Chapter 3, The Bayesian Network Representation

Seminar on Probabilistic Graphical Models
October 27th 2015
Lecture notes and a bit aftermath, by Omer Tabach.

Motivation

Bayesian Networks, as Haim presented last week, serve as a compact representation for joint distribution over many variables.

Goals for today:

Key terms Covered

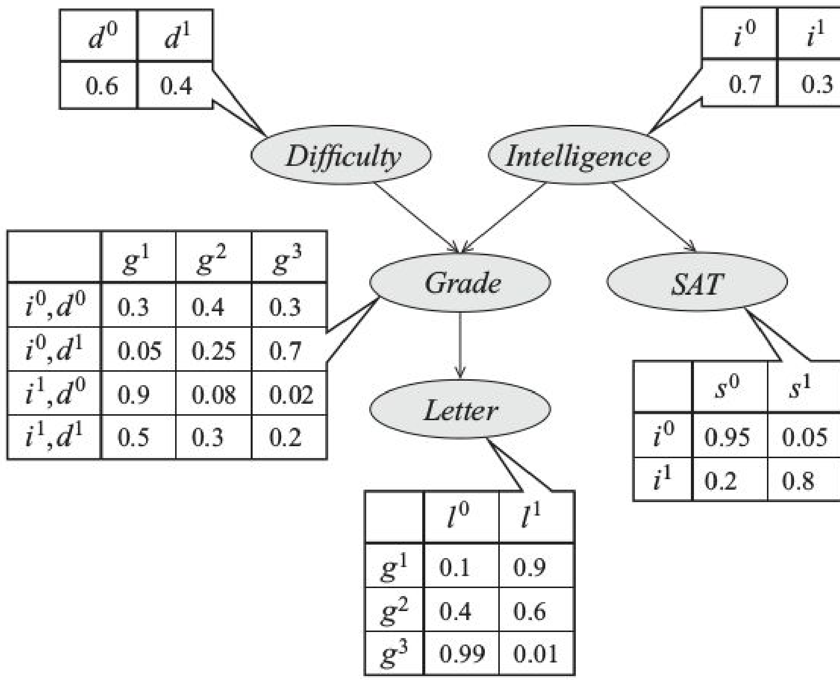
CPD,
Chain Rule,
Intercausal Reasoning.
I-Map
D-Separation
I-Equivalence
Minimal I-Map
Perfect Map

Student Example

Introduction

מיומנו של יועץ באקמי כלכלה.
בוחן את מועמדותו של ג'ורג' לחברה.
מנת משכל כהתאמה לתפקיד.
מתאם בין מנת משכל לציון פסיכומטרי.
מתאם בין מנת משכל לציון בקורס.
ידוע שהקורס לפעמים קל ולפעמים קשה כתלות במרצה.
ג'ורג' מעביר מכתב המלצה. ידוע שפרופסורים לא מכירים אף אחד ומחברים מכתב שתלוי רק בציון.

3.2. Bayesian Networks 53



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Figure 3.4 Student **Bayesian network** $\mathcal{B}^{student}$ **with CPDs**

Explaining the things on the graph:

Directed Dependencies and CPDs. [Conditional Probability Distribution.]
Each graph node depends only on its parents.

"Chain rule for Bayesian networks"

עוד רגע לפני שנעבור להגדרות פורמאליות נרצה להציג שאילתא.

Example of use of such network.
Query: ‘What are the odds that an intelligent student took an easy course, got a bad grade, a bad letter but still had a high score on his SAT?’
Answer, via evaluation of the graph’s nodes in some topoligcal ordering:

$$P(I=1,D=0,G=2,L=0,S=2)=P(I=1)\cdot P(D=0)\cdot P(G=2|I=1,D=0)\cdot P(L=0|G=2)\cdot P(S=2|I=1)$$

זו דוגמא לכלל השרשרת על רשת באייסיאנית. נחזור לזה בהמשך.

Bayesian Network

Definition 3.5

Bayesian network is a pair, BN=(G, P) where G is graph and P factorizes over G. P is specified as a set of CPDs associated with G's nodes. The distribution P is often annotated Pb.

Intuition:

The graph itself can be viewed in 2 different ways:
1. A data structure providing a skeleton for compact representation of the factorized joint distribution.
  _compact: Considering the 2^n alternative.
  _factorized: We look at random variables, disregarding their possible assignments (when building the graph.)
 - This will haunt us later.

כלומר, מבנה נתונים שאפשר לאכלס בו התפלגויות משותפות בצורה יעילה!

Factorized, הכוונה תוך התעלמות מההשמות האפשריות למשנים. - נחזור לזה אח"כ.

2. A compact representation for a set of conditional **in**dependence assumptions about a distribution.

-This is the, perhaps, conceptually harder direction, which involves looking at the anti-edges. We will later see it is quite useful.

- שאלה שעלתה בכיתה: אי תלות מותנית, כלומר אוספים של $X \perp Y | Z$, כאשר Z יכול להיות ריק.

* Each node represents a distribution that depends only on its parents.
  It is called "Local Probability Model".

(אלה הטבלאות שיושבות על כל צומת.)

3.2.2 Basic <u>**In**</u>dependencies in Bayesian Networks

אנחנו הולכים לפתוח את הצד היותר מסובך הזה, של האי-תלויות.

