Machine Learning: Foundations

Lecturer: Yishay Mansour

Lecture 2 – Bayesian Inference

Kfir Bar

Yaniv Bar

Marcelo Bacher

Based on notes by Shahar Yifrah, Keren Yizhak, Hadas Zur (2012)

Bayesian Inference

- Bayesian inference is a method of statistical inference that uses prior probability over some hypothesis to determine the likelihood of that hypothesis be true based on an observed evidence.

- Three methods:
  - ML - Maximum Likelihood rule
  - MAP - Maximum A Posteriori rule
  - Bayes Posterior rule

Bayes Rule

In general: $$\Pr[A|B] = \frac{\Pr[B|A] \cdot \Pr[A]}{\Pr[B]}$$

- In Bayesian inference:
  - *data* - a known information
  - *h* - an hypothesis/classification regarding the data distribution

We use Bayes Rule to compute the likelihood that our hypothesis is true:

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

Example 1: Cancer Detection

- A hospital is examining a new cancer detection kit.
  The known information (prior) is as followed:
  - A patient with cancer has a 98% chance for a positive result
  - A healthy patient has a 97% chance for a negative result
  - The Cancer probability in normal population is 1%

How reliable is this kit?

Example 1: Cancer Detection

- Let's calculate Pr[cancer|+]:

$$\Pr[+|cancer] = 0.98$$
$$\Pr[+|\neg cancer] = 0.97$$
$$\Pr[cancer] = 0.01$$

According to Bayes rule we get:

$$\Pr[cancer|+] = \frac{\Pr[+|cancer] \cdot \Pr[cancer]}{\Pr[+]}$$

$$\Pr[+] = \Pr[+|\text{cancer}] \cdot \Pr[\text{cancer}] + \Pr[+|\neg\text{cancer}] \cdot \Pr[\neg\text{cancer}]$$

$$= 0.01 \cdot 0.98 + 0.99 \cdot 0.03 = 0.0098 + 0.0297 = 0.0395$$

$$\Rightarrow \Pr[cancer|+] = \frac{0.98 \cdot 0.01}{0.0395} = 0.248 \approx 25\%$$

Example 1: Cancer Detection

- Surprisingly, the test, although it seems very accurate, with high detection probabilities of 97-98%, is almost useless

- 3 out of 4 patients found sick in the test, are actually not. For a low error, we can just tell everyone they do not have cancer, which is right in 99% of the cases

- The low detection rate comes from the low probability of cancer in the general population = 1%

Example 2: Normal Distribution

- A random variable Z is distributed normally with mean μ and variance σ^2

$$Z \sim N(\mu, \sigma^2)$$

$$\mu, \sigma \sim N(0,1)$$

Recall -

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

$$E[Z] = \mu$$

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

Example 2: Normal Distribution

- We have *m* i.i.d samples of a random variable *Z*

$$p[(\mu,\sigma)|z_1,z_2,\dots,z_m] = \frac{p[z_1,z_2,\dots,z_m|(\mu,\sigma)]\cdot p[(\mu,\sigma)]}{p[z_1,z_2,\dots,z_m]}$$

$$p[z_1,z_2,\dots,z_m|(\mu,\sigma)] = \prod_{i=1}^{m}\frac{1}{\sqrt{2\pi\sigma^2}}\cdot e^{-\frac{1}{2}(\frac{z_i-\mu}{\sigma})^2}$$

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

where $p[z_1,z_2,\dots,z_m]$ is a normalization factor

Example 2: Normal Distribution

- Maximum Likelihood (ML):

We aim to choose the hypothesis which best explains the sample, independent of the prior over the hypothesis space (the parameters that maximize the likelihood of the sample)

$$max_{h_i \in \mathcal{H}} Pr[D|h] \quad \text{where } D = Data$$

in our case -

$$ML = max_{\mu,\sigma} p[z_1, \ldots, z_m|(\mu, \sigma)] = max_{\mu,\sigma} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}(\frac{z_i-\mu}{\sigma})^2}$$

Example 2: Normal Distribution

- Maximum Likelihood (ML):

