 1$.
   (ב) אם $h(x) = 1$ ו־$c^\star(x) = 0$, אז $x_i^{t+1} \leftarrow \frac{w_i^t}{2}$: $\forall i. x_i = 1$.

הוכחה נסמן: $S = \{x_{i_1}, \ldots, x_{i_r}\}$ כך ש־$c^\star(x) = x_{i_1} \vee \cdots \vee x_{i_r}$. נסמן: $w_r = \{w_{i_1}, \ldots, w_{i_r}\}$. נגדיר: $w_i(t)$ – המשקל של x_i בזמן t. נסמן:

$$\mathrm{TW}(t) = \sum_{i=1}^{n} w_i(t)$$

אם יש שגיאה על דגימה חיובית (כלומר $c^\star(x) = 1$), אזי:

$$\exists i_j \in S. w_{i_j}(t+1) = 2w_{i_j}(t)$$

אם יש שגיאה על דגימה שלילית, אזי:

$$\forall i_j \in S. w_{i_j}(t+1) = w_{i_j}(t)$$

לכן, לכל $i_j \in S$, $w_{i_j}(t)$ היא פונקציה מונוטונית עולה. כמו כן, לכל $i_j \in S$, w_{i_j} לא יכול לגדול יותר מ־$1 + \log n$ פעמים (כי אם הוא גדל כמות כזאת של פעמים, כבר לא נטעה אם i_j מופיע ב־x). לכן:

$$M_t \leq r \cdot (1 + \log n)$$

אם $h(x) = 0$ ו־$c^\star(x) = 1$:

$$\begin{aligned} \mathrm{TW}(t+1) &= \mathrm{TW}(t) + \sum_{i=1}^{n} x_i w_i(t) \\ &\leq \mathrm{TW}(t) + n \end{aligned}$$

אם $h(x) = 1$ ו־$c^\star(x) = 0$:

$$\begin{aligned} \mathrm{TW}(t+1) &= \mathrm{TW}(t) - \frac{1}{2} \cdot \sum_{i=1}^{n} x_i w_i(t) \\ &\leq \mathrm{TW}(t) - \frac{n}{2} \end{aligned}$$

נשים לב כי $\mathrm{TW}(t) > 0$. ואז:

$$\begin{aligned} 0 &< \mathrm{TW}(t) \leq \mathrm{TW}(0) + n \cdot M_+ - \frac{n}{2} \cdot M_- \\ M_- &\leq 2M_+ + \frac{2}{n} \cdot \underbrace{\mathrm{TW}(0)}_{n} = 2M_+ + 2 \end{aligned}$$

לכן נסיק כי:

$$M = M_- + M_+ \leq 3r \cdot (1 + \log n) + 2$$

וסיימנו. □

פרק 5

Regret Minimization[1]

5.1 מבוא

לפי מודל Online, בזמן t:

1. מקבלים x_t.
2. נותנים תחזית b_t.
3. רואים את $c^\star(x_t)$.

ראינו את אלגוריתם החציה (HAL): אם $c^\star \in H$, אזי HAL יעשה לכל היותר $O(\log|H|)$ טעויות.
אם $c^\star \notin H$, נרצה לחזות כמו ההשערה הטובה ביותר ב־H.

5.1.1 המודל האלגוריתמי

המודל האלגוריתמי שלנו יהיה כזה: לכל $h \in H$, נשמור w_h, כך ש־$\sum_{h\in H} w_h = 1$ ו־$\forall h \in H. w_h \geq 0$. התחזית תהיה ממוצע משוקלל.
נסמן ב־ℓ_h^t את ההפסד של ההשערה h בזמן t. אזי $\ell_h^t \in [0,1]$, ותוחלת ההפסד בזמן t היא:

$$\sum_{h\in H} \ell_h^t \cdot w_h^t$$

5.1.2 External Regret

נניח כי:

$$\begin{aligned} L_h^T &= \sum_{1\leq t\leq T} \ell_h^t \\ L_{\text{best}}^T &= \min_{h\in H} L_h^T \end{aligned}$$

[1] שיעור שהתקיים בתאריך 18.11.2012.

אלגוריתם 1.5 Deterministic Greedy

1. עבור $t = 1$ נבחר h_1 (שרירותי).
2. עבור $t > 1$ נבחר:

$$h_t = \arg\min_{h \in H} L_h^{t-1}$$

אזי ההפסד של האלגוריתם $\mathcal{A}$ יהיה:

$$L_{\mathcal{A}}^T = \sum_{1 \le t \le T} \sum_{h \in H} \ell_h^t \cdot w_h^t$$

היעד שלנו יהיה למצוא אלגוריתם שיקיים את $L_{\mathcal{A}}^T - L_{\text{best}}^T \le \text{Regret}$. המטרה שלנו היא שה־Regret יהיה $o(T)$.

5.2 אלגוריתמים

5.2.1 אלגוריתם Deterministic Greedy ($\mathcal{G}$)

אלגוריתם 1.5 מתאר את מהלך אלגוריתם Deterministic Greedy.

משפט

$$L_{\mathcal{G}}^T \le |H| \cdot L_{\text{best}}^T + |H| - 1$$

הוכחה נגדיר את B_k להיות אוסף הזמנים בהם $L_{\text{best}}^t = k$.
בצעד הראשון, ב־B_k יש לכל היותר $|H|$ השערות עבורן $L_{\text{best}}^{t_0} = k$. כל שגיאה תוריד את מספר ההשערות עם $L_h = k$ בלפחות אחד.
אחרי לכל היותר $|H|$ שגיאות, L_{best} יגדל ל־$k + 1$. לכן:

$$L_{\mathcal{G}}^T \le |H| \cdot L_{\text{best}}^T + |H| - 1$$

וסיימנו. □

משפט לכל אלגוריתם דטרמיניסטי $\mathcal{D}$ יש סדרת הפסדים עבורה:

$$L_{\mathcal{D}}^T \ge |H| \cdot L_{\text{best}}^T + (T \mod |H|)$$

הוכחה נבחר סדרת הפסדים כך שבזמן t, $\mathcal{D}$ בוחר את h^t ומתקיים $\ell_{h^t}^t = 1$, ולכל $h \ne h^t$ מתקיים $\ell_h^t = 0$.
מבניית הסדרה,

$$L_{\mathcal{D}}^T = T$$

אלגוריתם 2.5 Randomized Greedy

בזמן t, נסמן:

$$H^t = \left\{ h \in H \mid L_h^t = L_{\text{best}}^t \right\}$$

נגדיר התפלגות מעל H באמצעות:

$$p_h^t = \begin{cases} \frac{1}{|H^{t-1}|} & h \in H^{t-1} \\ 0 & \text{otherwise} \end{cases}$$

האלגוריתם יבחר $h \in H$ לפי ההתפלגות שהגדרנו.

כמו כן:

$$\sum_{h \in H} L_h^T = T$$

לכן:

$$\exists h \in H . L_h^T \leq \left\lfloor \frac{T}{|H|} \right\rfloor$$

וסיימנו. □

5.2.2 אלגוריתם Randomized Greedy ($\mathcal{GR}$)

אלגוריתם 2.5 מתאר את מהלך אלגוריתם Randomized Greedy.

משפט בתוחלת:

$$L_{\mathcal{RG}}^T \leq (\ln |H| + 1) \cdot L_{\text{best}}^T + \ln |H|$$

הוכחה נגדיר את B_k כמו קודם. היריב יעדיף לבצע שגיאה אחת בכל צעד.
נניח שהוא עושה r שגיאות בבת אחת. אז ההפסד יהיה $\frac{r}{m}$ (כאשר $m = \left|H^{t-1}\right|$). אם הוא יעשה שגיאה אחת בכל פעם, נקבל:

$$\frac{1}{m} + \frac{1}{m-1} + \cdots + \frac{1}{m-r+1}$$

נשים לב כי:

$$\frac{r}{m} < \frac{1}{m} + \frac{1}{m-1} + \cdots + \frac{1}{m-r+1}$$

$$\begin{aligned} \mathbf{E}\left[L_{\mathcal{RG}}^{B_k}\right] &= \frac{1}{m} + \frac{1}{m-1} + \cdots + \frac{1}{m-r+1} \\ &= \sum_{i=1}^{|H|} \frac{1}{i} \leq \ln |H| + 1 \end{aligned}$$

אלגוריתם 3.5 Randomized Weighted Majority

נגדיר את המשקל שדל $h \in H$ בזמן t להיות:

$$w_h^t = (1-\eta)^{L_h^t}$$

נשים לב שאפשר לחשוב על הגדרה זו גם כעל הגדרה רקורסיבית:

$$w_h^{t+1} = \begin{cases} w_h^t & \ell_h^t = 0 \\ w_h^t \cdot (1-\eta) & \ell_h^t = 1 \end{cases}$$

(כאשר $w_h^0 = 1$).
נהפוך את המשקולות להסתברות: נגדיר:

$$W^t = \sum_{h \in H} w_h^t$$

נגדיר גם:

$$p_h^t = \frac{w_h^t}{W^t}$$

האלגוריתם יבחר השערה $h \in H$ לפי ההתפלגות שהגדרנו.

לכן:

$$L_{\mathcal{RG}}^T \leq (\ln|H|+1) \cdot L_{\text{best}}^T + \ln|H|$$

וסיימנו. □

5.2.3 אלגוריתם Randomized Weighted Majority ($\mathcal{RWM}$)

אלגוריתם 3.5 מתאר את אלגוריתם Randomized Weighted Majority.

משפט עבור $\eta \in (0, 1/2]$:

$$L_{\mathcal{RWM}}^T \leq (1+\eta) \cdot L_{\text{best}}^T + \frac{\log|H|}{\eta}$$

כמו כן, עבור $\eta = \min\left\{\frac{1}{2}, \sqrt{\frac{\log|H|}{T}}\right\}$:

$$L_{\mathcal{RWM}}^T \leq L_{\text{best}}^T + 2 \cdot \sqrt{T \cdot \log|H|}$$

הוכחה נסמן ב־F^t את ההסתברות לשגיאה של $\mathcal{RWM}$:

$$F^t = \frac{1}{W^t} \cdot \sum_{\substack{h \in H \\ \ell_h^t = 1}} w_h^t = \sum_{\substack{h \in H \\ \ell_h^t = 1}} p_h^t$$

אזי:

$$\begin{aligned}
W^{t+1} &= \sum_{h\in H} w_h^{t+1} \\
&= \sum_{\substack{h\in H \\ \ell_h^{t+1}=0}} w_h^{t+1} + \sum_{\substack{h\in H \\ \ell_h^{t+1}=1}} w_h^{t+1} \\
&= \sum_{\substack{h\in H \\ \ell_h^{t+1}=0}} w_h^{t} + \sum_{\substack{h\in H \\ \ell_h^{t+1}=1}} w_h^{t}\cdot(1-\eta) \\
&= \sum_{h\in H} w_h^t - \eta\cdot \sum_{\substack{h\in H \\ \ell_h^{t+1}=1}} w_h^t \\
&= W^t - \eta\cdot F^t\cdot W^t \\
&= W^t\cdot\left(1-\eta\cdot F^t\right) \\
&= W^1\cdot\prod_{\tau=1}^{t}\left(1-\eta\cdot F^\tau\right) \\
&= |H|\cdot\prod_{\tau=1}^{t}\left(1-\eta\cdot F^\tau\right)
\end{aligned}$$

לכן:

$$\forall h\in H. W^{T+1}\geq w_h^{T+1}\geq (1-\eta)^{L_{\text{best}}^T}$$

לכן:

$$(1-\eta)^{L_{\text{best}}^T}\leq |H|\cdot\prod_{t=1}^{T}\left(1-\eta\cdot F^t\right)$$

לכן:

$$L_{\text{best}}^T\cdot\ln(1-\eta)\leq\ln|H|+\sum_{t=1}^{T}\ln\left(1-\eta\cdot F^t\right)$$

נשתמש בזהות $-z-z^2\leq\ln(1-z)\leq -z$:

$$L_{\text{best}}^T\left(-1-\eta^2\right)\leq\ln|H|+\sum_{t=1}^{T}\left(-\eta\cdot F^t\right)$$

לכן:

$$\eta\cdot\sum_{t=1}^{T}F^t\leq\left(\eta+\eta^2\right)\cdot L_{\text{best}}^T+\ln|H|$$

לכן:

$$L^T_{\mathcal{RWM}} \leq (1+\eta) \cdot L^T_{\text{best}} + \frac{\log|H|}{\eta}$$

אם נבחר:

$$\eta = \min\left\{\frac{1}{2}, \sqrt{\frac{\ln|H|}{T}}\right\}$$

אזי $L^T_{\text{best}} \leq T$, ואז:

$$L^T_{\mathcal{RWM}} \leq L^T_{\text{best}} + \eta \cdot T + \frac{\ln|H|}{\eta}$$

וסיימנו. □

5.3 חסמים תחתונים לאלגוריתמי Online ממושקלים

אחרי שהראנו חסמים עליונים ל־External Regret, עלינו לבדוק כמה טוב יכול להיות אלגוריתם Online ממושקל.
נבדוק את זה עבור שני מקרים.

5.3.1 טווח קצר – $T = \frac{1}{2} \cdot \log|H|$

נבחר הפסדים כך ש:

$$\forall h \in H. \forall 1 \leq t \leq T. \Pr\left[\ell^t_h = 0\right] = \frac{1}{2} = \Pr\left[\ell^t_h = 1\right]$$

נראה שבהסתברות גבוהה קיים $h \in H$ כך ש־$L^T_h = 0$.
ואכן:

$$\forall h \in H. \Pr\left[L^T_h = 0\right] = \left(\frac{1}{2}\right)^T = \frac{1}{\sqrt{|H|}}$$

לכן:

$$\Pr\left[L^T_h \neq 0\right] = 1 - \frac{1}{\sqrt{|H|}}$$

לכן:

$$\Pr\left[\forall h \in H. L^T_h \neq 0\right] = \left(1 - \frac{1}{\sqrt{|H|}}\right)^{|H|} \leq e^{-\sqrt{|H|}}$$

לכן:

$$\Pr\left[\exists h \in H. L^T_h = 0\right] \geq 1 - e^{-\sqrt{|H|}}$$

נעבור לכתיב תוחלות:

$$\begin{aligned}
\mathbf{E}\left[L_{\text{best}}^{T}\right] &\leq T\cdot e^{-\sqrt{|H|}}=\frac{1}{2}\cdot\log|H|\cdot e^{-\sqrt{|H|}}\\
\mathbf{E}\left[L_{\text{ON}}\right] &= \frac{1}{2}\cdot T=\frac{1}{4}\cdot\log|H|\\
\mathbf{E}\left[\text{Regret}\right] &\geq \mathbf{E}\left[L_{\text{ON}}\right]-\mathbf{E}\left[L_{\text{best}}^{T}\right]\\
&\geq \frac{1}{4}\cdot\log|H|-\frac{1}{2}\cdot\log|H|\cdot e^{-\sqrt{|H|}}\\
&\approx \frac{1}{4}\cdot\log|H|
\end{aligned}$$

5.3.2 כתלות בזמן – $|H|=2$

נניח כי $H=\{h_1,h_2\}$. כמו כן, נניח כי סדרת ההפסדים היא $(0,1)$ בהסתברות $1/2$, ו־$(1,0)$ בהסתברות $1/2$.

אזי, אם האלגוריתם ON מחליט להחזיר h_1 בהסתברות p ו־h_2 בהסתברות $1-p$:

$$\mathbf{E}\left[\ell_{\text{ON}}^{t}\right]=p\cdot\frac{1}{2}+(1-p)\cdot\frac{1}{2}=\frac{1}{2}$$

לכן:

$$\mathbf{E}\left[L_{\text{ON}}^{T}\right]=\frac{1}{2}\cdot T$$

ואז:

$$\begin{aligned}
\mathbf{E}\left[L_{\text{best}}^{T}\right] &= \mathbf{E}\left[\min\left\{L_{h_1}^{T},L_{h_2}^{T}\right\}\right]\\
&= \mathbf{E}\left[\frac{T}{2}-\frac{\left|L_{h_1}^{T}-L_{h_2}^{T}\right|}{2}\right]\\
&= \frac{T}{2}-\frac{1}{2}\cdot\mathbf{E}\left[\left|L_{h_1}^{T}-L_{h_2}^{T}\right|\right]\\
&= \Omega\left(\sqrt{T}\right)
\end{aligned}$$

לכן:

$$\mathbf{E}\left[\text{Regret}\right]\geq c\cdot\sqrt{T}$$

5.4 Multi-Arm Bandit

המודל שלנו יהיה מעט שונה. בזמן t, כאשר מבצעים פעולה $a\in A$, מקבלים ℓ_a^t (ולא את ℓ_b^t אם $a\neq b$).

מודל סטוכסטי: נניח כי לכל פעולה יש מ"מ X_a, כך ש־$\mathbf{E}\left[\ell_a^t\right]=\mathbf{E}\left[X_a\right]=\mu_a$. המטרה שלנו היא לבחור את $a\in A$ כך ש־ℓ_a^t הכי נמוך.