Intuition

In the previous section and example we viewed Bayesian networks as a data structure, encoding local, conditional probabilities.
We're now going to look at Bayesian networks from a different perspective: the graph encodes a set of conditional **in**dependence assumptions.

להזכיר מה זו תלות או אי-תלות.

הגדרה מתמטית לאי תלות: $X \perp Y[\iff?]P(X|Y) = P(X)$.

Conditional Independencies, the Student example.

נחפש אי תלויות מותנות בדוגמאת הסטודנט.

1. L ⊥ I,D,S | G

הווה אומר, ברגע שאנו יודעים את הציון, הציפייה שלנו מאיכות המכתב שלו לא תושפע משום נתון אחר.

2. S ⊥ D,G,L | I

שוב, ברגע שנדע מה רמת האינטיליגנציה של ג'ורג'- אף אחד מהנתונים האחרים לא ישפיע על הציפייה שלנו לציון הפסיכומטרי שלו. או, אם I=1 אז S=0 בהסתברות כלשהי ו-S=1 בהסתברות האחרת ושום מידע על D, G או L לא ישנה את זה.

3. G

עד כה ראינו שכל צומת, בהינתן האבות הישירים שלו בלתי תלוי בשום צומת אחר.
מה לגבי בנים?
אינטואיטיבית: ברור שחוזק המכתב יכול להעיד על רמת הציון.
למשל, P g1 | d1i1 קטן ממש מ P g1 | d1i1s1 .
לטעמי קל יותר לראות בדוגמאת צעצוע,

X -> Y -> Z
Px1 = 0.5 ; P(y1 = x1) = 0.25 ; P(z1 | y1) = 0.9 P(z1|y0) = 0.5
X Y Z
0 0 0 0.0625
0 0 1 0.0625
0 1 0 0.3375
0 1 1 0.0375
1 0 0 0.1875
1 0 1 0.1875
1 1 0 0.0125
1 1 1 0.1125

P(y1|x1) = 0.125 / 0.125 + 0.375 ; P(y1|x1z1)=0.1125/0.1125+0.1875.

3. G ⊥ S | D,I
4. I?

ראינו כבר שכל צומת יכול להיות מושפע מהצאצאים שלו. האם יכול להיות קשר בין I ל-D?
בבקשה להתרכז, זה כנראה החלק הכי מבלבל בהרצאה.

Conditional Independencies: the case for roots, and Intercausal reasoning

מה קורה בין D ל-I?
ראשית, מהסתכלות על הגרף ועל ה-CPD של כל אחד מהם- באופן מובהק D ⊥ I.

5. D ⊥ I

כעת נציג מושג חשוב ומבלבל. Inter-Causal Reasoning מכונה בספר, כאשר בקאוז'אל הכוונה לחיפוש אחר הסיבות, וב-Inter הכוונה לאינטרקציה בין סיבות בלתי תלויות.

The window issue

דוגמא מהירה: אני גר בבית ישן בלי חלונות. לפעמים מופיעות שלוליות. סיכוי לפיצוץ בצנרת, סיכוי לגשם, הופיעה שלולית.

R1 = 0.2 ; X1 = 0.3 ; P1 | r1x1 = 0.9, P1 | r1x0 = 0.9 , P1 | r0x1 = 0.8 P1 | r0x0 = 0.

* בהינתן שהופיעה שלולית, אשמח לשמוע שירד גשם! כי זה מוריד את הסיכוי לפיצוץ בצנרת.

Back to George

השאלה שבאים לענות עליה היא האם תיתכן תלות בין מנת המשכל לקושי הקורס.
שואלים, מה הסיכוי, בלי לדעת כלום, שהסטודנט חכם? -מה-CPD.
כעת שואלים, בהינתן שהקורס קשה: האם זה שינה את דעתנו? לא.
כעת נתון ציון נמוך. מה עכשיו הסיכוי שהסטודנט חכם?
כעת בנוסף לציון הנמוך נתון שהקורס קשה. דעתנו השתנתה!

מסקנה:

D ⊥ I does not imply D ⊥ I | G

בסוף, הואיל ובגרף אין שום קשר בין D ל -S הרי ש-

D ⊥ I,S.

בהמשך נדבר על החשיבות של אי-תלויות מותנות מהסוג הזה.

3.2.3 Graphs and Distributions

כעת נציג משפט חשוב על הקשר בין רשתות להתפלגויות שהן מתיימרות לייצג, משפט שנותן תקפות מתמטית לרצון שלנו לייצג התפלגויות מותנות מרובות משתנים באמצעות BN.
נתעלם מה-CPD. נתייחס רק למבנה הגרף.
נחזור לעבוד מהצד של ההתפלגות המותנית, או התלויות שיש (להבדיל מהתלויות שאין, שאותן חיפשנו קודם), ונשתמש בהן כדי להוכיח תכונה חשובה על הקשר בין התפלגות לבין גרף שמתיימר לייצג אותה.

I-Maps introduction Independence Set Definition

I-Map: $I \subseteq \mathcal{X} \times \mathcal{X} \times 2^{\mathcal{X}}$

לכתוב את ההגדרה: שלשות של סטים של משתנים וניסוחי אי-תלויות או אי-תלויות-מותנות ביניהם.

כל התפלגות משרה סט של אי-תלויות שכאלה. נסמן I(P).

I-Map Definition

Let G be a graph implying a set of independencies I(G). We say that G is an I-map for a set of independencies I if I(G) ⊆ I.