We take a log to simplify the computation -

$$L_{ML} = \log ML = \sum_{i=1}^{m} -\frac{1}{2}\left(\frac{z_i - \mu}{\sigma}\right)^2 - \frac{m}{2}\log 2\pi - m\log\sigma$$

now we find the maximum for μ :

$$\frac{\partial}{\partial x} L_{ML} = \sum_{i=1}^{m} \frac{1}{\sigma}\left(\frac{z_i - \mu}{\sigma}\right) = 0$$

$$\sum_{i=1}^{m} z_i = m \cdot \mu$$

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} z_i$$

It's easy to see that the second derivative is negative, thus it's a maximum

Example 2: Normal Distribution

- Maximum Likelihood (ML):
  - Note that this value of μ is independent of the value of σ and it is simply the average of the observations
  - Now we compute the maximum for σ given that μ is $\hat{\mu}$:

$$\frac{d}{dx}L_{ML} = \sum_{i=1}^{m} \frac{(z_i - \hat{\mu})^2}{\sigma^3} - \frac{m}{\sigma} = 0$$

$$\sum_{i=1}^{m} (z_i - \hat{\mu})^2 = m \cdot \sigma^2$$

$$\hat{\sigma}^2 = \frac{1}{m} \cdot \sum_{i=1}^{m} (z_i - \hat{\mu})^2$$

Example 2: Normal Distribution

- Maximum A Posteriori (MAP):

  MAP adds the priors to the hypothesis. In this example, the prior distributions of μ and σ are $N(0,1)$ and are now taken into account

$$max_{h_i \in \mathcal{H}} Pr[h_i|D] = max_{h_i \in \mathcal{H}} \frac{\Pr[D|h_i] \cdot \Pr[h_i]}{\Pr[D]}$$

  and since $Pr[D]$ is constant for all $h_i \in \mathcal{H}$ we can omit it, and have the following:

$$MAP = max_{\mu,\sigma} \prod \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{z_i-\mu}{\sigma}\right)^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{\mu^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{\sigma^2}{2}}$$

Example 2: Normal Distribution

- Maximum A Posteriori (MAP):
  - How will the result we got in the ML approach change for MAP? We added the knowledge that σ and μ are small and around zero, since the prior is $\sigma,\mu \sim N(0,1)$
  - Therefore, the result (the hypothesis regarding σ and μ) should be closer to 0 than the one we got in ML

$$L_{MAP} = \log MAP = \sum -\frac{1}{2}\left(\frac{z_i - \mu}{\sigma}\right)^2 - \frac{1}{2}m\log 2\pi - m\log\sigma - \frac{1}{2}\log 2\pi - \frac{1}{2}\mu^2$$

$$-\frac{1}{2}\log 2\pi - \frac{\sigma^2}{2}$$

$$\frac{d}{d\mu}L_{MAP} = \sum_{i=1}^{m}\frac{z_i - \mu}{\sigma^2} - \mu = 0$$

$$\frac{d}{d\sigma}L_{MAP} = \sum\frac{z_i - \mu}{\sigma^3} - \frac{\mu}{\sigma} - \sigma = 0$$

Example 2: Normal Distribution

- Maximum A Posteriori (MAP):

Now we should maximize both equations simultaneously:

$$\frac{1}{m}\sum_{i=1}^{m} z_i = \hat{\mu}\left(\frac{\hat{\sigma}^2}{m} + 1\right)$$

$$\frac{1}{m}\sum_{i=1}^{m} (z_i - \hat{\mu})^2 = \hat{\sigma}^2\left(\frac{\hat{\sigma}^2}{m} + 1\right)$$

it can be easily seen that μ and σ will be closer to zero than in the ML approach, since

$$\frac{\hat{\sigma}^2}{m} > 0$$

Example 2: Normal Distribution

- Posterior (Bayes):

Assume μ~*N*(η,1) and *Z*~*N*(μ,1) and σ = 1.
We see only one sample of *Z*. What is the new posterior distribution of μ?

Pr[*Z*] is a normalizing factor, so we can drop it for the calculations:

$$\Pr[\mu] = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\mu-\eta)^2}$$

$$\Pr[Z|\mu] = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(Z-\eta)^2}$$

$$\Pr[Z|\mu] = \Pr[\mu] \cdot \Pr[Z|\mu]$$

Example 2: Normal Distribution

- Posterior (Bayes):

$$\Pr[\mu|Z] = \Pr[\mu] \cdot \Pr[Z|\mu]$$
$$\propto \exp\left\{-\frac{1}{2}(\mu^2 - 2\eta\mu + \eta^2) - \frac{1}{2}(Z^2 - 2Z\mu + \mu^2)\right\}$$
$$= \exp\left\{-\frac{1}{2}(2\mu^2 - 2\mu(\eta + Z) + \eta^2 + Z^2)\right.$$
$$= \exp\left\{\left(\mu - \frac{\eta + Z}{2}\right)^2 + \left(\frac{\eta + Z}{2}\right)^2 + \eta^2 + Z^2\right\}$$