אלגוריתם 4.5 Test & Play

1. שלב ה־Test: נדגום כל פעולה $a \in A$ סדר גודל של $O\left(\frac{1}{\varepsilon} \cdot \ln \frac{|A|}{\delta}\right)$ פעמים.
2. לכל פעולה $a \in A$ נגדיר את ממוצע ההפסד $\hat{\mu}_a$.
3. שלב ה־Play: נבחר:

$$\hat{a}^\star = \arg\min_{a \in A} \hat{\mu}_a$$

בשאר הפעולות, נבחר תמיד ב־$\hat{a}^\star$.

5.4.1 אלגוריתם Test & Play

אלגוריתם Test & Play הוא האלגוריתם הפשוט ביותר ללמידת המודל. אלגוריתם 4.5 מתאר את מהלכו.

משפט בהסתברות $1 - \delta$ נקבל:

$$\forall a \in A. |\mu_a - \hat{\mu}_a| \leq \frac{\varepsilon}{2}$$

הוכחה ישירות מחסמי Chernoff. □

כעת נרצה להעריך את $|\mu^\star - \mu_{\hat{a}^\star}|$. מהמשפט:

$$\mu_{\hat{a}^\star} - \frac{\varepsilon}{2} \leq \hat{\mu}_{\hat{a}^\star} \leq \hat{\mu}_{a^\star} \leq \mu_{a^\star} + \frac{\varepsilon}{2}$$

לכן:

$$|\mu^\star - \mu_{\hat{a}^\star}| \leq \frac{\varepsilon}{2}$$

נחשב חסם עליון על ה־Regret של Test & Play.

- בשלב ה־Test:

$$O\left(\frac{|A|}{\varepsilon^2} \cdot \ln \frac{|A|}{\delta}\right)$$

- בשלב ה־Play:
  - בהסתברות $1 - \delta$ בחרנו ε־טוב, ואז ה־Regret הוא $\varepsilon \cdot T$.
  - בהסתברות δ, ה־Regret הוא $\delta \cdot T$.

הסכום שלהם הוא תוחלת ה־Regret. נבחר $\delta = \frac{1}{T}$. נרצה להביא למינימום את:

$$\frac{|A|}{\varepsilon^2} \cdot \ln |A| \cdot T + \varepsilon \cdot T + 1$$

אלגוריתם 5.5 Upper Confidence Bound

1. שלב האתחול: נדגום כל $a \in A$ פעם אחת. מהדגימות האלה נקבל את $\hat{\mu}_a$, ו־$T_a = 1$.
2. עבור $t > |A|$, נבחר את a^t שמוגדר על ידי:

$$a^t = \arg\min_{a \in A} \hat{\mu}_a - \sqrt{\frac{2 \cdot \log t}{T_a}}$$

כמו כן, נעדכן את $\hat{\mu}_a$ ו־T_a בהתאם.

זה נותן לנו $\varepsilon = \frac{1}{T^{1/3}}$ בקירוב, או בדיוק:

$$\varepsilon = \left(\frac{2\,|A| \ln |A| \cdot T}{T} \right)^{1/3}$$

כלומר, בתנאים אופטימליים נבלה $O\left(T^{2/3}\right)$ ב־Test. ואז:

$$\mathbf{E}\left[\text{Regret}\right] \leq \left(2 \cdot |A| \ln |A| \cdot T\right)^{1/3} \cdot T^{2/3}$$

5.4.2 אלגוריתם Upper Confidence Bound ($\mathcal{UCB}$)

אלגוריתם 5.5 מתאר את מהלכו של אלגוריתם Upper Confidence Bound.

משפט יהי $\Delta_a = \mu_a - \mu^\star$. אזי:

$$\text{Regret}\left(\mathcal{UCB}\right) \leq 8 \cdot \ln T \cdot \sum_{\substack{a \in A \\ a \neq a^\star}} \frac{1}{\Delta_a} + 4 \cdot \sum_{\substack{a \in A \\ a \neq a^\star}} \frac{1}{\Delta_a}$$

הוכחה ניזכר כי:

$$\mathbf{E}\left[\text{Regret}\right] = \sum_{\substack{a \in A \\ a \neq a^\star}} \Delta_a \cdot \mathbf{E}\left[T_a\right]$$

לכן, נרצה לחסום את $\mathbf{E}\left[T_a\right]$:

$$\begin{aligned} T_a &= 1 + \sum_{t=|A|+1}^{T} I\left\{a^t = a\right\} \\ &\leq \ell + \sum_{t=|A|+1}^{T} I\left\{a^t = a, T_a \geq \ell\right\} \\ &\leq \ell + \sum_{t} \sum_{\ell \leq r \leq T} \sum_{1 \leq s \leq T} I\left\{\hat{\mu}_a - \sqrt{\frac{2\log T}{r}} \leq \hat{\mu}_{a^\star}^s - \sqrt{\frac{2\log t}{s}}\right\} \end{aligned}$$

האינדיקטור מתקיים בלפחות אחד מהתנאים הבאים:

$$\hat{\mu}^s_{a^\star} \geq \mu^\star + \sqrt{\frac{2 \cdot \log t}{s}} \tag{5.4.1}$$

$$\mu^r_a \leq \hat{\mu_a} - \sqrt{\frac{2 \cdot \log t}{r}} \tag{5.4.2}$$

$$\mu^\star > \mu_a + 2 \cdot \sqrt{\frac{2 \cdot \log t}{r}} \tag{5.4.3}$$

את 5.4.1 ו־5.4.2 נחסום עם Chernoff בהסתברות $\frac{1}{t^4}$. לגבי 5.4.3: אם $\ell = \left\lceil \frac{8 \cdot \log t}{\Delta^2} \right\rceil$, 5.4.3 לא מתקיים, ואז נישאר עם $T_a \leq \ell + \sum \frac{1}{t^2}$, ו־$\sum \frac{1}{t^2}$ קבוע. □

פרק 6

Boosting[1]

6.1 למידה חלשה וחזקה

במודל PAC, מקבלים דוגמאות $\langle x, c^{\star}(x)\rangle$, וקיימת התפלגות D עבור x. המטרה היא למצוא השערה $h \in H$ כך ש־$\text{error}(h, c^{\star}) \leq \varepsilon$ בהסתברות $1-\delta$.

δ נקרא פרמטר הבטחון, ו־ε נקרא פרמטר הדיוק. שני הפרמטרים ביחד קובעים את חוזק הלמידה שלנו.

6.1.1 שיפור בפרמטר הבטחון

ההנחה: נתון אלגוריתם $\mathcal{A}$ שבהסתברות לפחות $1/2$ מחזיר השערה $h \in H$ כך ש:

$$\text{error}(h, c^{\star}) \leq \varepsilon$$

אלגוריתם 1.6 מתאר את אלגוריתם Boost-Confidence, שמשפר את פרמטר הבטחון. ננתח את האלגוריתם:

1. לכל i:

$$\Pr\left[\text{error}(h_i) \leq \frac{\varepsilon}{2}\right] \geq \frac{1}{2}$$

לכן:

$$\Pr\left[\exists i.\text{error}(h_i) \leq \frac{\varepsilon}{2}\right] \geq 1 - \left(\frac{1}{2}\right)^k$$

עבור $k = \log \frac{2}{\delta}$:

$$\begin{aligned}
\Pr\left[\exists i.\text{error}(h_i) \leq \frac{\varepsilon}{2}\right] &\geq 1 - \left(\frac{1}{2}\right)^k \\
&= 1 - \left(\frac{1}{2}\right)^{\log \frac{2}{\delta}} \\
&= 1 - \frac{1}{2^{\log \frac{2}{\delta}}} \\
&= 1 - \frac{1}{2/\delta} = 1 - \frac{\delta}{2}
\end{aligned}$$

[1] שיעור שהתקיים בתאריך 25.11.2012.

אלגוריתם 1.6 Boost-Confidence

נגדיר את אלגוריתם Boost-Confidence ($\mathcal{BC}$) שפועל על האלגוריתם $\mathcal{A}$ (אותו הוא מקבל כקלט):

1. הרץ את האלגוריתם $\mathcal{A}$ במשך $k = \log \frac{2}{\delta}$ פעמים עם רמת דיוק $\varepsilon/2$. הפלט שנקבל הוא $h_1, \ldots, h_k$.
2. נקח דגימה נוספת S בגודל:

$$m = \frac{2}{\varepsilon^2} \cdot \ln \frac{4k}{\delta} = O\left(\frac{1}{\varepsilon^2} \cdot \ln \frac{k}{\delta}\right)$$

עבור כל h_i, נחשב את $\widehat{\text{error}}(h_i)$.

3. נחזיר את:

$$\hat{h}^\star = \arg\min_{h_i} \widehat{\text{error}}(h_i)$$

---

2. עבור כל h_i: ניזכר כי:

$$\widehat{\text{error}}(h_i, c^\star) = \frac{1}{m} \cdot \sum_{j=1}^{m} I\left(h_i(x_j) \neq c^\star(x_j)\right)$$

לכן, מ־Chernoff:

$$\Pr\left[\left|\widehat{\text{error}}(h_i, c^\star) - \text{error}(h_i, c^\star)\right| > \frac{\varepsilon}{2}\right] \leq 2 \cdot e^{-2 \cdot \left(\frac{\varepsilon}{2}\right)^2 \cdot m}$$

נרצה לחסום את ההסתברות למאורע ה"רע" הזה על ידי $\frac{\delta}{2}$, ולכן נדרוש:

$$2 \cdot k \cdot e^{-2 \cdot \left(\frac{\varepsilon}{2}\right)^2 \cdot m} \leq \frac{\delta}{2}$$

נפתור את המשוואה:

$$\begin{aligned}
2 \cdot k \cdot e^{-2 \cdot \left(\frac{\varepsilon}{2}\right)^2 \cdot m} &\leq \frac{\delta}{2} \\
4 \cdot k \cdot e^{-2 \cdot \left(\frac{\varepsilon}{2}\right)^2 \cdot m} &\leq \delta \\
\ln 4k - 2 \cdot \left(\frac{\varepsilon}{2}\right)^2 \cdot m &\leq \ln \delta \\
2 \cdot \left(\frac{\varepsilon}{2}\right)^2 \cdot m &\geq \ln 4k - \ln \delta \\
\frac{\varepsilon^2}{2} \cdot m &\geq \ln \frac{4 \cdot k}{\delta} \\
m &\geq \frac{2}{\varepsilon^2} \cdot \ln \frac{4 \cdot k}{\delta}
\end{aligned}$$

אלגוריתם 2.6 אלגוריתם לשקילת למידה חלשה וחזקה

הקלט: דגימות $x_1, \ldots, x_m$ וסיווגן על ידי $c^\star$.
נניח כי H היא מחלקת השערות (חלשות), ו־$\mathcal{RM}$ הוא אלגוריתם Regret-Minimization.

1. בכל שלב t, האלגוריתם $\mathcal{RM}$ יבחר פילוג D_t מעל $x_1, \ldots, x_m$.
2. בהינתן D_t, קיימת השערה $h_t \in H$ עבורה:
$$\text{error}_{D_t}(h_t) \leq \frac{1}{2} - \gamma$$
3. ההפסד יהיה 1 לכל סיווג נכון של h_t, ו־0 לכל סיווג מוטעה.
4. אחרי T שלבים, נחזיר את $\text{MAJ}(h_1, \ldots, h_T)$.

לכן:
$$\begin{aligned} \text{error}\left(\hat{h}^\star\right) &\leq \widehat{\text{error}}\left(\hat{h}^\star\right) + \frac{\varepsilon}{2} \\ &\leq \widehat{\text{error}}(h_i) + \frac{\varepsilon}{2} \\ &\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \end{aligned}$$

6.1.2 שיפור בפרמטר הדיוק

נראה שקיים $\gamma > 0$ כך שלכל פונקציית מטרה $c^\star \in C$, לכל התפלגות D ולכל פרמטר דיוק δ, אלגוריתם $\mathcal{A}$ מוצא השערה $h \in H$ כך ש:
$$\text{error}(h, c^\star) \leq \frac{1}{2} - \gamma$$

דוגמה אם $x_1 = x_2 = 1$, אז $c^\star(x)$ היא פונקציה קשה. אחרת, $c^\star(x) = 0$. נניח כי D היא ההתפלגות האחידה.
ניתן להגיע ל־87.5% דיוק בקלות. כדי להעלות את החוזי הדיוק, היינו משנים את D כך שנתמקד במקרה $x_1 = x_2 = 1$.
זאת הסיבה לכך שאנחנו נראה שקיים γ לכל D, ולא שלכל D קיים γ.

הוכחת שקילות למידה חלשה וחזקה אלגוריתם 2.6 הוא האלגוריתם שיעזור לנו להוכיח את השקילות.
ננתח אותו:

1. בכל שלב מצאנו h_t. ההפסד של $\mathcal{RM}$ הוא $\left(\frac{1}{2} + \gamma\right)$. לכן, סך כל ההפסד יהיה $\left(\frac{1}{2} + \gamma\right) \cdot T$.
2. אם קיים x_i כך ש־$\text{MAJ}(h_1, \ldots, h_T)$ טועה עליו, אזי ההפסד של x_i הוא לכל היותר $\frac{T}{2}$. ה־Regret יהיה:
$$\left(\frac{1}{2} + \gamma\right) \cdot T \leq \frac{T}{2} + 2 \cdot \sqrt{T \cdot \log m}$$

נחלץ את T:

$$\begin{aligned}\gamma \cdot T &\leq 2 \cdot \sqrt{T \cdot \log m} \\ \gamma^2 \cdot T^2 &\leq 4 \cdot T \cdot \log m \\ \gamma^2 \cdot T &\leq 4 \cdot \log m \\ T &\leq \frac{4 \cdot \log m}{\gamma^2}\end{aligned}$$

6.2 בנייה רקורסיבית

נמצא שלוש השערות h_1, h_2 ו־h_3 כך ש־$\mathrm{MAJ}(h_1, h_2, h_3)$ משפר את הדיוק. אלגוריתם 3.6 מתאר איך עושים את זה.

הערכת השגיאה נעריך את השגיאה של האלגוריתם.
מה תהיה השגיאה אם ההסתברויות ב"ת? כלומר, מה תהיה השגיאה אם ההסתברות לשגיאה של h_1, h_2 ו־h_3 ב"ת?

$$\text{error} = p^3 + 3p^2(1-p) = 3p^2 - 2p^3$$

ברור שההסתברויות האלה אינן ב"ת, מכיוון ש־h_2 נבנתה על סמך h_1, ו־h_3 נבנתה בהתבסס על h_1 ו־h_2. למרות שההסתברויות אינן ב"ת, גם במודל שלנו נקבל שגיאה כזו.
נחלק את הדוגמאות לארבע קבוצות:

$$\begin{aligned}S_{cc} &= \{x \mid h_1(x) = c^\star(x) = h_2(x)\} \\ S_{ce} &= \{x \mid h_1(x) = c^\star(x) \neq h_2(x)\} \\ S_{ec} &= \{x \mid h_1(x) \neq c^\star(x) = h_2(x)\} \\ S_{ee} &= \{x \mid h_1(x) \neq c^\star(x) \neq h_2(x)\}\end{aligned}$$

ההסתברויות של הקבוצות האלו ביחס ל־D_1 הן p_{cc}, p_{ce}, p_{ec} ו־p_{ee} בהתאמה. השגיאה הכוללת שלנו תהיה:

$$\text{error} = p_{ee} + (p_{ce} + p_{ec}) \cdot p$$

(נזכיר כי p היא הסתברות השגיאה של h_3).
נסמן:

$$\alpha = D_2(S_{ce})$$

לפי הגדרת D_2, ביחס ל־D_1:

$$p_{ce} = 2 \cdot (1-p) \cdot \alpha$$

כמו כן:

$$D_2(S_{ec}) = p - \alpha$$

כי:

$$p = D_2(S_{ce}) + D_2(S_{ec})$$

אלגוריתם 3.6 אלגוריתם לשיפור הדיוק

נניח כי $\mathcal{A}$ הוא אלגוריתם ללמידה חלשה, ו־p תהיה הסתברות השגיאה של $\mathcal{A}$.
כמו כן, נניח כי נתון לנו Oracle מעל מרחב הדגימות, שנותן לנו דגימה לפי ההתפלגות D.

1. נסמן: $D_1 = D$. נמצא השערה $h_1 \in H$ ביחס ל־D_1 (לפי $\mathcal{A}$). השגיאה חסומה על ידי $\frac{1}{2} - \gamma = p$.

2. נגדיר:

$$\begin{aligned} S_c &= \{x \mid h_1(x) = c^\star(x)\} \\ S_e &= \{x \mid h_1(x) \neq c^\star(x)\} \end{aligned}$$

3. נגדיר התפלגות חדשה D_2 כך ש־$D_2(S_c) = \frac{1}{2} = D_2(S_e)$, וחוץ מזה D_2 תהיה בדיוק כמו D_1. כלומר:

$$D_2 = \begin{cases} \frac{0.5}{1-p} \cdot D_1(x) & x \in S_c \\ \frac{0.5}{p} \cdot D_1(x) & x \in S_e \end{cases}$$

איך נייצור את D_2 בפועל? נתון לנו Oracle שמייצר D_1. אנחנו רוצים להביא ל־$1/2$את כמות הפעמים ש־h_1 צודקת ואת כמות הפעמים ש־h_1 טועה.