ההגדרה שהושמטה מההרצאה ונוספה בזמן אמת:
נגדיר באופן לוקאלי

I_I(G) = X ⊥ {Y | Y is not a descendant of X} | Parents of X

נבחין כי זה נובע מהגדרת ה-CPD.
נאמר שגרף מייצג התפלגות אם הוא לא משלה אותנו, כלומר, אם לא נוספו לו אי-תלויות מיותרות.
כלומר, אם לא חסרות בו קשתות חשובות.

I(G) ⊆ I(P).

מאידך, יתכן שבהתפלגות יש אי תלויות שאינן נובעות ממבנה הגרף.

נראה בהמשך הגדרה מתמטית שמכלילה את המושג הזה.

אבל קודם כל, דוגמא.

I-Maps Example

נשחק קצת עם מושג ה-I-Map דרך התפלגות משותפת על שני משתנים בינאריים.
יש בדיוק שלושה DAGים מכוונים שאפשר לבנות. שני אלה שיש בהם קשתות- בהם אנו מניחים תלות בין המשתנים. הגרף השלישי, שאין בו קשתות מקודד אי תלות בין המשתנים.
שימו לב שבגרף בלי הקשתות, לא משנה אילו CPD נשים עליו, אין תלות בין המשתנים.

I-Map example 1

0 0 08%
0 1 32%
1 0 12%
1 1 48%

פה למשל X ב"ת ב-Y כי Y=1 זה 80% או 1:4 וגם בהינתן X=0.
כל שלושת ה-DAGים תופסים, למשל הגרף שיש בו קשת מ-X ל-Y עם CPD

	Y=0	Y=1
X=0	2	8
X=1	2	8

או הגרף שאין בו שום קשת.

I-Map example 2

0 0 40%
0 1 30%
1 0 20%
1 1 10%

פה למשל יתקיים ש P X = 1 הוא 3:7 (או 30%) אבל אם נתון Y=1 ירדנו ל- 1:3 (או 25%)
העיקר פה הוא רק ש-Y ⊥ X אינו במפת-אי-התלות של הטבלא הימנית, לכן הגרף הריק (שכן משרה X ⊥ Y) אינו מפת-אי-תלות שלה.

I(G_\phi) \not ⊆ I(P)

Factorization: Bayesian Networks Chain Rule strikes again

P(X_1,...,X_n) = \Multiple i=1 to n P(X_i | Pa x_i)

נאמר ש-P, התפלגות מעל X_1 עד X_n מתפרקת, Factorizes, מעל מבנה גרף G אמ"מ ניתן לבטא את P בתור מכפלת התפלגויות מותנות לפי מבנה G.
נזכיר את כלל השרשרת של התפלגות מותנת:

P(X_1,...,X_n) = P(X_i | X_<i)

I-Maps and Factorization

המשפט שנותן את הבסיס לכל העבודה עם רשתות בייסיאניות.
If G is an I-map for P, then P factorizes according to G. (Theorem 3.1.)
And if P factorizes according to G, then G is an I-map for P. (Th. 3.2.)
הוכחה למשפט 3.1: נניח G הוא I-Map של P, אז כפי שראינו קודם,
I_l(G) = X ⊥ {Y | Y is not a descendant of X} | Parents of X
נניח X_1 עד X_n מיון טופולוגי של הצמתים. נפתח את הגדרת כלל השרשרת של P.
P(X_i) = P(X_i | X_1...X_i-1)
והרי ההורים של X_i נמצאים בקבוצה הזו של X_1 עד X_i-1 והצאצאים שלו לא. אז אין נתון על הצאצאים אז אין השפעה מצמתים שאינם ההורים. אז נשארים עם ה-CPD ומגיעים בדיוק להגדרת ה-Factorization.
הוכחה למשפט 3.2: שלצערי דילגתי עליה בהרצאה.
כרגע מה שנתון זה ההתפלגות המותנית המצומצמת, כלומר, כל התפלגות כל משתנה תלויה רק בהוריו.
ראשית, דוגמא של S: צריך להראות ש-

P(S | I, D, G, L) = P(S | I)

אז מתחילים מהגדרת התלות, מנה בין P(S, I, D, G, L) לבין P(I, D, G, L)
את המונה פותחים ישירות לפי הנתון, כלומר, P(I) P(D) P(S|I) P(G|I, D), P(L|G). את המכנה פותחים לפי כלל הסתברותי של

P(I, D, G, L) = \sum _S P(I, D, G, L, S)

פותחים גם את המונה לפי הכלל הזה, דוחפים את הסכימה ימינה ומצמצמים כי זה בעצם 1.

זו הייתה דוגמא עבור אי תלות מותנית של משתנה יחיד בדוגמא יחידה.
זו לא הוכחה מלאה, אבל שימו לב שעדיין לא נתנו הגדרה מתמטית ל-I(G), ושבדוגמא של ג'ורג' עבדנו אחד אחד.
כעת נחפש דרך למצוא את I(G).

3.3 Independencies in Graphs

נמשיך להתעלם מה-CPD. נתייחס רק למבנה הגרף.
אנו מעוניינים לדעת האם מבנה גרף כלשהו עשוי לתאום לאסוף של התלפגויות.
השתכנעו בחשיבות של אי התלויות בגרף. נחפש דרכים לדעת מה הן.