$\left(\frac{\eta + Z}{2}\right)^2 + \eta^2 + Z^2 =$ normalization factor, does not depend on μ

Example 2: Normal Distribution

- Posterior (Bayes):

The new posterior distribution $N\sim(\hat{\mu}, \hat{\sigma}^2)$ has:

$$\hat{\mu} = \frac{\eta + Z}{2}$$

$$\hat{\sigma}^2 = \frac{1}{2}$$

after taking into account the sample z, μ moves towards Z and the variance is reduced

Example 2: Normal Distribution

- Posterior (Bayes):

In general, for:

$$\mu \sim (\eta, S^2) \text{ and } Y \sim (\mu, \sigma^2)$$

given m samples $y_1, \cdots, y_m$ we have:

$$\hat{\mu} = \frac{\frac{1}{S^2}\eta + \frac{m}{\sigma^2}\bar{y}}{\frac{1}{S^2} + \frac{m}{\sigma^2}}$$

$$\hat{\sigma}^2 = \left(\frac{1}{S^2} + \frac{m}{\sigma^2}\right)^{-1}$$

Example 2: Normal Distribution

- Posterior (Bayes):

And if we assume $S=\sigma$, we get:

$$\hat{\mu} = \frac{\eta + \sum_{i=0}^{m} y_i}{m+1}$$

$$\hat{\sigma}^2 = \frac{\sigma^2}{m+1}$$

which is like starting with an additional sample of value μ, i.e.,

$$y_0 = \mu$$

Learning A Concept Family (1/2)

- We are given a Concept Family H.
- Our information consist of examples $< x, f(x) >$, where $f \in H$ is an unknown target function that classifies all samples.
- Assumptions:

  (1) The functions in H are deterministic functions ($\Pr[h(x) = 1] = \{1,0\}$).

  (2) The process that generates the input is independent of the target function *f*.
- For each $h \in H$ we will calculate $\Pr[S \mid h]$ where $S = \{< x_i, b_i >, 1 \le i \le n\}$ and $b_i = f(x_i)$.

  Case 1: $\exists_i : b_i \neq h(x_i) \Rightarrow \Pr[< x_i, b_i > \mid h] = 0 \Rightarrow \Pr[S \mid h] = 0$

  Case 2: $\forall_i : b_i = h(x_i) \Rightarrow \Pr[< x_i, b_i > \mid h] = \Pr[x_i] \cdot \Pr[b_i \mid h, x_i] = \Pr[x_i]$

  $$\Rightarrow \Pr[S \mid h] = \prod_{i=1}^{m} \Pr[x_i] = \Pr[S]$$

Learning A Concept Family (2/2)

- Definition: A **consistent function** $h \in H$ classifies all the samples S correctly ($\forall_{<x_i,b_i>\in S} h(x_i) = b_i$).
- Let $H' \subseteq H$ be all the functions that are consistent with S.

  There are three methods to choose H':

  ML - choose any consistent function, each one has the same probability.

  MAP - choose the consistent function with the highest prior probability.

  Bayes - weighted combination of all consistent functions to one predictor,

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

Example (Biased Coins)

- We toss a coin *n* times and the coin ends up heads k times.
- We want to estimate the probability *p* that the coin will come up heads in the next toss.
- The probability that *k* out of *n* coin tosses will come up heads is:

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

- With the Maximum Likelihood approach, one would choose *p* that maximize $\Pr[(k,n) \mid p]$ which is $p = \frac{k}{n}$.
- Yet this result seems unreasonable when *n* is small.

  (For example, if you toss the coin only once and get a tail, should you believe that it is impossible to get a head on the next toss?)

Laplace Rule (1/3)

- Let us suppose a uniform prior distribution on p. That is, the prior distribution on all the possible coins is uniform,

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

- We will calculate the probability to see k heads out of n tosses.

$$\Pr[(k,n)] = \int_0^1 \Pr[k \mid p] \cdot \Pr[p] dp = \int_0^1 \binom{n}{k} x^k (1-x)^{n-k} dx \underset{\substack{Integraion \\ by \\ parts}}{=}$$

$$\left[\binom{n}{k} \frac{x^{k+1}}{k+1} (1-x)^{n-k} \right]_0^1 + \int_0^1 \binom{n}{k} \frac{x^{k+1}}{k+1} (n-k)(1-x)^{n-k-1} dx \underset{\binom{n}{k}\frac{(n-k)}{k+1}=\binom{n}{k+1}}{=}$$

$$\int_0^1 \binom{n}{k+1} x^{k+1} (1-x)^{n-k-1} dx = \int_0^1 \Pr[k+1 \mid p] \cdot \Pr[p] dp = \Pr[(k+1,n)]$$