לכן, באופן הזה נבנה Oracle שפועל לפי D_2, כי נחליט (בהתפלגות אחידה) אם הפעם אנחנו רוצים ש־h_1 תצדק או ש־h_1 תטעה.

4. נרוץ עם $\mathcal{A}$ עם ההתפלגות D_2 ונמצא h_2 (עם שגיאה $\frac{1}{2} - \gamma = p$).

5. נגדיר התפלגות D_3 שתשים משקל על $\{x \mid h_1(x) \neq h_2(x)\}$. כלומר:

$$D_3(x) = \begin{cases} \frac{D_1(x)}{z} & h_1(x) \neq h_2(x) \\ 0 & h_1(x) = h_2(x) \end{cases}$$

כאשר $z = \Pr[h_1(x) \neq h_2(x)]$ הוא פוקטור נרמול.

קל לבנות Oracle שפועל לפי D_3, באופן דומה לבניית ה־Oracle שפועל לפי D_2.

6. מוצאים עם $\mathcal{A}$ השערה h_3 ביחס ל־D_3 (עם שגיאה $\frac{1}{2} - \gamma = p$).

7. הפלט שלנו יהיה:

$$H(x) = \begin{cases} h_1(x) & h_1(x) = h_2(x) \\ h_3(x) & h_1(x) \neq h_2(x) \end{cases}$$

או במילים אחרות, $H = \mathrm{MAJ}(h_1, h_2, h_3)$.

לכן:

$$p_{ee} = 2 \cdot p \cdot (p - \alpha)$$

באופן דומה, נקבל כי:

$$D_2\left(S_{ec}\right) = \frac{1}{2} - (p - \alpha)$$

ולכן:

$$p_{ec} = 2 \cdot p \cdot \left(\frac{1}{2} - p + \alpha\right)$$

נציב ב־error:

$$\begin{aligned}
\text{error} &= p_{ee} + (p_{ce} + p_{ec}) \cdot p \\
&= 2 \cdot p \cdot (p - \alpha) + \left(2 \cdot p \cdot \left(\frac{1}{2} - p + \alpha\right) + 2 \cdot (1 - p) \cdot \alpha\right) \cdot p \\
&= 2p^2 - \cancel{2 \cdot p \cdot \alpha} + p^2 - 2p^3 + \cancel{2 \cdot p^2 \cdot \alpha} + \cancel{2 \cdot p \cdot \alpha} - \cancel{2 \cdot p^2 \cdot \alpha} \\
&= 3p^2 - 2p^3
\end{aligned}$$

למעשה, היינו רוצים להראות ש־γ הולך וגדל עם הזמן. כלומר, אם

$$p_0 = \frac{1}{2} - \gamma_0$$

אזי

$$\frac{1}{2} - \gamma_{t+1} = 3 \cdot \left(\frac{1}{2} - \gamma_t\right)^2 - 2 \cdot \left(\frac{1}{2} - \gamma_t\right)^3$$

ולכן:

$$\frac{1}{2} - \gamma_{t+1} \leq \frac{1}{2} - \gamma_t \cdot \left(\frac{3}{2} - \gamma_t\right)$$

דרך אחרת להסתכל על זה היא באמצעות:

$$p_{t+1} \leq 3 \cdot p_t^2 - 2 \cdot p_t^3$$

לכן, אם γ היא שגיאת ההשערה החלשה, ו־ε היא שגיאת ההשערה החזקה, עומק הרקורסיה שלנו יהיה $O\left(\log\log\frac{1}{\varepsilon} + \log\frac{1}{\gamma}\right)$. גודל הרקורסיה שלנו יהיה 3^{depth}.

6.3 אלגוריתם AdaBoost

אלגוריתם 4.6 הוא אלגוריתם AdaBoost (Adaptive Boost).

אלגוריתם 4.6 AdaBoost (Adaptive Boost)

קלט: נקודות $x_1, \ldots, x_m$ מסווגות $\langle x_i, y_i \rangle$.
נגדיר: D_t – ההתפלגות בשלב t. $D_t(i)$ – המשקל של הנקודה x_i בשלב t.

1. אתחול:

$$\forall i. D_1(i) = \frac{1}{m}$$

2. בשלב t:

(א) נבחר $h_t \in H$ שמביאה למינימום את השגיאה ε_t ביחס ל־D_t:

$$\varepsilon_t = \Pr_{D_t}[h_t(x) \neq c^\star(x)]$$

(ב) נגדיר את D_{t+1}. נניח כי $h_t(x) \in \{\pm 1\}$. אזי:

$$D_{t+1}(i) = \frac{D_t(i)}{z_t} \cdot \begin{cases} e^{-\alpha \cdot t} & h_t(x_i) = y_i \\ e^{+\alpha \cdot t} & h_t(x_i) \neq y_i \end{cases}$$

כאשר:

$$\alpha_t = \frac{1}{2} \cdot \ln \frac{1 - \varepsilon_t}{\varepsilon_t}$$

ו־z_t הוא גורם מנרמל.

3. אחרי T שלבים:

$$H(x) = \operatorname{sign}\left(\sum_{t=1}^{T} \alpha_t \cdot h_t(x)\right)$$

משפט

$$\begin{aligned}
\widehat{\text{error}}\left(H\right) &\leq \prod_{t=1}^{T}\sqrt{4\cdot\varepsilon_t\cdot\left(1-\varepsilon_t\right)} \\
&= \prod_{t=1}^{T}\sqrt{4\cdot\left(\frac{1}{2}-\gamma_t\right)\cdot\gamma_t} \qquad (6.3.1)\\
&= \prod_{t=1}^{T}\sqrt{1-4\cdot\gamma_t^2} \\
&= \prod_{t=1}^{T}\left(1-4\cdot\gamma_t^2\right)^{1/2} \\
&\leq \prod_{t=1}^{T}\left(e^{-4\cdot\gamma_t^2}\right)^{1/2} \qquad (6.3.2)\\
&= \prod_{t=1}^{T}e^{-2\cdot\gamma_t^2} \\
&= e^{-2\cdot\sum_{t=1}^{T}\gamma_t^2}
\end{aligned}$$

נסביר את המעברים בשלבים:

- מעבר 6.3.1 בוצע כי $\varepsilon_t = 1-\gamma_t$.
- מעבר 6.3.2 בוצע לפי הזהות $1+x\leq e^x$.

הוכחה שלב א': נוכיח כי:

$$D_{T+1}\left(i\right)=\frac{D_1\left(i\right)\cdot e^{-y_i\cdot f(x_i)}}{\prod_{t=1}^{T}z_t}$$

כאשר:

$$f\left(x\right)=\sum_{t=1}^{T}\alpha_t\cdot h_t\left(x\right)$$

מהרקורסיה נקבל:

$$D_{t+1}\left(i\right)=\frac{D_t\left(i\right)}{z_t}\cdot e^{-y_i\cdot\alpha_t\cdot h_t(x_i)}$$

לכן:

$$\begin{aligned}
D_{T+1}\left(i\right) &= \frac{D_1\left(i\right)}{\prod_{t=1}^{T}z_t}\cdot e^{-y_i\cdot\sum_{t=1}^{T}\alpha_t\cdot h_t(x_i)} \\
&= \frac{D_1\left(i\right)}{\prod_{t=1}^{T}x_t}\cdot e^{-y_i\cdot f(x_i)} \\
&= \frac{e^{-y_i\cdot f(x_i)}}{m\cdot\prod_{t=1}^{T}z_t} \qquad (6.3.3)
\end{aligned}$$

כאשר מעבר 6.3.3 בוצע על סמך העובדה:

$$D_1(i) = \frac{1}{m}$$

שלב ב': נוכיח כי:

$$\widehat{\text{error}}(H) \leq \prod_{t=1}^{T} z_i$$

נחזור להגדרה של $\widehat{\text{error}}(H)$:

$$\begin{aligned}
\widehat{\text{error}}(H) &= \frac{1}{m} \cdot \sum_{i=t}^{m} I(y_i \cdot f(x_i) \leq 0) \\
&\leq \frac{1}{m} \cdot \sum_{i=1}^{m} e^{-y_i \cdot f(x_i)} \\
&= \frac{1}{\cancel{m}} \cdot \sum_{i=1}^{m} \cancel{m} \cdot D_{T+1}(i) \cdot \prod_{t=1}^{T} z_t \qquad (6.3.4) \\
&= \prod_{t=1}^{T} z_t \cdot \sum_{i=1}^{m} D_{T+1}(i) \\
&= \prod_{t=1}^{T} z_t \qquad (6.3.5)
\end{aligned}$$

כאשר המעבר 6.3.4 בוצע על סמך הטענה מחלק א', והמעבר 6.3.5 בוצע על סמך העובדה:

$$\sum_{i=1}^{m} D_t(i) = 1$$

כי D_t מתאר התפלגות (עם נרמול).
שלב ג': נוכיח כי:

$$z_t = 2 \cdot \sqrt{\varepsilon_t \cdot (1 - \varepsilon_t)}$$

z_t הוא הגורם המנרמל של D_t. כלומר, לפי ההגדרה:

$$\begin{aligned}
z_t &= \sum_{i=1}^{m} D_t(i) \cdot e^{-y_i \cdot \alpha_t \cdot h_t(x_i)} \\
&= \sum_{\substack{i=1 \\ h_t(x_i) = y_i}}^{m} D_t(i) \cdot e^{-\alpha_t} + \sum_{\substack{i=1 \\ h_t(x_i) \neq y_i}}^{m} D_t(i) \cdot e^{\alpha_t} \\
&= (1 - \varepsilon_t) \cdot e^{-\alpha_t} + \varepsilon_t \cdot e^{\alpha_t}
\end{aligned}$$

נחשוב על z_t כפונקציה של α_t, ונגזור:

$$\begin{aligned}
\frac{\partial}{\partial \alpha_t} z_t = -(1-\varepsilon_t)\cdot e^{-\alpha_t} + \varepsilon_t \cdot e^{\alpha_t} &= 0 \\
-e^{-\alpha_t} + \varepsilon_t \cdot e^{-\alpha_t} + \varepsilon_t \cdot e^{\alpha_t} &= 0 \\
-1 + \varepsilon_t + \varepsilon_t \cdot e^{2\cdot\alpha_t} &= 0 \\
\varepsilon_t \cdot e^{2\cdot\alpha_t} &= 1-\varepsilon_t \\
e^{2\cdot\alpha_t} &= \frac{1-\varepsilon_t}{\varepsilon_t} \\
2\cdot\alpha_t &= \ln\frac{1-\varepsilon_t}{\varepsilon_t} \\
\alpha_t &= \frac{1}{2}\cdot\ln\frac{1-\varepsilon_t}{\varepsilon_t}
\end{aligned}$$

אם נציב את α_t שקיבלנו, נקבל:

$$\begin{aligned}
z_t &= (1-\varepsilon_t)\cdot e^{-\alpha_t} + \varepsilon_t \cdot e^{\alpha_t} \\
&= (1-\varepsilon_t)\cdot e^{-\frac{1}{2}\cdot\ln\frac{1-\varepsilon_t}{\varepsilon_t}} + \varepsilon_t \cdot e^{\frac{1}{2}\cdot\ln\frac{1-\varepsilon_t}{\varepsilon_t}} \\
&= (1-\varepsilon_t)\cdot\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)^{-1/2} + \varepsilon_t\cdot\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)^{1/2} \\
&= (1-\varepsilon_t)\cdot\left(\frac{\varepsilon_t}{1-\varepsilon_t}\right)^{1/2} + \varepsilon_t\cdot\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)^{1/2} \\
&= (1-\varepsilon_t)\cdot\frac{\varepsilon_t^{1/2}}{(1-\varepsilon_t)^{1/2}} + \varepsilon_t\cdot\frac{(1-\varepsilon_t)^{1/2}}{\varepsilon_t^{1/2}} \\
&= \varepsilon_t^{1/2}\cdot(1-\varepsilon_t)^{1/2} + \varepsilon_t^{1/2}\cdot(1-\varepsilon_t)^{1/2} \\
&= \sqrt{\varepsilon_t\cdot(1-\varepsilon_t)} + \sqrt{\varepsilon_t\cdot(1-\varepsilon_t)} \\
&= 2\cdot\sqrt{\varepsilon_t\cdot(1-\varepsilon_t)}
\end{aligned}$$

מהחיבור של שלושת השלבים הוכחנו את המשפט. □

פרק 7

Nearest Neighbor[1]

7.1 הקדמה

נתון מרחב מדגם X. המטרה שלנו היא לסווג נקודה $x \in X$. כמו כן, נתון מגדם $S = \{\langle x_i, b_i \rangle\}$.
אנחנו נמצא נקודה $y \in S$ כך ש־$\|x - y\| = \min$, ואם $\langle y, b \rangle$ אז ההשערה עבור x היא b.

7.2 שיטות כלליות

יש לנו כמה שיטות כלליות:

1. הנקודה הקרובה ביותר.
2. ממוצע של k נקודות קרובות ביותר.
3. משקל על הנקודות לפי מרחק.

7.3 מודלים ל־Nearest Neighbor

כמו כן, ישנם מספר מודלים לייצור הנקודות:

- נניח שישנה התפלגות D מעל X. תחילה דוגמים מ־X לפי D.
  בהינתן $x \in X$, ישנה פונקציה $p(x)$ כך ש:

$$\Pr[b = 1 \mid x] = p(x)$$

- נניח שישנן שתי התפלגויות D_0 ו־D_1, ופרמטר $\alpha \in [0, 1]$.
  תחילה נדגום את b לפי α (כלומר $\Pr[b = 1] = \alpha$), ואז נדגום את $x \in X$ לפי D_b.
  מחזירים את $\langle x, b \rangle$.

$$p(x) = \frac{\alpha \cdot D_1(x)}{\alpha \cdot D_1(x) + (1 - \alpha) \cdot D_0(x)}$$

[1] שיעור שהתקיים בתאריך 1.12.2012.

- נניח שקיימת התפלגות D מעל $X \times \{0,1\}$.
  כשנדגום את D נקבל $\langle x, b \rangle$ ישירות.

$$p(x) = \frac{D(\langle x,1\rangle)}{D(\langle x,1\rangle) + D(\langle x,0\rangle)}$$

אנחנו רוצים להביא למינימום את מספר השגיאות.

7.3.1 0-1 Loss

$$h(x) = \begin{cases} 1 & p(x) \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

7.3.2 Bayes Risk

ההפסד של המסווג האופטימלי (בהינתן ההתפלגות):

$$r(x) = \min\{p(x), 1-p(x)\}$$

ואז:

$$R^\star = \mathbf{E}_x[\text{Loss}(h,x)] = \int_x r(x) \cdot D(x)\, dx$$

נסמן: $P = \mathbf{E}_x[\text{Loss}(\text{NN},x)]$ (Nearest Neighbor – NN). נוכיח כי:

$$R^\star \leq P \leq 2 \cdot R^\star \cdot (1 - R^\star)$$

אם $R^\star \sim \varepsilon$, אזי צ"ל $P \leq 2 \cdot \varepsilon \cdot (1-\varepsilon)$. אם $R^\star \sim \frac{1}{2} - \varepsilon$, אזי $P \leq 2 \cdot \left(\frac{1}{2} - \varepsilon\right) \cdot \left(\frac{1}{2} + \varepsilon\right) = \frac{1}{2} - 2 \cdot \varepsilon^2$.

7.3.2.1 מקרה פשוט

רוצים לסווג את $x \in X$. במדגם יש נקודה אחת $\langle x_i, b_i \rangle$ כך ש־$x_i \equiv x$.
נחשב:

$$\begin{aligned} \Pr[b_i \neq b] &= 2 \cdot p(x) \cdot (1 - p(x)) \\ &= 2 \cdot r(x) \cdot (1 - r(x)) \end{aligned}$$

נסביר את המעבר הראשון: נייצר שני מקרים: בראשון $b = 0$ ו־$b_i = 1$, ובשני $b = 1$ ו־$b_i = 0$.
נמצא את התוחלת למקרה הזה:

$$\begin{aligned} \mathbf{E}[b_i \neq b] &= \int_x 2 \cdot p(x) \cdot (1 - p(x)) \cdot D(x)\, dx \\ &= \int_x 2 \cdot r(x) \cdot (1 - r(x)) \cdot D(x)\, dx \\ &= 2 \cdot \int_x r(x) \cdot D(x)\, dx - 2 \cdot \int_x r^2(x) \cdot D(x)\, dx \\ &\leq 2 \cdot R^\star - 2 \cdot (R^\star)^2 = 2 \cdot R^\star \cdot (1 - R^\star) \end{aligned}$$

7.3.2.2 המקרה הכללי

נניח:

- השכן הקרוב ביותר מתכנס ל־x כאשר $m \to \infty$, ואז:

$$\mathbf{E}_x \mathbf{E}_{x_1,\ldots,x_m} \min_i \|x_i - x\| \underset{m\to\infty}{\to} 0$$

- הסיווג בסביבה של x דומה לזה של x:

$$|p(x) - p(z)| \leq \alpha \cdot \|x - z\|$$

נוסיף עוד שתי הנחות מפשטות:

- $D(x)$ לא אפס על כל X, ואין נקודת מסה.
- $p(x)$ פונקציה רציפה.