3.3.1 D-separation

בשלב ראשון נבקש לדעת האם אפשר **להבטיח** $X \perp Y | Z$

נענה על שאלה זו באמצעות התייחסות לשאלה, האם ייתכן ש-X ישפיע על Y בהינתן Z.
אם ניתן לבנות דוגמא, כולל CPDs בה יש השפעה ביניהם- התשובה תהיה לא.
D stands for Directed,
אך אין הכוונה לבדיוק לפי כיווני הגרף אלא שה-D הפרדה מושפעת מכיווני הגרף.
נתחיל במקרה פרטי ונפתח למקרה כללי.
מקרה פרטי פשוט ביותר:

X -> Y

1. קשר ישיר: אם יש קשת מ-X ל-Y אזי לא משנה מה תהיה הקבוצה Z- אפשר לבנות התפלגות בה X יהיה תלוי ב-Y, או להיפך.

X <-> Z <-> Y ?

2. קשר עקיף: נפתור באופן כללי עבור כל השלשות מהצורה X, Y, Z כך שיש ביניהם קשתות.

Causal Trail: X -> Z -> Y

Evidential Trail: X <- Z <- Y

ראינו בדוגמאת הסטודנט ש-X יכול להשפיע על Y, אמ"מ Z לא ידוע.

Common Cause X <- Z -> Y

שוב, כפי שראינו בדוגמאות, מידע לגבי X יכול להשפיע על ההסתברות של Y אמ"מ Z לא ידוע.
(לא הובהר מספיק בהרצאה: נכנה מסלול "פעיל" להיות מסלול כך שתיתכן התפלגות בה X ישפיע על Y או להיפך. במקרה של Causal, Evidental או Common Cause יתקיים שמסלול פעיל אמ"מ Z אינו נתון.)

Common Effect X -> Z <- Y

מקרה יחיד בו המסלול יהיה פעיל אמ"מ Z, או צאצא כלשהו של Z, נתונים.
מבנה של Common Effect כונה בעבר Intercausal Reasoning ויכונה בהמשך V-Structure. כך נהוג גם בספר.
(אם כן, מבנה V פעיל אמ"מ Z *נתון*. בכך הוא מתנהג בדיוק להיפך משלושת המבנים הקודמים.)

General Case

ההרחבה לכך שמסלול כלשהו יהיה **פעיל** תהיה לבקש שכל שלושה צמתים עוקבים במסלול יהיו פעילים.
נאמר, אם כן, שמסלול כלשהו X_1 עד X_n פעיל אמ"מ בכל המבנים מהצורה V - הצומת האמצעי, או צאצא של הצומת האמצעי, פעיל- ואלה הצמתים היחידים הפעילים.
דוגמא מסכמת:
לצייר על הלוח כעין

N Z?

N X Y

N N

N

להסביר שאם כלום לא נתון- X יכול להשפיע על כל הצאצאים שלו, ועל כל האבות שלו, ועל כל הצאצאים של האבות שלו.
כל נתון על מסלול ישיר לאב- עוצר השפעה על מסלול זה. כל נתון על מסלול ישיר לצאצא- עוצר השפעה ממנו למטה, אולם מאפשר השפעות של מבני V.

אם **אין שום** מסלול פעיל בין שני צמתים- נאמר שיש ביניהם הפרדה מכוונת, או **D-Separation**.

D-separation to I-Maps

כעת ניתן לעבור חזרה ל-I-Map:
כעת נגדיר באופן מלא (ולא לוקאלי כמו קודם)

Definition 3.7
I(G)={(X ⊥ Y | Z) ; d-sep(X;Y|Z)}

את הקשר בין I-Map לוקאלי ל-I-Map כללי נראה בתרגיל הבית.

3.2.2 Soundness and Completeness

הגדרנו I(G).
אבל I(G) זו תכונה של הגרף, כלומר, כל רשתות הבאייס בעלות אותו המבנה.
ע"מ שבאמת נוכל להשליך מן ה-D-Separation על ה-I(G) נדרוש תכונות של נאותות ושלמות.

Soundness

If X and Y are d-sep given some Z then we're guaranteed they're conditionally independent given Z over all possible distributions, which factorize over G.

מכך ינבע לנו ש-G הוא I-Map של P.
בהוכחה הם עוברים דרך רשתות מארקוב כך שאולי תראו את זה בשבוע הבא ואולי לא.

[Lack-of] Completeness

הגדרה של נאותות הייתה אם D-מופרדים אזי תמיד בת"ל.
הגדרה של שלמות יכולה להיות: אם לא D-מופרדים אזי תמיד תלויים.
תמיד תלויים, כלומר, תלויים בכל התפלגות שמתפרקת על הגרף.
האם מובטח לנו שבהינתן X ו-Y *שאינם* d-separated אז הם תלויים? ברור שלא, אפשר לדמיין דוגמא מנוונת.

0 0 0.4
0 1 0.4
1 0 0.6
1 1 0.6

ואפשר לבנות את הגרף A ל-B שבו הם לא D-מופרדים למרות שהם כן בת"ל.
על כן נאמץ הגדרה חלשה יותר של שלמות:

Completeness

הגדרה של נאותות הייתה אם **לכל התפלגות** D-מופרדים אזי **לכל התפלגות** תמיד בת"ל.
הגדרה של שלמות תהיה, אם לא לכל התפלגות D-מופרדים אזי קיימת התפלגות שמתפרקת על G בה הם תלויים.
הוכחה דרך בנייה שגם היא לא מובאת בספר אולם קל לדמיין.