שתי ההנחות האחרונות אינן הכרחיות, אך מפשטות מאוד את המודל.
בהינתן $x_1, \ldots, x_m$, נגדיר את $\mathrm{NN}_1^m(x)$ להיות השכן הקרוב ביותר ל־x:

$$\mathrm{NN}_1^m(x) = \arg\min_{x_i} \|x_i - x\|$$

משפט עבור כל x:

$$\mathrm{NN}_1^m(x) \underset{m\to\infty}{\to} x$$

הוכחה לכל $\varepsilon > 0$, נגדיר: $B_\varepsilon(x)$ – כדור ברדיוס ε שמרכזו x. אזי:

$$\Pr[x_i \in B_\varepsilon(x)] = q > 0$$

אם נדגום m פעמים:

$$\Pr[x_1, \ldots, x_m \notin B_\varepsilon(x)] = (1-q)^m \underset{m\to\infty}{\to} 0$$

וזה נכון לכל $\varepsilon > 0$. □
למעשה, הוכחנו כי:

$$\forall \varepsilon > 0, \delta > 0. \exists m_D(\varepsilon, \delta) \,.\, \Pr[\mathrm{NN}_1^m(x) \notin B_\varepsilon(x)] \leq \delta$$

משפט

$$r(\mathrm{NN}_1^m(x)) \underset{m\to\infty}{\to} r(x)$$

הוכחה נובעת מרציפות $p(x)$. □

נגדיר: e_x^m – המאורע ש־$\mathrm{NN}_1^m(x)$ עשה שגיאה עם מדגם בגודל m על הנקודה x. אזי:

$$\Pr[e_x^m] \underset{m\to\infty}{\to} r(x)\cdot(1-r(x))$$

לכן:

$$\begin{aligned}\lim_{m\to\infty}\int \Pr[e_x^m]\cdot D(x)\,dx &= \int \lim_{m\to\infty}\Pr[e_x^m]\cdot D(x)\,dx \\ &= \int r(x)\cdot(1-r(x))\cdot D(x)\,dx\end{aligned}$$

7.4 שיטת k שכנים קרובים (k-NN)

נגדיר:

$$\mathrm{NN}_k^m = \left\{ \begin{array}{c} k\text{nearest} \\ \text{neighbors of } m \end{array} \right\}$$

אזי $\hat{p}(x) = \frac{\ell}{k}$ אם ℓ מתוך k השכנים הקרובים חיוביים. נבחר:

1.
$$\frac{k}{m} \to 0$$

2.
$$k \to \infty$$

למשל: $k = \sqrt{m}$ ו־$k = \log m$.
נרצה להראות:

1. כל k השכנים הקרובים ביותר מתכנסים ל־x:

$$\forall z \in \mathrm{NN}_k^m(x).\, z \to x$$

2.
$$\hat{p}(x) \to p(x)$$

7.4.1 מקרה פשוט

כל k הקרובים ביותר זהים ל־x (כל השאר שונים מ־x).
יש לנו k מטבעות עם הסתברות $\hat{p}(x)$:

$$\Pr[|\hat{p}(x) - p(x)| \geq \lambda] \leq 2\cdot e^{-\lambda^2\cdot k}\cdot \hat{p}(x)$$

לכל $\lambda > 0$, ומכיוון ש־$k \to \infty$, מתקיים $\hat{p}(x) \to p(x)$.

7.4.2 מקרה כללי

התכנסות ל־x.
כמו קודם, נגדיר את $B_\varepsilon(x)$. אזי:

$$\Pr[x_i \in B_\varepsilon(x)] = q > 0$$

מספר הנקודות z בתוחלת שיפלו בתוך $B_\varepsilon(x)$ הוא $q \cdot m$. לכן:

$$\begin{aligned}\Pr[\text{at most } k-1 x_i\text{'s in } B_\varepsilon(x)] &= \Pr\left[\left|\frac{z}{m} - q\right| \geq q - \frac{k-1}{m}\right] \\ &\leq 2 \cdot e^{-\left(q-\frac{k-1}{m}\right)^2 \cdot m} \underset{m\to\infty}{\to} 0\end{aligned}$$

7.5 מדידת המרחק

המרחק תלוי בגודל של המימדים, או ביחידות שלהם (למשל ק"ג או גרמים, שניות או דקות).
רצוי להביא את המדידות לסקלה אחידה. איך עושים את זה? למשל:

- נניח שהמדידות התקבלו בהתפלגות נורמלית:

$$\frac{v_i - \mu_i}{\sigma_i}$$

- נניח שהמדידות התקבלו בהתפלגות אחידה:

$$\frac{v_i}{\max_i - \min_i}$$

אנחנו צריכים לבחור מרחק, למשל:

$$L_2(x,y) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$$

או לפי כל נורמה אחרת שנבחר.

7.6 Locality Sensitive Hashing

השאלה שתעניין אותנו: בהינתן x ו־R, צריך למצוא $y \in S$ כך ש־$\|x-y\| \leq R$ או להגיד שאין כזה, כלומר:

$$\forall y \in S. \|x - y\| > R$$

הגדרה תהי משפחה H. אם לכל $p, q \in \mathbb{R}^d$:

1. אם $\|p-q\| \leq R$ אזי $p_1 \leq \Pr_H [h(p) = h(q)]$.
2. אם $\|p-q\| > R$ אזי $\Pr_H [h(p) = h(q)] \geq p_2$.
3. $p_1 < p_2$.

אזי נאמר כי H היא משפחה טובה. זו ההגדרה ל־*Locality Sensitive Hashing*.

דוגמה $X = \{0,1\}^d$. נגדיר:

$$h_i(x_1, \ldots, x_d) = x_i$$

אזי:

$$\Pr_H [h(x) = h(y)] = 1 - \frac{d_{\mathbf{H}}(x,y)}{d}$$

כאשר $d_{\mathbf{H}}$ הוא מספר המינגטון:

$$d_{\mathbf{H}}(x,y) = \#\text{ corr. diff}$$

7.6.1 שלב א' – Amplification

נגדיר:

$$g(x) = \langle h_1(x), \ldots, h_k(x) \rangle$$

כאשר h_i נבחרים באקראי מתוך H.
עבור g:

1. אם $\|p-q\| \leq R$, אזי:

$$\Pr [g(p) = g(q)] \geq p_1^k$$

2. אם $\|p-q\| > R$, אזי:

$$\Pr [g(p) = g(q)] \leq p_2^k$$

נבחר את k כך ש־$p_2^k = \frac{1}{n}$. אזי:

$$k = \frac{\ln n}{\ln 1/p_2}$$

לכן:

$$\begin{aligned} p_1^k &= p_1^{\frac{\ln n}{\ln 1/p_2}} \\ &= n^{-\frac{\ln 1/p_1}{\ln 1/p_2}} \\ &= n^{-\rho} \end{aligned}$$

כאשר:

$$\rho = \frac{\ln 1/p_1}{\ln 1/p_2}$$

7.6.2 שלב ב'

נבחר $g_1, \ldots, g_L$, ונטען שההסתברות שקיים i כך ש־$g_i(p) = g_i(q)$ כאשר $\|p - q\| \leq R$ גדולה.

ההסתברות עבור g יחיד היא $1 - n^{-\rho}$. לכן, ההסתברות עבור $g_1, \ldots, g_L$ היא:

$$\left(1 - n^{-\rho}\right)^L \approx e^{-n^\rho \cdot L} = \delta$$

לכן:

$$L = n^\rho \cdot \ln \frac{1}{\delta}$$

7.6.3 האלגוריתם

בוחרים L פונקציות $g_1, \ldots, g_L$ כך ש־$g_i(x) = \langle h_{i,1}(x), \ldots, h_{i,k}(x) \rangle$ כאשר $h_{i,j} \in H$ אקראיים.

עבור כך g_i: לכל x מחשבים את $g_i(x)$, ומכניסים לטבלת Hash. יש לנו L טבלאות. בסך הכל, נתפוס $O(L \cdot n)$ מקום.

חיפוש: בהינתן q, לכל i נחשב את $g_i(q)$, ונבדוק אם יש $g_i(p) = g_i(q)$ בטבלה ה־i. זמן החישוב יהיה:

$$\underbrace{\frac{1}{n} \cdot n \cdot L}_{\substack{\text{Bad} \\ \text{points}}} + \underbrace{L \cdot \begin{pmatrix} \text{Hash} \\ \text{lookup} \end{pmatrix} \cdot \begin{pmatrix} \# \text{ Nearesst} \\ \text{points} \end{pmatrix}}_{\text{Good points}}$$

איך נעשה רדוקציות כאלה? למשל אם $X = \{0, \ldots, S-1\}^d$. אזי כל $x \in X$ יהיה בייצוג אונרי[2].

במקרה הזה, L_1 יהיה $d_{\mathbf{H}}$ (המינגטון) על $\{0,1\}^{d \cdot \log S}$, כאשר:

$$L_1(x, y) = \sum_{i=1}^{d} (x_i - y_i)$$

דוגמה מרחק L_1 מעל $[0,1]^d$. נניח כי $R \ll 1$. נבחר באקראי $s_1, \ldots, s_d \in [0,1]$. אזי:

$$h_{s_1,\ldots,s_d}(x) = \langle \operatorname{sign}(x_1 - s_1), \ldots, \operatorname{sign}(x_d - s_d) \rangle$$

לכן:

$$\Pr_s \left[\operatorname{sign}(x_i - s_i) \neq \operatorname{sign}(y_i - s_i)\right] = |x - y|$$

כאשר $x, y \in [0,1]^d$.

[2] למשל, 3 ייוצג כך: $\underbrace{00\cdots0}_{S-3\text{times}}\ \underbrace{111}_{3\text{times}}$.

$$\begin{aligned}\Pr\left[h\left(x\right)=h\left(y\right)\right] &= \prod_{i=1}^{d}\left(1-\left|x_i-y_i\right|\right)\\ &= 1-\prod_{i=1}^{d}\left|x_i-y_i\right|\\ &= 1-L_1\left(x,y\right)\end{aligned}$$

פרק 8

VC Dimension[1]

8.1 מודל PAC (חזרה)

במודל ה־PAC נתונה לנו D, ההתפלגות מעל הקלטים X. ההנחות שלנו לגביה:

1. D אינה משתנה (עם התקדמות הלמידה).
2. D אינה ידועה (ללומד).
3. הדוגמאות נדגמות מ־D באופן i.i.d, גם עבור הלמידה וגם עבור הבדיקה.

יש לנו פונקציית מטרה $c^\star \in C$. הדוגמאות הן מהצורה $\langle x, c^\star(x)\rangle$ כאשר $x \in X$.
המטרה: למצוא $h \in H$ כך שהשגיאה, המוגדרת על ידי:

$$\text{error}(h) = D\left[c^\star(x) \neq h(x)\right]$$

תהיה קטנה ככל הניתן.
יש לנו שני פרמטרים: ε – פרמטר הדיוק, ו־δ – פרמטר הבטחון. מאורע של הצלחה מוגדר כך:

$$\text{error}(h) \leq \varepsilon$$

פרמטר הבטחון נותן לנו:

$$\Pr[\text{success}] \geq 1 - \delta$$

אלגוריתם PAC מקבל את ε ו־δ, ולפעמים יש לו גישה לאורקל שדוגם את X לפי D:

$$\mathbf{EX}(D, c^\star) \mapsto \langle x, c^\star(x)\rangle$$

אלגוריתם PAC $\mathcal{A}$ לומד משפחה C (קבוצת פונקציות מטרה) אם לכל $c^\star \in C$ ולכל התפלגות D (מעל X), $\mathcal{A}$ מוצא השערה $h \in H$ כך ש:

$$\Pr[\text{error}(h) \leq \varepsilon] \geq 1 - \delta$$

נאמר כי האלגוריתם $\mathcal{A}$ יעיל אם הוא רץ בזמן פולימומי ב־$1/\varepsilon$, $\ln 1/\delta$ וקידוד $c^\star$.

[1] שיעור שהתקיים בתאריך 16.12.2012.

8.2 מימד VC

8.2.1 מוטיבציה

נתעניין במקרה שבו $C = H$. אנחנו נרצה לראות כמה דוגמאות m נחוצות כדי שהשערה עקבית (שגיאה 0 על המדגם) תלמד PAC.
אם C מחלקה סופית, ראינו כי:

$$m \geq \frac{1}{\varepsilon} \cdot \ln \frac{|C|}{\delta}$$

מה קורה אם C אינסופית? החסם הזה לא אומר לנו כלום. האם זה אומר שאנחנו לא נצליח ללמוד מחלקות אינסופיות?
ראינו כבר את הדוגמה של למידת מלבנים מקבילים לצירים, שבה:

$$m = O\left(\frac{1}{\varepsilon} \ln \frac{1}{\delta}\right)$$

אבל בהוכחה הזו הסתמכנו על תכונות של המלבנים.
היינו רוצים למצוא דרך גנרית יותר כדי להבין מהו m.

8.2.2 הגדרות

נניח כי C מחלקת השערות מעל X.
ניתן לחשוב על $c \in C$ בשתי דרכים:

1. $c(x) \in \{0,1\}$.
2. $\{x \mid c(x) = 1\}$.

עבור C מעל X ו־$S \subseteq X$ סופית, נגדיר:

$$\Pi_C(S) = \{c \cap S \mid c \in C\}$$

או באופן שקול, אם $S = \{x_1, \ldots, x_m\}$:

$$\Pi_C(S) = \{\langle c(x_1), \ldots, c(x_m)\rangle \mid c \in C\}$$

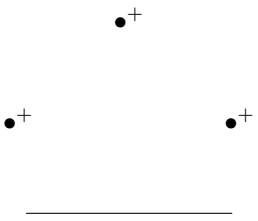
אנחנו נתעניין בגודל שצ $\Pi_C(S)$. נקרא ל־$\Pi_C(S)$ הטלה של C על קבוצה S.
נתעניין בשאלה הבאה: האם C משרה את כל הפונקציות מעל S? כלומר, האם $|\Pi_C(S)| = 2^m$?

הגדרה מחלקה C מנתצת את S אם $|\Pi_C(S)| = 2^m$ כאשר $m = |S|$.

הגדרה נגדיר את מימד VC:

$$\mathrm{VCdim}(C) = \max\left\{d \mid \exists S.\, |S| = d \wedge |\Pi_C(S)| = 2^d\right\}$$

אם התנאי מתקיים לכל d, אזי $\mathrm{VCdim}(C) = \infty$.



איור 8.1: סיווג אפשרי של 3 נקודות שאינן על אותו הישר, עם המפריד הלינארי שלהן

8.2.3 דוגמאות

8.2.3.1 סיפא על קו

נגדיר את C_1 להיות מחלקת הסיפאות על הקווים, $X = [0,1]$, ונגדיר:

$$c_\alpha(x) = \begin{cases} 0 & x < \alpha \\ 1 & x \geq \alpha \end{cases}$$

נשים לב כי C_1 אינסופית (אפילו לא בת־מניה). נראה כי $\text{VCdim}(C_1) = 1$.

שלב ראשון נוכיח כי $\text{VCdim}(C_1) \geq 1$.
נבחר $S = \{1/2\}$. אזי:

$$\begin{aligned} c_{\frac{3}{4}}\left(\frac{1}{2}\right) &= 0 \\ c_{\frac{1}{4}}\left(\frac{1}{2}\right) &= 1 \end{aligned}$$

לכן, $|\Pi_C(\{1/2\})| = 2$. לכן, $\text{VCdim}(C_1) \leq 1$.

שלב שני נראה כי $\text{VCdim}(C_1) < 2$.
יהיו $x, y \in [0,1]$. נניח כי $y \geq x$. אזי לא קיימת $c \in C_1$ כך ש־$c(x) = 1$ ו־$c(y) = 0$.
לכן, $\text{VCdim}(C_1) < 2$.
לכן, $\text{VCdim}(C_1) = 1$.

8.2.3.2 מפריד לינארי במישור

נגדיר את C_2 להיות מחלקת המפרידים הלינארים במישור, $X = \mathbb{R}^2$.
אם $x = (x_1, x_2) \in X$, עבור $w = (\alpha_1, \alpha_2, \theta)$ נגדיר:

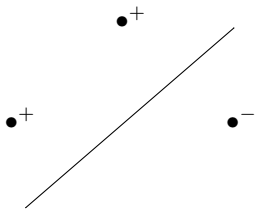
$$c_w(x) = 1 \Leftrightarrow \sum_{i=1}^{2} \alpha_i \cdot x_i \geq \theta$$

נראה כי $\text{VCdim}(C_2) = 3$.

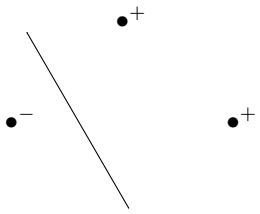
1. נראה כי $\text{VCdim}(C_2) \geq 3$.

   נבחר שלוש נקודות שאינן על אותו הישר. כפי שניתן לראות באיורים 8.1, 8.2, 8.3, 8.4, 8.5, 8.6, 8.7 ו־8.8, לכל סיווג אפשרי של שלוש הנקודות נוכל למצוא מפריד לינארי.

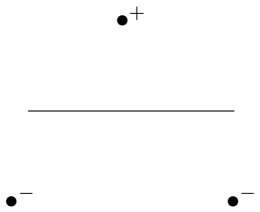
   לכן, $\text{VCdim}(C_2) \geq 3$.



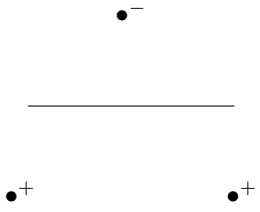
איור 8.2: סיווג אפשרי של 3 נקודות שאינן על אותו הישר, עם המפריד הלינארי שלהן



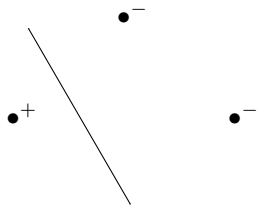
איור 8.3: סיווג אפשרי של 3 נקודות שאינן על אותו הישר, עם המפריד הלינארי שלהן



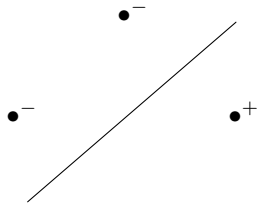
איור 8.4: סיווג אפשרי של 3 נקודות שאינן על אותו הישר, עם המפריד הלינארי שלהן



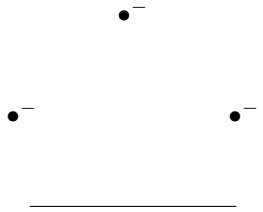
איור 8.5: סיווג אפשרי של 3 נקודות שאינן על אותו הישר, עם המפריד הלינארי שלהן



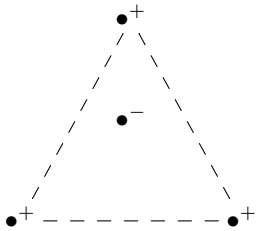
איור 8.6: סיווג אפשרי של 3 נקודות שאינן על אותו הישר, עם המפריד הלינארי שלהן



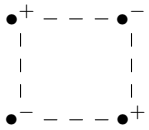
איור 8.7: סיווג אפשרי של 3 נקודות שאינן על אותו הישר, עם המפריד הלינארי שלהן



איור 8.8: סיווג אפשרי של 3 נקודות שאינן על אותו הישר, עם המפריד הלינארי שלהן



איור 8.9: סיווג בלתי ניתן לסיפוק של 4 נקודות במישור



איור 8.10: סיווג בלתי ניתן לסיפוק של 4 נקודות במישור

2. נראה כי $\text{VCdim}(C_2) < 4$.

בהינתן קבוצה כלשהי של 4 נקודות במישור, אנחנו נמצאים באחד המקרים שמתוארים באיורים 8.9, 8.10 או 8.11. כל אחד מהאיורים מראה סיווג בלתי ניתן לסיפוק על ידי C_2.

לכן, $\text{VCdim}(C_2) < 4$.

לכן, $\text{VCdim}(C_2) = 3$.

נוכיח בהמשך כי לכל מימד $d \geq 2$, מפריד לינארי במימד d מקיים $\text{VCdim}\left(C^d\right) = d+1$.[2]

8.2.3.3 מלבנים מקבילים לצירים

נגדיר את C_3 להיות המלבנים המקבילים לצירים. נראה כי $\text{VCdim}(C_3) = 4$.

הערה לא ניתן לנתץ כל קבוצה בגודל 4. למשל, הקבוצה שמתוארת באיור 8.12 אינה ניתנת לניתוץ.

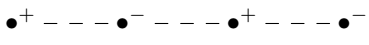
אבל אנחנו יכולים לבחור 4 נקודות כלשהן, ולא בהכרח את הנקודות האלה.

נבחר 4 נקודות כמו באיור 8.13. קל לראות שלכל אחת מההצבות לסיווגי הנקודות ניתן לייצר מלבן חוסם.לכן, $\text{VCdim}(C_3) \geq 4$.

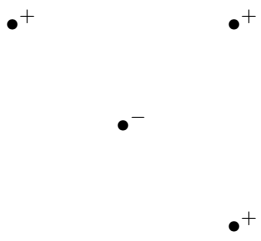
עבור 5 נקודות: לכל 5 נקודות, קיימת אחת שאינה קיצון באף מימד. נסמן אותה כשלילית, ואת שאר הנקודות כחיוביות. לא קיימת $c \in C_3$ שתנתץ את 5 הנקודות. לכן, $\text{VCdim}(C_3) < 5$.

לכן, $\text{VCdim}(C_3) = 4$.

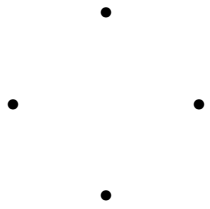
[2] ראה קטע 8.2.5.3.



איור 8.11: סיווג בלתי ניתן לסיפוק של 4 נקודות במישור



איור 8.12: דוגמה למקרה שלא ניתן לנתץ קבוצה בגודל 4



איור 8.13: דוגמה לנקודות שכל הצבה שלהן מנתצת את C_3

8.2.3.4 מספר סופי של אינטרוולים

נגדיר את C_4 להיות מספר סופי של אינטרוולים. כלומר: בהינתן $c \in C_4$, קיימת סדרת קטעים $\{[a_i, b_i]\}_{i=1}^{k_c}$ כך ש:

$$c(x) = 1 \Leftrightarrow x \in \bigcup_{i=1}^{k_c} [a_i, b_i]$$

נראה ש־$\mathrm{VCdim}(C_4) = \infty$.
לכל d נקודות ולכל הצבה אפשר לשים אינטרוול קטן סביב הנקודות החיוביות. כלומר, לכל הצבה יש פונקציה ב־C_4 שהיא עקבית איתה. לכן, $\mathrm{VCdim}(C_4) = \infty$.

8.2.3.5 פוליגון קונבקסי במישור

נגדיר את C_5 להיות פוליגון קונבקסי במישור. נראה ש־$\mathrm{VCdim}(C_5) = \infty$.
ניקח d נקודות על מעגל היחידה. לכל הצבה בנקודוץ, נוכל לחבר את הנקודות החיוביות ולקבל פוליגון קונבקסי. כל הנקודות השליליות יהיו מחוץ לפוליגון.
לכן, לכל d נקודות ולכל הצבה יש פוליגון קונבקסי שעקבי עם ההצבה. לכן, $\mathrm{VCdim}(C_5) = \infty$.

8.2.4 חסם תחתון על גודל הדגימה

משפט אם $\mathrm{VCdim}(C) = d + 1$ אזי:

$$m\left(\varepsilon, \delta = \frac{1}{2}, d+1\right) \geq \frac{d}{16 \cdot \varepsilon} = \Omega\left(\frac{d}{\varepsilon}\right)$$

הוכחה נבחר $T = \{z_0, \ldots, z_d\}$ כך ש־$|\Pi_C(T)| = 2^{d+1}$. נבנה את ההתפלגות D:

$$D(x) = \begin{cases} 1 - 8 \cdot \varepsilon & x = z_0 \\ \frac{8 \cdot \varepsilon}{d} & x = z_i, i \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

נבחר פונקציית מטרה:

$$c^\star(x) = \begin{cases} 0 \text{or } 1 \text{with probability } 1/2 & x \in T \\ \text{don't care} & x \notin T \end{cases}$$

קיימת $c \in C$ שזהה ל־$c^\star$ ב־T: נשתמש ב־c עבור $x \notin T$.
נשים לב כי $c^\star$ אינה פונקצייה אקראית, אלא הבנייה שלה אקראית.
נגדיר את הקבוצה $SEEN$ כקבוצת כל הנקודות ב־T שראינו (במהלך הלמידה). כמו כן, נגדיר: $UNSEES = T \setminus SEEN$.
לכל $z \in UNSEEN$, הסתברות השגיאה היא בדיוק $1/2$. בהנחה שראינו את z_0:

$$\begin{aligned} D[\text{error}] &\geq \frac{1}{2} \cdot |UNSEEN| \cdot \frac{8 \cdot \varepsilon}{d} \\ &= \frac{4 \cdot \varepsilon}{d} \cdot |UNSEEN| \end{aligned}$$

כמו כן:

$$\mathbf{E}[|SEEN|] \leq m \cdot 8 \cdot \varepsilon$$

נניח בשלילה כי $m < \frac{d}{16 \cdot \varepsilon}$. אזי $\mathbf{E}[|SEEN|] < \frac{d}{2}$. לכן:

$$\Pr\left[|SEEN| \leq \frac{d}{2}\right] \geq \frac{1}{2}$$

לכן, בהסתברות $1/2$, $|UNSEEN| \geq \frac{d}{2}$, ולכן $D[\text{error}] \geq 2 \cdot \varepsilon$, בסתירה להגדרת m. □

8.2.5 עוד דוגמאות

8.2.5.1 פונקציית Pairity

נגדיר את C_6 להיות פונקציות Pairity. כלומר, $X = \{0,1\}^n$ ועבור $S \subseteq \{1, \ldots, n\}$ מגדירים:

$$\chi_S(x) = \bigoplus_{i \in S} x_i$$

נראה ש־$\text{VCdim}(C_6) = n$.

1. נראה ש־$\text{VCdim}(C_6) \geq n$.

נבחר את n וקטורי הבסיס $\{e_i\}_{i=1}^n$, כאשר:

$$e_i = \left(\underbrace{0, \ldots, 0}_{i-1}, 1, \underbrace{0, \ldots, 0}_{n-i}\right)$$

לכל הצבה $b_1, \ldots, b_n$ עבור $e_1, \ldots, e_n$ נבחר $S = \{i \mid b_i = 1\}$, ואז:

$$\chi_S(e_j) = \begin{cases} 1 & j \in S \\ 0 & j \notin S \end{cases}$$

כאשר $j \in S \Leftrightarrow b_j = 1$.

לכן, $\text{VCdim}(C_6) \geq n$.

2. נראה ש־$\text{VCdim}(C_6) < n + 1$.

יש שתי דרכים להראות את זה:

(א) משיקולים קומבינטוריים: $|C_6| = 2^n$.

כמו כן, באופן כללי, $\text{VCdim}(C) \leq \log_2 |C|$. למה? אם ל־$C$ יש $\text{VCdim}(C) = d$, אז יש לפחות 2^d פונקציות שונות.

(ב) בהינתן וקטורים $x_1, \ldots, x_{n+1}$, קיים וקטור x_i כך ש־$x_i = \bigoplus_{j \in T} x_j$.

אם נבחר $b_j = 0$ לכל $j \in T$, אז בהכרח $\chi_S(x_i) = 0$. לכן, ההצבה $b_i = 1$ אינה אפשרית.

לכן, $\text{VCdim}(C_6) < n + 1$.

לכן, $\text{VCdim}(C_6) = n$.

8.2.5.2 OR של ליטרלים

נגדיר את C_7 להיות OR של ליטרלים כאשר $X = \{0,1\}^n$.
נראה ש־$\text{VCdim}(C_7) = n$.

1. נראה ש־$\text{VCdim}(C_7) \geq n$.

נבחר את וקטורי הבסיס $e_1, \ldots, e_n$. בהינתן הצבה $b_1, \ldots, b_n$, נגדיר את S כמו קודם: $S = \{i \mid b_i = 1\}$. ואז:

$$c_S(x) = \bigvee_{i \in S} x_i$$

2. נראה ש־$\text{VCdim}(C_7) \leq n$.

לצורך ההוכחה, נשתמש בטענת עזר:

טענה אם למחלקה C יש אלגוריתם Online שעושה לכל היתר d שגיאות, אזי $\text{VCdim}(C) \leq d$.

הוכחה נניח בשלילה כי $\text{VCdim}(C) \geq d + 1$. אזי קיימות $d + 1$ נקודות S שמקבלות את כל ההצבות. נסמן: $S = \{x_1, \ldots, x_{d+1}\}$.

נריץ את האלגוריתם על S, ותמיד נגיד לו שטעה. האלגוריתם יעשה $d + 1$ שגיאות, בסתירה להנחה. □

למה בהינתן 3 וקטורים (או יותר), ישנם 2 וקטורים במרחק 2 לפחות.

הוכחה יהיו $z_1, z_2, z_3 \in \{0,1\}^n$ 3 וקטורים שונים. נניח כי z_1 ו־z_2 נבדלים בביט i, ונניח כי z_2 ו־z_3 נבדלים בביט j.

אם $i = j$, אזי $z_1 = z_3$, בסתירה להנחה כי הווקטורים שונים. לכן, z_1 ו־z_3 נבדלים בביטים i ו־j. □

נראה כעת כי $\text{VCdim}(C_7) \leq n$: נראה שיש סדר כך שהאלגוריתם ELIM[3] מבצע לכל היותר n שגיאות.

בהינתן $n+1$ נקודות, נבחר את a ו־b להיות שני וקטורים במרחק 2 לפחות. נריץ את ELIM על הסדרה הבאה: a, b ואז כל האחרים. כאשר עוברים על a, מקבלים שגיאה ופוסלים לכל היותר n ליטרלים. כשעוברים על b פוסלים עוד 2 ליטרלים. כשעוברים על הווקטורים האחרים, עושים עוד $n-2$ איטרציות, שבכל אחת פוסלים לכל היותר ליטרל בודד.

לכן, קיבלנו לכל היותר n שגיאות. לכן, $\text{VCdim}(C_7) \leq n$.

לכן, $\text{VCdim}(C_7) = n$.

8.2.5.3 מפריד לינארי במימד n

נגדיר את C_8 להיות מפריד לינארי במימד n, $X = \mathbb{R}^n$ ($n \geq 2$).
עבור $w = (\alpha_1, \ldots, \alpha_n, \theta)$ נגדיר את $c_w \in C_8$:

$$c_w(x) = 1 \Leftrightarrow \sum_{i=1}^{n} \alpha_i \cdot x_i \geq \theta$$

נוכיח כי $\text{VCdim}(C_8) = n + 1$.

1. נראה כי $\text{VCdim}(C_8) \geq n + 1$.

נגדיר: $E = \left\{\vec{0}, \vec{e}_1, \ldots, \vec{e}_n\right\}$. בהינתן $b_0, \ldots, b_n$ נבחר:

$$\alpha_i = \begin{cases} +1 & b_i = 1 \\ -1 & b_i = 0 \end{cases}$$

$$\theta = \begin{cases} -\frac{1}{2} & b_0 = 1 \\ +\frac{1}{2} & b_0 = 0 \end{cases}$$

נסמן: $w = (\alpha_1, \ldots, \alpha_n, \theta)$. אזי:

$$c_w(e_i) = 1 \Leftrightarrow \alpha_i \geq \theta \Leftrightarrow b_i = 1$$

$$c_w\left(\vec{0}\right) = 1 \Leftrightarrow 0 \geq \theta \Leftrightarrow b_0 = 1$$

לכן, c_w עיקבית עם $b_0, \ldots, b_n$. □

2. נראה כי $\text{VCdim}(C_8) < n + 2$.

[3] ראה אלגוריתם 1.3.

הגדרה תת־קבוצה $A \subseteq \mathbb{R}^d$ היא קונבקסית אם:

$$\forall x_1, x_2 \in A. \forall \lambda \in [0,1] \,. x_1 + (1-\lambda)\, x_2 \in A$$

ה־*Convex Hull* של קבוצת נקודות S הוא הקבוצה הקונבקסית הקטנה ביותר שמכילה את S. ה־Convex Hull של S מסומן ב־$\mathrm{Conv}(S)$.

משפט *Radon* תהי $E \subseteq \mathbb{R}^d$ כך ש־$|E| = d+2$. אזי קיימת $S \subseteq E$ כך ש־$S \neq \emptyset$ וגם $E \setminus S \neq \emptyset$ וגם $\mathrm{Conv}(S) \cap \mathrm{Conv}(E \setminus S) \neq \emptyset$.

הוכחה נסמן: $E = \{x_0, \ldots, x_{d+1}\}$. אי קיימים $\alpha_0, \ldots, \alpha_{d-1}$ כך ש־$\sum_{i=0}^{d-1} \alpha_i \cdot x_i = \vec{0}$. מכיוון ש־$\mathbb{R}^d$ ממימד d ו־$|E| = d+2$, אזי אפשר להוסיף את האילוץ כי $\sum_{i=0}^{d+1} \alpha_i = 0$.
קיבלנו בסך הכל $d+1$ משוואות עבור $d+1$ נעלמים ($\alpha_0, \ldots, \alpha_{d-1}$), ולכן קיים פתרון שאינו טריוויאלי.
נניח בה"כ כי $\alpha_0, \ldots, \alpha_p$ חיוביים ו־$\alpha_{p+1}, \ldots, \alpha_{d+1}$ שליליים. נסמן:

$$\alpha = \sum_{i=0}^{p} \alpha_i > 0$$

עבור $i \leq p$ נגדיר:

$$\beta_i = \frac{\alpha_i}{\alpha} > 0$$

עבור $i \geq p+1$ נגדיר:

$$\gamma_i = \frac{-\alpha_i}{\alpha} > 0$$

נסמן:

$$y = \sum_{i=0}^{p} \beta_i \cdot \vec{x}_i = \sum_{i=p+1}^{d+1} \gamma_i \cdot \vec{x}_i$$

נסיק כי:

$$\sum_{i=0}^{p} \beta_i = 1 = \sum_{i=p+1}^{d+1} \gamma_i$$

נקבע: $S = \{x_0, \ldots, x_p\}$, $E \setminus S = \{x_{p+1}, \ldots, x_{d+1}\}$. אזי $y \in \mathrm{Conv}(S) \cap \mathrm{Conv}(E \setminus S)$. לכן, $\mathrm{Conv}(S) \cap \mathrm{Conv}(E \setminus S) \neq \emptyset$. □

נוכיח כעת כי $\mathrm{VCdim}(C_8) < n+2$: נניח בשלילה שלא. תהי E קבוצה של $n+2$ נקודות. לפי משפט *Radon*, קיימת $S \subseteq E$ כך ש־$\mathrm{Conv}(S) \cap \mathrm{Conv}(E \setminus S) \neq \emptyset$.
ניתן ל־S סיווג חיובי ול־E סיווג שלילי. אזי קיים c_w שמפריד בין S לבין $E \setminus S$. נסתכל על $x \in \mathrm{Conv}(S) \cap \mathrm{Conv}(E \setminus S)$:

- x חיובי כי S חיובית.
- x שלילי כי $E \setminus S$ שלילית.