D-Separation Conclusion

בהמשך בספר מובא משפט שאומר שכמעט כל התפלגות על גרף תקיים ש I(G)=I(P), כלומר, המקרים בהם יש תלות מיותרת בגרף הם קבוצה ממידה 0 במרחב ההתפלגויות מעל הגרף.
For almost all parametrizations P of the graph G
(For almost all choices of CPDs for the variables)
The d-separation test precisely characterizes the independencies that hold for P.
מה שאנחנו לומדים מזה זה שמבחן ה-D-Separation מייצג את תכונות האי תלות של ההתפלגות באופן נאמן.

3.3.3 Linear algorithm for d-separation.

לצערי לא הוצג בכיתה.

Algorithm 3.1
G, X, Z -> Nodes reachable from X given Z in G.

אלגוריתם נאיבי ימנה את כל המסלולים שיוצאים מ-X לכל Y אפשרי בגרף ולא עוברים ב-Z ועבור כל מבנה V- יבדוק אם יש לו צאצא ב-Z. אולם זה יכול להיות אקספ' בגודל הגרף.
כאמור, לא מחפשים סתם מסלולים או אפילו סתם מסלולים על גרף הבסיס, אלא מסלולים שלא חוסמים השפעה, כלומר, או שהם לא נתונים, או שהם ממבנה של Common-Effect ואז הם או אחד הצאצאים שלהם נתון.

ראינו ש-d הפרדה אמ"מ כל המבנים של Common-Effect נתונים והם המבנים היחידים שנתונים.
אלגוריתם:
1. לכל צומת מ-Z, סמן את כל האבות שלו בתור צמתים פעילים עבור מבנה של Common-Effect.
2. בצע BFS החל מ-X תוך התעלמות מכיווני הקשתות.
שמור בפלט כל צומת אליו הגעת שאינו ב-Z. אם הגעת לצומת ב-Z, עצור. אם לא: הוסף את כל האיברים המושפעים מהצמתים אליהם הגעת.

Run example over Figure 3.6

בנייה: גרף עם הצמתים:

W -> Y
Z -> Y
Z -> X

Y -> X
X | Y ??

אם הריצה תתחיל מ-X ל-Y ורק אז X ל-Z ל-Y ל-W.

פתרון: במהלך ה-BFS התנהג אחרת אם הגעת לצומת מבן או מאב.
דוגמא 3.4 ואיור 3.6

I-Equivalence

Definition

G is I-equivalent to K ⇔ I(G)=I(K).
עכשיו שאנחנו מרגישים בנוח עם ההגדרה של I-Map ועם הקשר בין גרף למפת האי-תלות-המותנית שלו, ניתן להגדיר שקילות.

Motivation

יש לנו את הכלי החזק של ה-I(G). נבחין כי לגרפים שונים יכולים להיות מיפויי אי-תלויות זהים. ניתן לדבר על אי התלויות ולשכוח מהגרף!

כמו כן, כמו כל שקילות, הרי שI-Equivalence מחלקת את מרחב הגרפים מעל n משתנים למחלקות שקילות, כך שכל איבר שייך בדיוק למחלקת שקילות יחידה.
* From P to Equivalence class.
מהשקילות בין ה-I-map ל-Factorization נובע שאם התפלגות P היא מעל גרף כלשהו- אזי היא גם מעל כל גרף ששקול לו.
אין שום תכונה של P שגורמת לנו לבחור בגרף מסויים ע"מ לייצג אותה ולא בגרף אחר.
לתכונה זו יש השלכות חשובות, למשל, בנוגע ליכולת שלנו לקבוע כיווניות של השפעה.
(יכול להיות שאיכות המכתב הכתיבה את קושי הקורס?...)

קיבלנו את הטענה שמה שמעניין התפלגות מסויימת זו מחלקת השקילות של הגרפים מעליה היא מפולגת ולא הגרף הספציפי.
כעת ננסה לאפיין מחלקת שקילות.
נחפש תכונות הכרחיות לשקילות בין גרפים.

Skeleton

קל להיווכח שאם שני גרפים לא חולקים את אותו שלד, כלומר, ב-G יש קשת X -> Y שב-K אין, אזי ניתן לבנות דוגמא שבה, באמצעות כיווני הקשתות ב-K אין מסלול השפעה בין X ל-Y וב-G יש.
למשל, אם ב-K יש מסלול מ-X ל-Y אזי עדיין, בהינתן ההורים שלו, Y בת"ל ב-X. לא כך ב-G.

מאידך, כפי שראינו מאיור 3.5,

X -> Z <- Y
X -> Z -> Y

חולקים את אותו שלד אך ה-I-Map שלהם שונה. Y|Z תלוי ב-X באחד ולא כך בשני.

V-Structure

השתכנענו שאם לשני גרפים יש שלד שונה אז הם לא שקולים,
השתכנענו שאם לשני גרפים יש שלד דומה- זה לא מספיק.
האם שלד דומה ואוסף מבני V זהים זה מספיק?

[הבהרת מבני V?]

Theorem 3.7, If G and K have the same skeleton and same set of v structures then they are I-Equivalent

נשמע סביר. האם זה מספיק?
נראה שמדובר בכלל חזק מדי.
אז פה עוברים לדוגמא של גרפים (DAGים) מלאים.
בהכרח אם G ו-K הם גרפים מלאים על, למשל, n=4, אבל שונים: אזי יש מבנה V כלשהו ששונה ביניהם.
מאידך, שניהם מקודדים את אותו אוסף התלויות (אוסף אי התלויות הריק… בין כל שני משתנים יש יחס הורות.)