סתירה! לכן, לא קיימות $n+2$ נקודות שניתנות לניתוץ. לכן, $\mathrm{VCdim}(C_8) < n+2$.
לכן, $\mathrm{VCdim}(C_8) = n+1$.

פרק 9

מימד VC (המשך)[1]

9.1 חזרה

במודל PAC, יש לנו התפלגות D על הקלטים X, ומחלקת השערות C. פונקציית המטרה שלנו היא $c^\star \in C$, ומניחים כי $C = H$.
האלגוריתם מוצא $h \in H$ כך שבהסתברות $1-\delta$, $\text{error}(h, c^\star) \leq \varepsilon$, כאשר δ ו־ε הם קלטים לאלגוריתם.
עבור $S \subseteq X$ ומחלקה C נגדיר:

$$\Pi_C(S) = \{c \cap S \mid c \in C\}$$

או באופן שקול:

$$\Pi_C(S) = \{\langle c(x_1), \ldots, c(x_m)\rangle \mid c \in C\}$$

כאשר $S = \{x_1, \ldots, x_m\}$. אזי $1 \leq |\Pi_C(S)| \leq 2^m$.
נרמא שמחלקה C מנתצת (*Shatters*) את S אם $|\Pi_C(S)| = 2^m$.
נגדיר את מימד VC:

$$\text{VCdim}(C) = \max\left\{d \mid \exists S.\, |S| = d \wedge |\Pi_C(S)| = 2^d\right\}$$

אם לכל d קיימתץ קבוצה S בגודל d ש־C מנתצת, אזי:

$$\text{VCdim}(C) = \infty$$

אם $d = \text{VCdim}(C)$, עבור מחלקה סופית:

$$|C| \geq |\Pi_C(S)|$$

לכן:

$$|C| \geq 2^d$$

ולכן:

$$\log_2(|C|) \geq d$$

[1] סיכום לשיעור שהתקיים בתאריך 23.12.2012.

9.2 חסמים לגודל הדגימה

בשיעור הקודם ראינו חסם תחתון לגודל הדגימה:

$$m\left(\varepsilon, \delta = \frac{1}{2}, d\right) \geq \Omega\left(\frac{d}{\varepsilon}\right)$$

היום נראה חסם עליון.
נדגום את S. מספר ההשערות ב־$\Pi_C(S)$ הוא סופי. לכן, יש לנו מחלקת השערות סופיץ. לכן:

$$m \leq \frac{1}{\varepsilon} \cdot \log \frac{|\Pi_C(S)|}{\delta}$$

הבעיה שלנו בטענה הזאת היא שקיבלנו את S, ורק אחר כך קבענו את החסם. החסם לא אמור להיות תלוי במדגם שלנו.

הגדרה נגדיר את קבוצת ההשערות הרעות:

$$B_\varepsilon(c^\star) = \{h \in H \mid \text{error}(h, c^\star) > \varepsilon\}$$

נרצה להראות שאם גודל המדגם m מספיק גדול, אזי $h \notin B_\varepsilon(c^\star)$ בהסתברות $1 - \delta$.

הגדרה קבוצת נקודות S היא ε-*Net* עבור $c^\star \in C$ והתפלגות D אם:

$$\forall h \in B_\varepsilon(c^\star). \exists x \in S. c^\star(x) \neq h(x)$$

נראה שבהסתברות $1 - \delta$, S היא ε-Net, ולכן זוהי למידת PAC.
נגדיר מדגם $S = S_1 \uplus S_2$ כאשר $|S_1| = m = |S_2|$. נגדיר את המאורע A: S_1 הוא לא ε-Net. נגדיר את המאורע B: ישנה פונקציה $h \in B_\varepsilon(c^\star)$ שהיא עקבית עם S_1, ול־h יש לפחות $\frac{\varepsilon \cdot m}{2}$ שגיאות על S_2.
ברור ש־$A \subseteq B$. כמו כן, $\Pr[B \mid A] \geq 1/2$. נחשב:

$$\begin{aligned} \Pr[B] &= \Pr[B \mid A] \cdot \Pr[A] \\ &\geq \frac{1}{2} \cdot \Pr[A] \\ 2 \cdot \Pr[B] &\geq \Pr[A] \end{aligned}$$

נחסום את $\Pr[B]$. נגדיר:

$$\begin{aligned} F &= \Pi_C(S_1 \cup S_2) \\ \text{ER}(h) &= \{x \in S_1 \cup S_2 \mid c^\star(x) \neq h(x)\} \end{aligned}$$

המאורע A אומר $\text{ER}(h) \cap S_1 = \emptyset$. המאורע B אומר $\text{ER}(h) \cap S_1 \neq \emptyset$ ו־$\text{ER}(h) \cap S_2 = \text{ER}(h)$ כאשר $|\text{ER}(h)| \geq \frac{m \cdot \varepsilon}{2}$.
יש לנו בעיה קומבינטורית: יש לנו $2 \cdot m$ כדורים, שמתוכם לפחות $\frac{m \cdot \varepsilon}{2}$ שחורים (השאר לבנים). נסמן ב־ℓ את מספר הכדורים השחורים.

מספר הקונפיגורציות לצביעת הכדורים: $\binom{2\cdot m}{\ell}$. מספר הקונפיגורציות בהן ℓ רק ב־S_2: $\binom{m}{\ell}$. לכן, ההסתברות שכל ℓ הכדורים יהיו ב־S_2:

$$\begin{aligned}\frac{\binom{m}{\ell}}{\binom{2\cdot m}{\ell}} &= \prod_{i=0}^{\ell-1}\frac{m-i}{2\cdot m-i} \\ &= \prod_{i=0}^{\ell-1}\left(\frac{1}{2}-\frac{i/2}{2\cdot m-i}\right)\leq\frac{1}{2^{\ell}}\end{aligned}$$

לכן:

$$\Pr\left[B\right]\leq|F|\cdot 2^{-\ell}\leq|F|\cdot 2^{-\frac{m\cdot\varepsilon}{2}}\leq\delta$$

לכן:

$$m=O\left(\frac{1}{\varepsilon}\cdot\ln\frac{|F|}{\delta}\right)$$

כאשר

$$|F|=\left|\Pi_C\left(\underbrace{S_1\cup S_2}_{2\cdot m\text{points}}\right)\right|$$

השלב הבא יהיה לחסום את גודל F כפונקציה של מימד ה־VC. נגדיר:

$$J\left(m,d\right)=J\left(m-1,d\right)+J\left(m-1,d-1\right)$$

כאשר:

$$\begin{aligned}\forall d &\quad . \quad J\left(0,d\right)=1\\ \forall m &\quad . \quad J\left(m,0\right)=1\end{aligned}$$

לכן, הנוסחה הסגורה של $J\left(m,d\right)$ היא:

$$J\left(m,d\right)=\sum_{i=0}^{d}\binom{m}{i}$$

טענה אם $\mathrm{VCdim}\left(C\right)=d$ ו־$|S|=m$, אזי:

$$\left|\Pi_C\left(S\right)\right|\leq J\left(m,d\right)$$

הוכחה נוכיח באינדוקציה של m ו־d (למעשה באינדוקציה על $m+d$).
בסיס: אם $m=1$ או $d=1$ אנחנו בסדר (מהגדרת J).
צעד האינדוקציה: נניח נכונות לכל m' ו־d' כך ש־$m'+d'<m+d$. כלומר, $\left|\Pi_C\left(S\right)\right|\leq J\left(m',d'\right)$ כאשר $|S|=m'$ ו־$\mathrm{VCdim}\left(C\right)=d'$.

נוכיח עבור m ו־d: נסמן:

$$\begin{aligned} S &= \{x_1, \ldots, x_m\} \\ C_S &= \Pi_C(S) \end{aligned}$$

נסתכל על $T = \{x_1, \ldots, x_{m-1}\}$. אזי:

$$\Pi_C(T) = \{y \mid y0 \in \Pi_C(S) \vee y1 \in \Pi_C(S)\}$$

נסמן:

$$C_\star = \{y \in \Pi_C(T) \mid y0 \in \Pi_C(S) \wedge y1 \in \Pi_C(S)\}$$

נראה כי $|C_S| = |C_T| + |C_\star|$. מהאינדוקיה:

$$|C_T| \leq J(m-1, d)$$

נראה ש־$|C_\star| \leq J(m-1, d-1)$. אם $C_\star$ מנתצת את $\{x_1, \ldots, x_i\}$, אזי C מנתצת את $\{x_1, \ldots, x_i, x_m\}$, כי כל הצבה ל־$\{x_1, \ldots, x_i\}$ ב־$C_\star$ יש לה שני המשכים ב־C עבור x_m. מכיוון ש־$\mathrm{VCdim}(C) = d$, $\mathrm{VCdim}(C_\star) \leq d-1$. לכן, מהנחת האינדוקציה:

$$|C_\star| \leq J(m-1, d-1)$$

בסך הכל:

$$\begin{aligned} |C_S| &= |C_T| + |C_\star| \\ &\leq J(m-1, d) + J(m-1, d-1) \\ &= J(m, d) \end{aligned}$$

□

למעשה, הראנו כי $|F| \leq J(2 \cdot m, d)$. נסתכל על $J(m, d)$:

$$J(m, d) \leq \begin{cases} 2^m & d \geq m \\ 2 \cdot m^d & d < m \end{cases}$$

מכך ש־$m = O\left(\frac{1}{\varepsilon} \cdot \ln \frac{|F|}{\delta}\right)$ נסיק כי:

$$M = O\left(\frac{1}{\varepsilon} \cdot \ln \frac{1}{\delta} + \frac{d}{\varepsilon} \cdot \ln \frac{d}{\varepsilon}\right)$$

9.3 סיבוכיות Radamacker

9.3.1 ממוצעי Radamacker

הגדרה עבור $S = \{x_1, \ldots, x_m\}$ ומחלקה H כאשר $H = \{h \colon X \to \{\pm 1\}\}$, נגדיר את ממוצעי *Radamacker*:

$$R_S(H) = \mathbf{E}_\sigma \left[\max_{h \in H} \frac{1}{m} \cdot \sum_{i=1}^{m} \sigma_i \cdot h(x_i)\right]$$

כאשר:

$$\forall i.\Pr\left[\sigma_i = 1\right] = \frac{1}{2} = \Pr\left[\sigma_i = -1\right]$$

באופן כללי, עבור $h \in H$ כלשהו, $\mathbf{E}\left[\sigma_i \cdot h\left(x_i\right)\right] = 0$. במקרה שלנו, $\sigma_i = \pm 1$ ו־$h\left(x_i\right) = \pm 1$, ולכן:

$$\sigma_i = h\left(x_i\right) \Leftrightarrow \sigma_i \cdot h\left(x_i\right) = 1$$

מה הקשר של זה למימד VC? אם H מנתצת את S:

$$\forall \sigma_i . \exists h \in H . \frac{1}{m} \cdot \sum_{i=1}^{m} \sigma_i \cdot h\left(x_i\right) = 1$$

לכן, $R_S\left(H\right) \leq 1$. אם H מכילה h וגם $\bar{h}$, אזי $R_S\left(H\right) \geq 0$.

דוגמה $H_1 = \{h_0, h_1\}$, $H_2 = \{h_0, h_2\}$ כאשר:

$$\begin{aligned} h_0 &= -1, \ldots, -1 \\ h_1 &= 1, \ldots, 1 \\ h_2 &= -1, \ldots, -1, 1 \end{aligned}$$

אזי:

$$\begin{aligned} R_S\left(H_1\right) &\approx \frac{\sqrt{m}}{m} \\ R_S\left(H_2\right) &= \frac{1}{m} \end{aligned}$$

נשים לב כי $\mathrm{VCdim}\left(H_1\right) = 1 = \mathrm{VCdim}\left(H_2\right)$, אבל $R_S\left(H_1\right) \neq R_S\left(H_2\right)$.

הגדרה

$$R_D\left(H\right) = \mathbf{E}_{S \sim D}\left[R_S\left(H\right)\right]$$

9.3.2 אי שוויון McDiarmid

יהיו $X_1, \ldots, X_m$ מ"מ ב"ת. תהי $\Phi\left(x_1, \ldots, x_m\right)$ פונקציה. נניח כי:

$$\forall i. \left|\Phi\left(x_1, \ldots, x_i, \ldots, x_m\right) - \Phi\left(x_1, \ldots, x_i', \ldots, x_m\right)\right| \leq c_i$$

אזי לכל $\varepsilon > 0$:

$$\Pr\left[\Phi\left(x\right) > \mathbf{E}\left[\Phi\left(x\right)\right] + \varepsilon\right] \leq \exp\left(\frac{-2 \cdot \varepsilon^2}{\sum_{i=1}^{m} c_i^2}\right)$$

9.3.3 סיבוכיות Radamacker

משפט בהסתברות $1-\delta$, לכל $h \in H$:

$$\begin{aligned}\text{error}_D(h) &\leq \text{error}_S(h) + R_D(H) + \sqrt{\frac{\ln 2/\delta}{m}} \\ &\leq \text{error}_S(h) + R_S(D) + 3 \cdot \sqrt{\frac{\ln 2/\delta}{m}}\end{aligned}$$

כאשר $|S| = 2 \cdot m$.

הוכחה נגדיר:

$$\text{MAXGAP}(S) = \max_{h \in H} |\text{error}_D(h) - \text{error}_S(h)|$$

צעד 1: נראה שבהסתברות $1-\delta$, $\text{MAXGAP}(S) \leq \varepsilon$.
במקרה שלנו:

$$c_i = \frac{1}{|S|} = \frac{1}{m}$$

לכן, בהסתברות $1-\delta$:

$$\text{MAXGAP}(S) \leq \mathbf{E}[\text{MAXGAP}(S)] + \underbrace{\sqrt{\frac{\ln 2/\delta}{m}}}_{\varepsilon}$$

צעד 2: נראה כי $R_D(S) \geq \mathbf{E}_S[\text{MAXGAP}(S)]$.
נוסיף דגימה S' בגודל m. אזי:

$$\mathbf{E}[\text{error}_{S'}(h)] = \text{error}_D(h)$$

לכן:

$$\begin{aligned}\mathbf{E}[\text{MAXGAP}(S)] &= \mathbf{E}_S\left[\max_{h \in H} \mathbf{E}_{S'}[\text{error}_{S'}(h) - \text{error}_S(h)]\right] \\ &\leq \mathbf{E}_{S,S'}\left[\max_{h \in H}\{\text{error}_{S'}(h) - \text{error}_S(h)\}\right] \\ &= \mathbf{E}_{S,S'}\left[\max_{h \in H} \frac{1}{m} \cdot \sum_{i=1}^{m} \left(\text{error}_{x_i'}(h) - \text{error}_{x_i}(h)\right)\right]\end{aligned}$$

אם $\sigma_i = -1$ אז נבחר $x_i \in S$ ו־$x_i' \in S'$. אם $\sigma_i = 1$ אז נבחר $x_i' \in S$ ו־$x_i \in S'$.
ואז:

$$\begin{aligned}\mathbf{E}[\text{MAXGAP}(S)] &\leq \mathbf{E}_S\left[\max_{h \in H} \frac{1}{m} \cdot \sum_{i=1}^{m} \left(\text{error}_{x_i'}(h) - \text{error}_{x_i}(h)\right)\right] \\ &\leq \frac{2}{m} \cdot \mathbf{E}_{S,\sigma}\left[\max_{h \in H} \sum_{i=1}^{m} \sigma_1 \cdot \text{error}_{x_i}(h)\right]\end{aligned}$$

נשים לב כי:

$$\sigma_i \cdot (1 - 2 \cdot \text{error}_{x_i}(h)) = \sigma_i \cdot \underbrace{(c^\star(x_i) \cdot h(x_i))}_{\text{error}} \simeq \sigma_i \cdot \text{error}_{x_i}(h)$$

השקילות היא בין שגיאה בטווח ± 1 לבין שגיאה בטווח $\{0, 1\}$.
לכן:

$$\begin{aligned} R_D(H) &= \mathbf{E}_\sigma \left[\max_{h \in H} \frac{1}{m} \cdot \sum_{i=1}^{m} \sigma_i \cdot (1 - 2 \cdot \text{error}_{x_i}(h)) \right] \\ &= 2 \cdot \mathbf{E} \left[\max_{h \in H} \frac{1}{m} \cdot \sum_{i=1}^{m} \sigma_i \cdot \text{error}_{x_i}(h) \right] \end{aligned}$$

ולכן:

$$\mathbf{E}_S [\text{MAXGAP}(S)] \leq R_D(H)$$

□

פרק 10

שיעור 10[1]

סיכום השיעור אינו זמין.