אינטואיטיבית, מה שאנחנו מחפשים זו התכונה ש-X ו-Y בת"ל, אך תלויים בהינתן Z.

Definition 3.11: Immorality

A v-structure X -> Z <- Y is an Immorality if there is no direct edge between X and Y.
If there is such an edge, we call it a ‘covering edge’.
נבחין כמובן כי לא כל מבנה V הוא אימוראליטי, למשל, בגרף המלא אין שום אימוראליטי.

לי, לפחות, הפריעה ההגדרה הזאת:
למה אי אפשר שיהיה

X -> Z <- Y , Y -> R -> X

ואז אין קשת מכסה אבל יש מסלול מכסה?
אז העניין עם קשת מכסה הוא שבגללה בטוח יש השפעה.
אם יש רק מסלול מכסה Y ל-R ל-X, אזי בהינתן R: אנחנו חוזרים למצב של מבנה V, כלומר בהינתן רק R יתקיים ש-X ו-Y בת"ל אולם בהינתן R ו-Z יתקיים ש-X ו-Y תלויים!

אז העניין עם Immorality היא אי-תלות מותנית באבות אבל תלות מותנית בצאצא.
$X \perp Y \mid R$ as well as $X \perp Y \mid R,Z$
Whereas
$X \perp Y$ but $\neg X \perp Y|Z$.

The. 3.8: I-Equivalence

G and K are I-Equivalent iff they have the same skeleton and set of Immoralities.

3.4 From Distribution to Graph.

לצערי בשלב זה דילגנו לסוף.

Motivation

ראינו שאם P מתפלגת על G אזי ניתן להסיק את אוסף אי התלויות בין משתני P, רק מבחינת G.
היינו רוצים, בהינתן P, להתחיל מלבנות גרף ואז לנסות ללמוד ממנו על P.

כעת נבחן את השאלה, בהינתן P, עד לאיזו רמה ניתן לבנות G שמדמה את P?
העניין הוא שבד"כ לא נקבל P התפלגות מלאה ומפורטת: כפי שאמרנו, התפלגות מלאה זה אובייקט גדול ומפורט מדי.

לפיכך, הפרק הנוכחי הוא בגדר תרגיל מחשבתי, שיקל עלינו בבואנו לבנות רשת באייז שתתאר מודל של העולם.

Naive Approach

נבחר גרף שרירותי ש-P מתפלגת עליו.
ובכן, הגרף המלא הוא I-Map לכל התפלגות שהיא, אבל הוא לא מלמד אותנו שום דבר לגבי אי התלויות בהתפלגות.

3.4.1 Minimal I-Maps

הניסיון הבא שלנו הוא הגדרה של מינימום.
* A graph K is a minimal I-Map for a set of independencies I if it is an I-Map for I, and the removal of even a single edge from K- renders it not an I-map.
נבחין שזו הגדרה כללית ומועילה: ניתן לבחור בתור ה-I הימני גם את ה-I-Map של גרפים. כמו כן, מובטח בספר שבהמשך ישתמשו בהגדרה דומה גם לתיאור עוד סוגים של גרפים.
כמו כן, מקובל להגביל את ההגדרה שניתנה קודם לרשת באייז לכך ש-G הוא I-Map מינימאלי של P (ולא רק I-Map כללי, כמובטח מזה ש-P מתפלגת על G).

Minimal I-Map is not unique

אם נחשוב למשל על התפלגות על פני A, B, C כך ש-A=B וכן יש מתאם בין A ל-C.
ברור שכל מבנה בעל שתי קשתות: A משפיע על B או להיפך וכן A או B משפיעים על C יהיה בגדר Minimal I-Map של P.

Minimal I-Map Fails

ברמה האינטואיטיבית, הבעיה עם המונח של I-Map מינימאלי היא שמדובר במינימום לוקאלי ולא מובטח שאין מינימומים קטנים יותר.

Minimal I-Map Algorithm

1. For X from i to n,
2. Find the minimal subset U ⊆ {1...i-1} s.t X ⊥ {i...i-1 \ U} | U, and set them to be X_i ‘s parents.

הוכחת הנכונות של האלגוריתם הזה היא כמו ההוכחה ש-P מתפלגת על G.
לפי הוכחת משפט 3.1, P מתפלגת על הגרף שחוזר מן האלגוריתם הזה. לפי משפט 3.2, הגרף שחוזר הוא I-Map של הגרף הזה.
לפי הבנייה, G הוא I-Map מינימאלי (הוספנו קשתות שהיינו חייבים).

Algorithm 3.2 Procedure to build a minimal I-map given an ordering

Procedure Build-Minimal-I-Map (
$X_1, \ldots, X_n$ // an ordering of random variables in $\mathcal{X}$
$\mathcal{I}$ // Set of independencies
)

Set $\mathcal{G}$ to an empty graph over $\mathcal{X}$
for $i = 1, \ldots, n$
$U \leftarrow \{X_1, \ldots, X_{i-1}\}$ // U is the current candidate for parents of X_i
for $U' \subseteq \{X_1, \ldots, X_{i-1}\}$
if $U' \subset U$ and $(X_i \perp \{X_1, \ldots, X_{i-1}\} - U' \mid U') \in \mathcal{I}$ **then**
$U \leftarrow U'$
// At this stage U is a minimal set satisfying $(X_i \perp \{X_1, \ldots, X_{i-1}\} - U \mid U)$
// Now set U to be the parents of X_i
for $X_j \in U$
Add $X_j \rightarrow X_i$ to $\mathcal{G}$
return $\mathcal{G}$

I-Map Algorithm over the Student example.

קודם כל, נבחין שהרשת שבנינו בתחילת הפרק היא רשת מינימאלית, למשל, היא חוזרת מהאלגוריתם בהינתן הסידור הטופולוגי של עצמה.
כעת ננסה סידור פחות אינטואיטיבי.
להתחיל מ-L, דרך, S, ואז G ואז I ואז D.

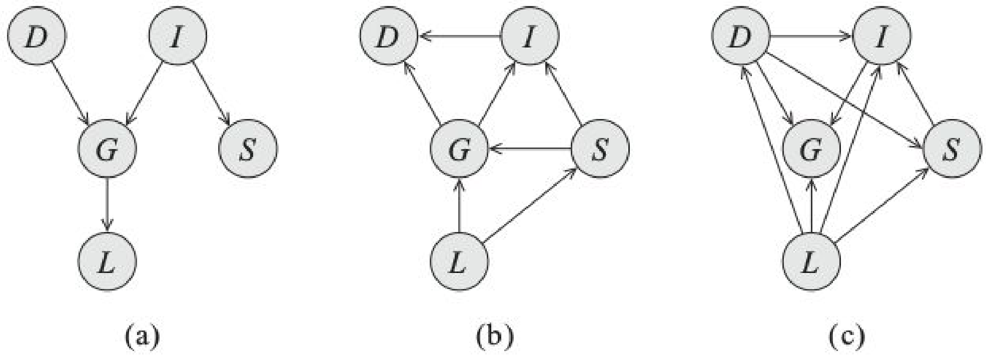


Figure 3.8 Three minimal I-maps for $P_{\mathcal{B}^{student}}$, induced by different orderings: (a) D, I, S, G, L; (b) L, S, G, I, D; (C) L, D, S, I, G.

Minimal I-Maps Conclusion

לסיכום, Minimal I-Map זה מושג שמגדיר מינימום לוקאלי, ולא מבטיח לתפוס את כל אי התלויות שבהתפלגות.

3.4.2 Perfect Maps

We aim to find a graph G that precisely captures the independencies in a given distribution P.
עודנו מחפשים גרף שתופס באופן מדוייק את אי התלויות בהתפלגות P.

Definition 3.14: Perfect Map

We say that G is a P-Map for a set of independencies I if we have that I(G) = I.
נאמר ש-G הוא מיפוי מושלם של I אם אוסף האי-תלויות-המותנות שלהם זהה.
שינינו את דרישת ההכלה לדרישה של שוויון.

Perfect Map Fails: Regularity

הבעיה העיקרית שמובאת בספר בנוגע למיפויים מושלמים היא שלא תמיד ניתן למצוא כאלה.
* P x,y,z = 1/12 iff x \xor y \xor z ; 1/6 otherwise.
מה שקורה כאן זה ש-X תלוי ב-Z בהינתן Y, אבל אם לא יודעים כלום על Y אז X לא משפיע על Z.
כלומר,

X ⊥ Y

אז ניסיון לבנות I-Map מינימאלי יתן לנו מבנה שבו נוסיף קודם את X, ואז את Y.
בבואנו להוסיף את Z, אי אפשר לטעון ש-Z לא תלוי ב-Y בהינתן X, וכן אי אפשר לטעון ש-Z לא תלוי ב-X בהינתן Y. על כן Y ו-X יהיו ההורים של Z בגרף.
אבל קל לראות ש-Z ב"ת ב-Y באופן ישיר, ועובדה זו לא מבוטאת בגרף שבנינו.

Perfect Map Fails, Cases where BN is simply not appropriate

נראה עוד דוגמא להתפלגות שעבורה אין מיפוי מושלם.

כשדיברנו על מגבלות המודל דיברנו על זה שאנו מתארים תלויות במשתנים תוך התעלמות מן הערכים שהם עלולים לקבל.
למשל,

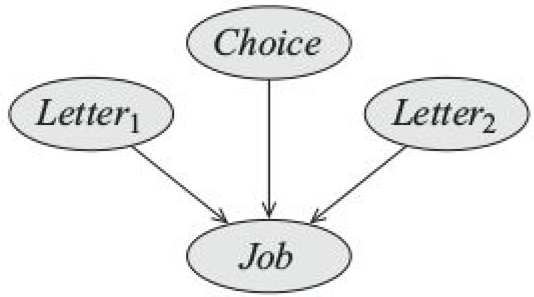


Figure 3.9 Network for the OneLetter example

למשל J כתלות ב-C.
אם לג'ורג' יש שני מכתבים, אחד מכל מרצה, אולם הוא מתבקש להציג רק אחד. (ונניח שהוא פשוט מטיל מטבע כדי להחליט.)
קודם כל, אפשר להגות בכך שברגע שהוא בוחר מכתב כלשהו- ההשפעה של המכתב השני נעלמת.
כמו כן, הרי שהמכתבים ב"ת בהינתן הבחירה, והשאלה האם הוא קיבל את העבודה.
(והם אינם ב"ת בגרף, שכן יש Immorality בינם לבין ה-J.)
ניתן לחשוב על כך אם נפרוט למקרים: אם C=1, הקשת בין L2 ל-J כאילו אמורה להיעלם.