[1] שיעור שהתקיים בתאריך 30.12.2012.

פרק 11

שיעור 11[1]

סיכום השיעור אינו זמין.

[1] שיעור שהתקיים בתאריך 06.01.2013.

פרק 12

רגרסיה[1]

12.1 הקדמה

נבחן את הדוגמה הבאה: נרצה לדעת לחזות כמה גשם ירד (בעתיד). נניח שיש מדגם $S = \{(x_i, y_i)\}$, כאשר x_i נדגם לפי D ו־$y_i \in \mathbb{R}$.
נסתכל על פונקציה $f(x) = \mathbf{E}[y \mid x]$. המטרה היא לבנות $h(x)$ כך ש־$|f(x) - h(x)|$ יהיה קטן. בפועל, נרצה ש־$(f(x) - h(x))^2$ יהיה קטן.
בגלל שאין לנו באמת גישה ל־$f(x)$, כל האלגוריתמים יסתכלו על $(y - h(x))^2$. נרצה להוכיח שזה טוב באותה מידה.

טענה

$$\underbrace{\mathbf{E}_{x,y}\left[(y - h(x))^2\right]}_{\text{Measured error}} = \underbrace{\mathbf{E}_x\left[(f(x) - h(x))^2\right]}_{\text{Estimation error}} + \underbrace{\mathbf{E}_{x,y}\left[(y - f(x))^2\right]}_{\text{Uncertainty of the problem}}$$

הוכחה

$$\begin{aligned}
\mathbf{E}_{x,y}\left[(y - h(x))^2\right] &= \mathbf{E}\left[(y - f(x) + f(x) - h(x))^2\right] \\
&= \mathbf{E}\left[(y - f(x))^2\right] + \mathbf{E}\left[(f(x) - h(x))^2\right] \\
&\quad +2 \cdot \mathbf{E}\left[\underbrace{(y(x) - f(x))}_{=0} \cdot (f(x) - h(x))\right] \\
&= \mathbf{E}\left[(y - f(x))^2\right] + \mathbf{E}\left[(f(x) - h(x))^2\right]
\end{aligned}$$

□

בהינתן מגדם $S = \{(x_i, y_i)\}$ נרצה להביא למינימום את:

$$L = \frac{1}{m} \cdot \sum_{i=1}^{m} (y_i - h(x_i))^2$$

[1] סיכום לשיעור שהתקיים בתאריך 13.01.2013.

דוגמה אם $h(x) = x \cdot w$, אזי:

$$\begin{aligned} L &= \frac{1}{m} \cdot \sum_{i=1}^{m} (y_i - h(x_i))^2 \\ &= \frac{1}{m} \cdot \sum_{i=1}^{m} (y_i - x_i \cdot w)^2 \end{aligned}$$

נגזור:

$$\frac{d}{dw} L = \frac{1}{m} \cdot \sum_{i=1}^{m} (-2 (y_i - x_i \cdot w) \cdot x_i) = 0$$

ולכן:

$$\frac{1}{m} \cdot \sum_{i=1}^{m} y_i \cdot x_i = \left(\frac{1}{m} \cdot \sum_{i=1}^{m} x_i^2 \right) \cdot w$$

ולכן:

$$w = \frac{\sum_{i=1}^{m} x_i \cdot y_i}{\sum_{i=1}^{m} x_i^2}$$

12.2 רגרסיה לינארית

נניח כי $h_w(x) = x \cdot w$. נרצה להביא את:

$$L = \frac{1}{m} \cdot \sum_{i=1}^{m} (y_i - h(x_i))^2 = \frac{1}{m} \cdot \sum_{i=1}^{m} (y_i - x_i \cdot w)^2$$

למינימום.
נסתכל על זה כמטריצות. נסמן:

$$W = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix}$$

$$X = \begin{pmatrix} -x_1- \\ -x_2- \\ \vdots \\ -x_m- \end{pmatrix}$$

ואז:

$$X \cdot W - y = \begin{pmatrix} x_1 \cdot W - y_1 \\ x_2 \cdot W - y_2 \\ \vdots \\ x_m \cdot W - y_m \end{pmatrix}$$

כאשר $x_i \in \mathbb{R}^n$, $W \in \mathbb{R}^n$ ו־$y_i \in \mathbb{R}$.
אזי:

$$\begin{aligned}\min_w L &= \min_w (X \cdot W - y)^T \cdot (X \cdot W - y) \\ &= \min_w (X \cdot W)^T \cdot (X \cdot W) - 2 \cdot (X \cdot W) \cdot y + y \cdot y\end{aligned}$$

הגרדיאנט:

$$\begin{aligned}\nabla_W L = 2 \cdot X^T \cdot X \cdot W - 2 \cdot X^T \cdot y &= 0 \\ w &= \left(X^T \cdot X\right)^{-1} \cdot X^T \cdot y\end{aligned}$$

ראינו כי לרגרסיה לינארית יש פתרון סגור. ניתן להכליל את הפתרון גם לרגרסיה מסדר יותר גבוה, בעזרת שימוש ב־Kernel של מרחב וקטורי כלשהו.

12.2.1 חוסר יציבות הפתרון

אחת הבעיות המשמעותיות של הרגרסיה הלינארית היא חוסר יציבות הפתרון. היינו מצפים ששני קלטים מאוד דומים יתנו תוצאות דומות.
למשל, $x_1 = (1, 0)$, $x_2 = (1, \varepsilon)$, $y_1 = 1 = y_2$. אזי:

$$\begin{aligned}L &= \min_w \left[\frac{1}{2} \cdot (w_1 - 1)^2 + \frac{1}{2} \cdot (w_1 + \varepsilon \cdot w_2 - 1)^2\right] \\ \frac{d}{dw_1} L &= (w_1 - 1) + (w_1 + \varepsilon \cdot w_2 - 1) = 0 \\ \frac{d}{dw_2} L &= (w_1 + \varepsilon \cdot w_2 - 1) \cdot \varepsilon = 0 \\ w &= (w_1, w_2) = (1, 0)\end{aligned}$$

נשנה את y_1 להיות $y_1 = 1 + \varepsilon$. עכשיו:

$$\begin{aligned}L &= \min_w \left[\frac{1}{2} \cdot (w_1 - 1 - \varepsilon)^2 + \frac{1}{2} \cdot (w_1 + \varepsilon \cdot w_2 - 1)^2\right] \\ \frac{d}{dw_1} L &= (w_1 - 1 - \varepsilon) + (w_1 + \varepsilon \cdot w_1 - 1) = 0 \\ \frac{d}{dw_2} L &= \varepsilon \cdot (w_1 + \varepsilon \cdot w_2 - 1) = 0\end{aligned}$$

הפתרון יהיה $w_1 = 1 + \varepsilon$, $w_2 = -1$.

12.3 רגולריזציה

12.3.1 Ridge Regression

נגדיר *Ridge Regression*: נרצה להביא למינימום את:

$$\frac{1}{2} \cdot \sum_{i=1}^{m} \underbrace{(w \cdot x_i - y_i)^2}_{SQ=\text{Quadratic Loss}} + \frac{\lambda}{2} \cdot \|w\|^2$$

כאשר $\frac{\lambda}{2} \cdot \|w\|^2$ הוא המשקל של w, ו־λ הוא קבוע ל־trade off בין שני החלקים ברגרסיה.
בכתיב מטריציוני:

$$\begin{aligned} L &= \frac{1}{2} \cdot (X \cdot w - y)^T \cdot (X \cdot w - y) + \frac{\lambda}{2} \cdot W^T \cdot W \\ \nabla_W L &= X^T \cdot X \cdot W - X^T \cdot y + \lambda \cdot W = 0 \\ \underbrace{\left(X^T \cdot X + \lambda \cdot \mathbb{I}\right)}_{\text{Eign values} \geq \lambda} \cdot W &= X^T \cdot y \end{aligned}$$

$$W = \underbrace{\left(X^T \cdot X + \lambda \cdot \mathbb{I}\right)^{-1}}_{\text{Eign values} \leq 1/\lambda} \cdot X^T \cdot y$$

12.3.2 Lasso Regression

יש עוד דרך לייצב את הרגרסיה: *Lasso Regression*. נרצה להביא למינימום את $SQ - \lambda \cdot \|w\|_1$.

קשה לנמק את זה מבחינה אנליטית, ולכן אין פתרון סגור. לעומת זאת, קל לפתור את זה בתוכנה. הפתרון יהיה דליל במספר המשקולות.

אי אפשר להוכיח את זה בצורה מתמטית, אבל אפשר להבין את זה על סמך ניסיון.

12.3.3 חסם הכללה ל־Ridge Regression (או משהו שדומה לו)

נחפש את:

$$\min_w \sum_{i=1}^{m} \psi_i^2$$

כאשר:

$$\psi_i^2 = w \cdot x_i - y_i$$

ונדרוש:

$$\|w\|_2^2 \leq \Lambda^2$$

נניח גם כי:

$$\|x_i\| \leq R$$

למה (*Ledoux-Talag*) נניח כי $\Phi_i : \mathbb{R} \to \mathbb{R}$ היא M־ליפשיץ, כלומר:

$$\forall a, b. \left|\Phi_i(a) - \Phi_i(b)\right| \leq M \cdot |a - b|$$

אזי:

$$\begin{aligned} \hat{R}_S(\Phi \circ H) &= \mathbf{E}_\sigma \left[\sup_{h \in H} \frac{1}{m} \cdot \sum_{i=1}^{m} \sigma_i \cdot \Phi(h(x_i)) \right] \\ &\leq M \cdot \mathbf{E}_\sigma \left[\sum_{h \in H} \left| \frac{1}{m} \cdot \sum_{i=1}^{m} \sigma_i \cdot h(x_i) \right| \right] \\ &= M \cdot \hat{R}_S(H) \end{aligned}$$

במקרה שלנו, נניח כי $\psi_i = w \cdot x_i - y_i$ חסום על ידי M. אזי $\Phi(\psi) = \psi^2$ היא $2M$-ליפשיץ. מהלמה, נקבל כי:

$$\hat{R}_S(SQ) \leq 2 \cdot M \cdot \hat{R}_S(\text{linear}) \leq 2 \cdot M \cdot \frac{R \cdot \Lambda}{\sqrt{m}}$$

מכאן, בהסתברות $1-\delta$:

$$\mathbf{E}[SQ] \leq SQ(S) + 2 \cdot \hat{R}_S(SQ) + O\left(\sqrt{\frac{\log 1/\delta}{m}}\right)$$

12.3.4 נקודת מבט בייסיאנית

נגדיר: $y = w \cdot x + \text{Noise}$, כאשר $\text{Noise} \sim N(0, \sigma)$ ו־$w \sim N(0, \sigma_0)$.
לפי חוק Bayes:

$$\Pr[w \mid \{(x_i, y_i)\}] = \frac{\Pr[\{(x_i, y_i)\} \mid w] \cdot \Pr[w]}{\Pr[\{(x_i, y_i)\}]}$$

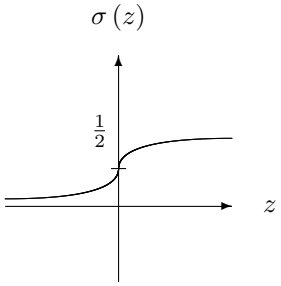
לפי שיטת Maximum Likelihood:

$$\begin{aligned} w_{\text{ML}} &= \arg\max_w \Pr[\{(x_i, y_i)\} \mid w] \\ &= \arg\max_w \prod_{i=1}^{m} \Pr[(x_i, y_i) \mid w] \\ &= \arg\max_w e^{-\sum_{i=1}^{m} \frac{(y_i - w \cdot x_i)^2}{2 \cdot \sigma^2}} \\ &= \arg\min_w \sum_{i=1}^{m} \frac{1}{2 \cdot \sigma^2} \cdot (y_i - w \cdot x_i)^2 \end{aligned}$$

קיבלנו מינימום של שגיאה ריבועית, כלומר רגרסיה לינארית.
לפי שיטת MAP:

$$\begin{aligned} w_{\text{MAP}} &= \arg\max_w \Pr[\{(x_i, y_i)\} \mid w] \cdot \Pr[w] \\ &= \arg\max_w e^{-\sum_{i=1}^{m} \frac{(y_i - w \cdot x_i)^2}{2 \cdot \sigma^2}} \cdot e^{-\frac{\|w\|_2^2}{2 \cdot \sigma_0^2}} \\ &= \arg\min_w \underbrace{\sum_{i=1}^{m} \frac{(y_i - x_i \cdot w)^2}{2 \cdot \sigma^2}}_{\text{Quadratic error}} + \underbrace{\frac{\|w\|_2^2}{2 \cdot \sigma_0^2}}_{\text{Weight of } w} \end{aligned}$$

קיבלנו Ridge Regression עם $\lambda = \frac{\sigma^2}{\sigma_0}$.
התרגיל הזה הראה לנו שאפשר לגזור את הרגרסיות גם מתוך עולם בייסיאני.



איור 12.1: תיאור של $\sigma(z)$

12.4 Logistic Regression

נחזור לבעיות סיווג: נניח כי $Y = \{0, 1\}$ או $Y = \{+1, -1\}$. עדיין יש משמעות ללדבר על $\Pr[Y = 1 \mid X]$.
אזי:

$$\mathbf{E}\left[(w \cdot x_i - y_i)^2\right] = \mathbf{E}\left[w \cdot x - \Pr[y = 1 \mid x]\right]$$

הרגרסיה תבנה מודל שבו

$$\Pr[Y = 1 \mid X] \simeq w \cdot x$$

עד כאן הכל עובד כמו שצריך. אבל לא מובטח ש־$w \cdot x \in \{0, 1\}$. נרצה להחזיר את הפתרון שלנו חזרה לעולם של בעיות הסיווג. לכן, נאמר כי

$$\Pr[Y = 1 \mid X] = \sigma(w \cdot x)$$

כאשר

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

נשים לב לתכונה הבאה: $\sigma(z) + \sigma(-z) = 1$. למה זה מתקיים?

$$\begin{aligned}
\sigma(z) + \sigma(-z) &= \frac{1}{1 + e^{z}} + \frac{1}{1 + e^{-z}} \\
&= \frac{\frac{1+e^{z}}{1+e^{-z}} + 1}{1 + e^{z}} \\
&= \frac{1 + e^{z} + 1 + e^{-z}}{(1 + e^{z}) \cdot (1 + e^{-z})} \\
&= \frac{2 + e^{z} + e^{-z}}{1 + \cancel{e^{z} \cdot e^{-z}} + e^{z} + e^{-z}} \\
&= \frac{\cancel{2 + e^{z} + e^{-z}}}{\cancel{2 + e^{z} + e^{-z}}} = 1
\end{aligned}$$

למה זה טוב? אם $\Pr[Y=1 \mid X] = \sigma(w \cdot x)$ אזי:

$$\begin{aligned} 1 - \Pr[Y=1 \mid X] &= \Pr[Y=-1 \mid X] \\ &= \sigma(-w \cdot x) = 1 - \sigma(w \cdot x) \end{aligned}$$

נרצה למצוא w שיתן לנו את

$$\min_w \sum_{i=1}^{m} \left(\sigma(y_i \cdot w \cdot x_i) - y_i\right)^2$$

נחפש מינימום מקומי. אזי:

$$\begin{aligned} w &= \arg\max_w \prod_{i=1}^{m} \Pr[y_i \mid x_i \cdot w] \\ &= \arg\max_w \sum_{i=1}^{m} \log \Pr[y_i \mid x_i \cdot w] \\ &= \arg\max_w \sum_{i=1}^{m} \log \sigma(y_i \cdot x_i \cdot w) \\ &= \arg\min_w \sum_{i=1}^{m} -\log \sigma(y_i \cdot x_i \cdot w) \\ &= \arg\min_w \sum_{i=1}^{m} \log\left(1 + e^{y_i \cdot x_i \cdot w}\right) \end{aligned}$$

איך נשערך את ההסתברויות הבאות:

$$\begin{aligned} \Pr[Y=1 \mid X] &= \sigma(w \cdot x_i) \\ \Pr[Y=-1 \mid X] &= \sigma(-w \cdot x_i) \end{aligned}$$

נהפוך את זה למסווג: $\sigma(w \cdot x) \geq 1/2$ או $w \cdot x \geq 0$. קיבלנו בחזרה מפריד לינארי.