Perfect Map: Conclusion

לסיכום, מיפוי מושלם הוא אידיאל, לעיתים בר השגה. בהגדרה של BN נדרוש מיפוי מינימאלי ונקווה למיפוי מושלם.

Box 3.C: Skill: Knowledge Engineering

מה שראינו זה את המעבר מהתפלגות משותפת לרשת בייז.
החיים האמיתיים יותר מסובכים.
יש לנו מושג עמום לגבי המציאות ואנו מנסים להשתמש בו כדי לבנות רשת.
לרוע המזל, לטעויות במודל עלולות להיות השלכות חמורות על טיב התוצאות, או למחיר החישובי של שימוש ברשת.

Picking Variables

האתגר הראשון בבואנו לתאר בעיה כלשהי באמצעות רשת באייז היא בבחירת המשתנים.
בחירה במשתנה כמו "ירד גשם" נשמעת תמימה, אולם לא ברור באיזה סוג של גשם מדובר, וכמה זמן, ומתי.
* Clarity Test

בספר הם מציעים את מבחן הבהירות.
הם אומרים, האם בהינתן ידע אין-סופי ניתן יהיה לבחון את העבר ולקבוע האם המשתנה התקיים או לא?
האם ירד גשם לאורך לפחות עשר דקות במהלך הלילה?

* Relevance Test

בעיה נוספת היא כמה משתנים להוסיף.
הדוגמא היחידה שאביא מהספר היא המשתנה B, מס' הבירות שג'ורג' שתה ערב הפסיכומטרי.
ברור שזה משפיע על איכות המבחן, אולם ניתן להניח שכבר התחשבנו במשתנה זה במהלך הגדרתנו את ה-CPD של S.
כמו כן, אפשר להתחשב בטווח הערכים שמאפשרים למשתנים. ציון פסיכומטרי יכול לקבל כל ערך בין 200 ל-800 . מכתב המלצה יכול להיות בכל אורך ולכלול כל סופרלטיב אפשרי.
קידדנו אותם באופן בינארי ופשוט ובכך הפכנו את המודל לשימושי.

Picking Structure

מסתבר שלתת מבנה שעקבי עם תפיסת הסיבתיות שלנו יעילה.
בספר אומרים, כנראה בגלל המקומיות של ההשפעה.
למשל ברשת של ביטוח רכב, חברת הביטוח תופסת את הבעיה בתור "תאונות קודמות" -> ,נהג טוב". כך הם חושבים על הבעיה, אולם במציאות, קל יותר לדמיין נהג טוב משפיעה על תאונה קודמת.
כמו שראינו בנושא של חיפוש מפה מינימאלית, סידור בעייתי של המשתנים עלול להוסיף לנו קשתות מיותרות בגרף.

* Butterfly Effect

בבואנו לתכנן מודל, אין לנו אלא להתעלם מהשפעות חלשות בין משתנים.
בדוגמא של ג'ורג' הסטודנט, יכול להיות שיש קשר בין קושי הבחינה למידת האינטילגנציה: המרצה ראה שג'ורג' ניגש ויודע שהוא חכם אז הוא מקשה את הקורס. או שהוא מכיר את ג'ורג' באופן אישי ובוחר להקל עלי.

אם נכניס את כל ההשפעות- נאבד מהערך של המודל כי יהיה קשה ואיטי להשתמש בו וכי יהיה קשה יותר לבדבג אותו.

אפקט הפרפר אמיתי ועל כן תמיד נהיה חייבים להתעלם מהשפעות צדדיות, חלשות יותר.

Picking Probabilities

אחד השלבים הקשים בבניית רשת באייס היא בניית ה-CPD.
אנשים לא אוהבים להתחייב למידה מדוייקת של הסתברות.
בעבר ניסו להשתמש בתיאורים, כמו "סביר מאוד", "לא סביר", אולם התברר שאצל חלק מהאנשים "לא סביר" זה 50% בעוד שאצל אחרים זה 5%.
כמו כן, מתברר ששימוש בשפה טבעית הטעה אנשים.
דוגמא בספר ניתן להשוות לדוגמא כמו, "הפועל ת"א תזכה באליפות"- 0, לעומת "הפועל ת"א תנצח במשחק הראשון", "הפועל ת"א תנצח במשחק השני", וכו'.

עד כמה זה חשוב לדייק בהערכות?
בספר אומרים שההבדל בין 70 ל-75% זניח. הם נותנים את הטעויות החשובות הבאות:

* 0% probability

הם אומרים בספר שיש לאנשים נטייה לייחס 0% סיכוי לאירועים שכן יכולים לקרות.

* Order of Magnitude

לאנשים קשה להעריך את ההבדל בין 10 במינוס 4 ל-10 במינוס 5, אולם יש לכך השפעות גדולות על התוצאות מהמודל.

Conclusion [terms]

Key terms Covered

CPD,
Chain Rule,
Intercausal Reasoning.
I-Map
D-Separation
I-Equivalence
Minimal I-Map
Perfect Map

Home work

יועלו בנפרד.

Bravely Skipped

נושאים שהושמטו מלכתחילה. למרחיקי לכת, לקריאה בספר:

Dummy/Naive Bayesian Network
3.4.3 Finding perfect maps.