פרק 13

[1]Model Selection

13.1 הקדמה

בהינתן מספר דוגמאות נתון, נרצה למצוא השערה טובה ביותר.
נניח כי $c^\star$ היא פונקציית המטרה ו־H היא מחלקת ההשערות. לשם פשטות (לפחות לאינטואיציה בהתחלה), נניח כי $c^\star \in H$. נניח כי H בת־מניה: $H = \{h_1, \ldots, h_i, \ldots\}$.
אלגוריתם 1.13 הוא אלגוריתם Model Selection. ננתח אותו:
ראשית, האלגוריתם תמיד עוצר. מכיוון ש־$c^\star \in H$, אזי $\exists j.h_j = c^\star$. אם נגיע לפאזה j, אז בוודאי נעצור, כי $c^\star = h_j$ תהיה עקבית. לכן, מספר הפאזות חסום על ידי j.
כמו כן, אם האלגוריתם עוצר ומחזיר את h_i, אזי מתקיים:

$$\Pr\left[h_i \text{is consistent} \mid \text{error}\left(h_i\right) > \varepsilon\right] \leq \delta_i$$

למה? ההסתברות שנעצור היא ההסתברות שב־m_i לא ראינו שגיאות. ההסתברות לכך היא:

$$\left(1 - \text{error}\left(h_i\right)\right)^{m_i} \leq \left(1 - \varepsilon\right)^{m_i} \leq e^{-\varepsilon \cdot m_i} = \delta_i$$

נבחן את פרמטר הביטחון:

$$\delta = \sum_{i=1}^{\infty} \delta_i$$

למשל, $\delta_i = \frac{1}{i^2}$ או $\delta_i = \frac{1}{2^i}$ נותנים $\delta = 1$ או $\delta = \frac{\pi^2}{6}$ בהתאמה. לרוב נגדיר $\delta_i = \frac{\delta}{i^2} \cdot \frac{6}{\pi^2}$ ונקבל פרמטר ביטחון δ.

[1] סיכום לשיעור שהתקיים ב־20.01.2013.

אלגוריתם 1.13 Model Selection

נרוץ בפאזות. בפאזה ה־i:

- נדגום $m_i = \frac{1}{\varepsilon} \cdot \ln \frac{1}{\delta_i}$ דוגמאות.
  - אם h_i עקבי, נחזיר את h_i.
  - אחרת, נמשיך לפאזה הבאה.

	החלפות	שגיאות
+ + + − + + − − − − + − − + − − −	דגימה	
−	0	7
+ −	1	3
+ |−| + − |+| − |+| −	7	0

טבלה 13.1: דוגמאות לדגימות והשערות עליהן

נבחן את מספר הדוגמאות: בלי מחזור של דוגמאות, נשתמש ב־$\sum_{i=1}^{j} m_i$ דוגמאות. עם מחזור:

$$\max_{1\leq i\leq j}\{m_i\} = m_j$$

כאשר:

$$m_j = \frac{1}{\varepsilon}\cdot\ln\left(\frac{j^2}{\delta}\cdot\frac{6}{\pi^2}\right)$$

כלומר, מספר הדוגמאות תלוי באינדקס של פונקציית המטרה.

13.1.1 דוגמה

ניתן לראות בטבלה 13.1 דוגמה לדגימות והשערות עליהן.
היינו מעדיפים לקבל את האפשרות השנייה: האפשרות הראשונה מפספסת יותר מדי, ואילו האפשרות השלישית יותר מדי מותאמת לדגימה. האפשרות השנייה היא משהו טוב ביניהן.

13.1.2 המודל

נתונה משפחת השערות H, כך ש־$H_1 \subseteq H_2 \subseteq H_3 \subseteq \cdots \subseteq H_i \subseteq \cdots \subseteq H$. כמו כן:

$$H = \bigcup_{i=1}^{\infty} H_i$$

הסיבוכיות של H_i: $\mathrm{VCdim}(H_i) = i$, $|H_i| = 2^{i-1}$, $|H_1| = 1$.
ננית כי פונקציית המטרה $c^\star$ מקיימת $c^\star \notin H$. נסמן:

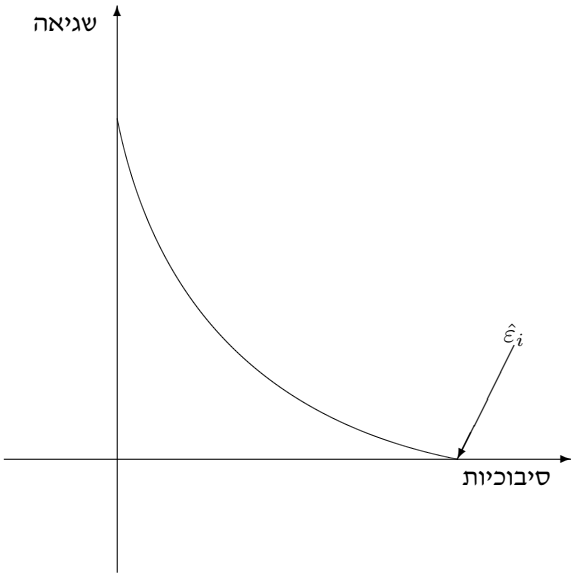
$$\begin{aligned}\varepsilon(h) &= \Pr[h \neq c^\star]\\ \varepsilon_i &= \min_{h\in H_i}\varepsilon(h_i)\end{aligned}$$

בגלל ש־$H_i \subseteq H_{i+1}$, אזי $\varepsilon_i \geq \varepsilon_{i+1}$, ולכן נגדיר:

$$\varepsilon^\star = \inf_i \varepsilon_i$$

כמו כן, נגדיר:

$$\begin{aligned}\hat{\varepsilon}(h) &= \frac{1}{m}\cdot\sum_{x\in S}\mathbf{I}(h(x)\neq c^\star(x))\\ \hat{\varepsilon}_i &= \min_{h\in H_i}\hat{\varepsilon}(h)\end{aligned}$$



איור 13.1: השגיאה כפונקציה של ההסתברות כאשר מסתכלים על $\hat{\varepsilon}_i$.

מכיוון ש־$\text{VCdim}(H) = \infty$, אזי כל m נקודות יכולות לקבל את כל הסיווגים. לכן, $\hat{\varepsilon}_m = 0$.

היינו יכולים לבחור את $g = \arg\min_h \hat{\varepsilon}(h)$, אבל תמיד קיים $h \in H_m$ כך ש־$\hat{\varepsilon}(h) = 0$, ולכן אין בהכרח משמעות ל־g (ניתן לראות את זה באיור 13.1).

אפשר לבחור: $g^\star = \arg\min_h \{\hat{\varepsilon}(h) + p(h)\}$. נגדיר:

$$d(h) = \min\{i \mid h \in H_i\}$$

13.2 Structural Risk Minimization

ב־Structural Risk Minimization (או בקיצור SRM), נבחר:

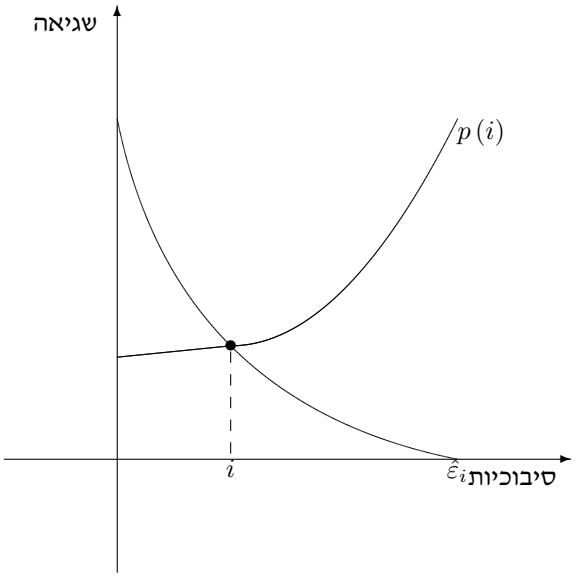
$$\text{Penalty}(h) = p(h) = \sqrt{\frac{2 \cdot d(h) \cdot \ln 2 + \ln 1/\delta}{m}}$$

כאשר m הוא מספר הדוגמאות ו־δ הוא פרמטר הבטחון.

נבחר:

$$g^\star = \arg\min_{h \in H} \hat{\varepsilon}(h) + \sqrt{\frac{2 \cdot d(h) \cdot \ln 2 + \ln 1/\delta}{m}}$$

נסמן: $h^\star = \arg\min_{h \in H} \varepsilon(h)$.



איור 13.2: השגיאה כפונקציה של ההסתברות כאשר מוסיפים את Penalty להשערות ארוכות

משפט בהסתברות $1-\delta$:

$$\varepsilon(g^\star) \leq \varepsilon(h^\star) + 2 \cdot \text{Penalty}(h)$$

הוכחה ניזכר בטענה כי $\Pr[|\varepsilon(h) - \hat{\varepsilon}(h)| \geq \lambda] \leq 2 \cdot e^{-\lambda^2 \cdot m}$ (נובעת ישירות מ־Chernoff). נבחר:

$$g_i = \arg\min_{h \in H_i} \hat{\varepsilon}(h)$$

אזי:

$$\begin{aligned} \Pr[|\varepsilon(g_i) - \hat{\varepsilon}(h_i)| \geq \lambda_i] &\leq \Pr[\exists h \in H_i. |\varepsilon(h) - \hat{\varepsilon}(h)| \geq \lambda_i] \\ &\leq |H_i| \cdot 2 \cdot e^{-\lambda_i^2 \cdot m} = 2^i \cdot e^{-\lambda_i^2 \cdot m} = \delta_i \end{aligned}$$

נקבע:

$$\lambda_i = \sqrt{\frac{i \cdot \ln 2 + \ln 2^i/\delta}{m}}$$

נחשב את ההסתברות עבור כל ה־i־ים:

$$\begin{aligned} \Pr[\forall i. \forall h \in H_i. |\varepsilon(h) - \hat{\varepsilon}(h)| \leq \lambda_i] &= 1 - \Pr[\exists i. \exists h \in H_i. |\varepsilon(h) - \hat{\varepsilon}(h)| \geq \lambda_i] \\ &= 1 - \sum_{i=1}^{\infty} \delta_i = 1 - \delta \end{aligned}$$

אפשר לבחור $\delta_i = \frac{\delta}{2^i}$ או $\delta_i = \frac{\delta}{i^2} \cdot \frac{6}{\pi^2}$. לכן, בהסתברות $1-\delta$, לכל $h \in H$ מתקיים $|\varepsilon(h) - \hat{\varepsilon}(h)| \leq \lambda_{d(h)}$.
נציב את δ_i ב־λ_i:

$$\lambda_i = \sqrt{\frac{i \cdot \ln 2 + \ln {}^{2^i}\!/_{\delta}}{m}} = \sqrt{\frac{2 \cdot i \cdot \ln 2 + \ln {}^{1}\!/_{\delta}}{m}}$$

נניח כעת כי $d(h^\star) = i$ ו־$d(g^\star) = j$. אזי:

$$\begin{aligned} \hat{\varepsilon}(h^\star) &\leq \varepsilon(h^\star) + \lambda_i \\ \varepsilon(g^\star) - \lambda_j &\leq \hat{\varepsilon}(g^\star) \end{aligned}$$

כמו כן, מבחירת $g^\star$:

$$\begin{aligned} \hat{\varepsilon}(g^\star) + p_j &\leq \hat{\varepsilon}(h^\star) + p_i \\ &\Downarrow \\ \varepsilon(g^\star) - \lambda_j + p_j &\leq \varepsilon(h^\star) + \lambda_j + p_j \end{aligned}$$

נשים לב כי $\lambda_k = p_k$. לכן: $\varepsilon(g^\star) \leq \varepsilon(h^\star) + 2 \cdot p_i$. □

13.3 Cross Validation ($\mathcal{CV}$)

נשים לב: ב־SRM בשלב א', מכל H_i בוחרים g_i, ובשלב ב' בוחרים בין g_i בעזרת p_i. ב־$\mathcal{CV}$, שלב א' זהה, ובשלב ב' מערבים דוגמאות חדשות.
נתון מדגם S בגודל m. נחלק ל־S_1 ו־S_2, כאשר $|S_1| = (1-\gamma) \cdot m$ ו־$|S_2| = \gamma \cdot m$.
S_1 ישמש את שלב א', ו־S_2 ישמש את שלב ב'.

שלב א' לכל H_i נבחר $g_i = \arg\min_{h \in H_i} \hat{\varepsilon}_1(h)$, כאשר $\hat{\varepsilon}_1$ פועל על S_1. נגדיר:

$$G = \{g_i \mid \hat{\varepsilon}_{i,1}(g_i) < \hat{\varepsilon}_{i-1,1}(g_i)\}$$

נשים לב כי $|G| \leq m$.

שלב ב' נבחר מתוך G בעזרת S_2. כלומר:

$$g^\star = \arg\min_{g \in G} \hat{\varepsilon}_2(g)$$

כאשר $\hat{\varepsilon}_2$ פועל על S_2.
יהי $\mathcal{A}$ אלגוריתם כללי שבוחר מתוך G.

משפט (*Cross-Validation*)

$$\varepsilon_{\mathcal{CV}}(m) \leq \varepsilon_{\mathcal{A}}((1-\gamma) \cdot m) + 2 \cdot \sqrt{\frac{\ln \frac{2 \cdot m}{\delta}}{m}}$$

הוכחה נניח כי g_i נבחר על ידי $\mathcal{CV}$, ונשווה אותו ל־$g_k \in G$ כללי. נרצה להראות:

$$\varepsilon\left(g_i\right) \leq \varepsilon\left(g_{\mathcal{A}}\right) + 2 \cdot \sqrt{\frac{\ln \frac{2 \cdot m}{\delta}}{m}}$$

טענת ההכללה:

$$\begin{array}{rcl} \forall g \in G. \Pr\left[\left|\varepsilon\left(g\right) - \hat{\varepsilon}\left(g\right)\right| \geq \lambda\right] & \leq & 2 \cdot e^{-\lambda^2 \cdot \gamma \cdot m} \\ & \Downarrow & \\ \Pr\left[\exists g \in G. \left|\varepsilon\left(g\right) - \hat{\varepsilon}\left(g\right)\right| \geq \lambda\right] & \leq & 2 \cdot |G| \cdot e^{-\lambda^2 \cdot \gamma \cdot m} \\ & \leq & 2 \cdot m \cdot e^{-\lambda^2 \cdot \gamma \cdot m} = \delta \\ & \Downarrow & \\ \lambda & = & \sqrt{\frac{\ln \frac{2 \cdot m}{\delta}}{\gamma \cdot m}} \\ & \Downarrow & \\ \Pr\left[\exists g \in G. \left|\varepsilon\left(g\right) - \hat{\varepsilon}\left(g\right)\right| \geq \lambda\right] & \leq & \delta \end{array}$$

לכן, בהסתברות $1 - \delta$:

$$\begin{array}{rcl} \varepsilon\left(g_i\right) - \lambda & \leq & \hat{\varepsilon}_2\left(g_i\right) \\ \hat{\varepsilon}_2\left(g_k\right) & \leq & \varepsilon\left(g_k\right) + \lambda \\ \hat{\varepsilon}_2\left(g_i\right) & \leq & \hat{\varepsilon}_2\left(g_k\right) \end{array}$$

לכן, $\varepsilon\left(g_i\right) - \lambda \leq \varepsilon\left(g_k\right) + \lambda$, ולכן $\varepsilon\left(g_i\right) \leq \varepsilon\left(g_k\right) + 2 \cdot \lambda$.
מההגדרה, $\mathbf{E}\left[\varepsilon\left(g_i\right)\right] = \varepsilon_{\mathcal{CV}}\left(m\right)$. כמו כן:

$$\mathbf{E}\left[\min_{g_k \in G} \varepsilon\left(g_k\right)\right] \leq \varepsilon_{\mathcal{A}}\left(\left(1 - \gamma\right) \cdot m\right)$$

לכן:

$$\varepsilon_{\mathcal{CV}}\left(m\right) \leq \varepsilon_{\mathcal{A}}\left(\left(1 - \gamma\right) \cdot m\right) + 2 \cdot \sqrt{\frac{\ln \frac{2 \cdot m}{\delta}}{m}}$$

□

10-fold Cross Validation מחלק ל־$\{S_i\}_{i=1}^{10}$, כאשר $|S_i| = \frac{m}{10}$. האלגוריתם משתמש ב־9 S_i־ים, ובודק על הנותר.

13.4 Minimum Description Length ($\mathcal{MDL}$)

בשלב הראשון נמצא קידוד ל־h, ובשלב השני נתקן. אם יש הרבה תיקונים, עדיף למצוא קידוד אחר ל־h.
לכן, נחפש:

$$g^{\star} = \arg\min_{h} \left\{ \underbrace{\text{size}\left(h\right)}_{p(h)} + \underbrace{\text{size}\left(\text{corrections}\right)}_{f(\hat{\varepsilon}(h))} \right\}$$

13.4.1 בעזרת MAP

לפי ההגדרה:

$$\Pr[h \mid D] = \frac{\Pr[D \mid h] \cdot \Pr[h]}{\Pr[D]}$$

לכן:

$$\begin{aligned} g^\star &= \arg\max_h \Pr[D \mid h] \cdot \Pr[h] \\ &= \arg\max_h \log \Pr[D \mid h] + \log \Pr[h] \\ &= \arg\min_h \frac{1}{\log \Pr[D \mid h]} + \frac{1}{\log \Pr[h]} \end{aligned}$$