Laplace Rule (2/3)

- Comparing both ends of the above sequence of equalities we realize that all the probabilities are equal, and therefore, for any *k*,

$$\Pr[(k,n)] = \frac{1}{n+1}$$

  Intuitively, it means that for a random choice of the bias *p*, any possible number of heads in a sequence of *n* coin tosses is equally likely.

- We want to calculate the posterior expectation, $E[p \mid s(k,n)]$ where $s(k,n)$ is a specific sequence with *k* heads out of *n*.

- We have,

$$\Pr[s(k,n) \mid p] = p^k (1-p)^{n-k}$$

$$\Pr[s(k,n)] = \int_0^1 p^k (1-p)^{n-k}\, dp = \frac{1}{n+1} \cdot \frac{1}{\binom{n}{k}}$$

Laplace Rule (3/3)

- Hence,

$$E[p \mid (k,n)] = \int_0^1 p \cdot \frac{\Pr[s(k,n) \mid p] \cdot \Pr[p]}{\Pr[s(k,n)]}\, dp = \frac{\int_0^1 p \cdot p^k (1-p)^{n-k}\, dp}{\frac{1}{n+1} \cdot \frac{1}{\binom{n}{k}}} =$$

$$\frac{\frac{1}{n+2} \cdot \frac{1}{\binom{n+1}{k+1}}}{\frac{1}{n+1} \cdot \frac{1}{\binom{n}{k}}} = \frac{k+1}{n+2}$$

- Intuitively, Laplace correction $\frac{k+1}{n+2}$ is like adding two samples to the ML estimator, one with value 0 and one with value 1.

Loss Function

- In lecture #1 we defined a few loss functions, among them was the logarithmic loss function. We will use it to compare our different approaches.

- When considering a loss function we should note that there are two causes for the loss:

  1. Bayes Risk - the loss we cannot avoid since we bound to have it even if we know the target concept.

  Example: Bias coin problem revisited.

  Even if we knew the bias p we would probably always predict 0 (if $p < \frac{1}{2}$) which, on the average, should result in $p \cdot n$ mistakes.

  2. Regret - the loss due to incorrect estimation of the target concept (having to learn an unknown model.)

Using Log Loss Function (1/3)

- A commonly use loss function is the LogLoss which states for the bias coin problem that if the learner guesses that the bias is *p* then the loss will be

  $\log \frac{1}{p}$ when the outcome is 1 (=head)

  $\log \frac{1}{1-p}$ when the outcome is 0 (=tail)

- If the true bias is θ then the expected LogLoss is $\theta \cdot \log \frac{1}{p} + (1-\theta) \cdot \log \frac{1}{1-p}$ which attains it's minima when $p = \theta$.
- Let's consider the loss at $p = \theta$, $H(\theta) = \theta \cdot \log \frac{1}{\theta} + (1-\theta) \cdot \log \frac{1}{1-\theta}$
- which is known in the Information Theory literature as the
- Entropy of , this is essentially the Bayes Risk.

Using Log Loss Function (2/3)

- Recall that we cannot do any better then $H(\theta)$ since Bayes Risk is the loss we cant avoid.
- How far are we from the Bayes Risk when using the guess of *p* according to the Laplace Rule ?

$$E[LogLoss] = \int_0^1 \sum_{n=1}^{T} \sum_{k=0}^{n} (\theta \cdot \log \frac{n+2}{k+1} + (1-\theta) \cdot \log \frac{n+2}{n-k+1}) \cdot \binom{n}{k} \theta^k (1-\theta)^{n-k} \, d\theta =$$

$$\sum_{n=1}^{T} \sum_{k=0}^{n} \binom{n}{k} \log \frac{n+2}{k+1} + \int_0^1 \theta \cdot \theta^k (1-\theta)^{n-k} \, d\theta + \sum_{n=1}^{T} \sum_{k=0}^{n} \binom{n}{k} \log \frac{n+2}{n-k+1} + \int_0^1 \theta^k (1-\theta)^{n-k} \, d\theta =$$

$$\sum_{n=1}^{T} \sum_{k=0}^{n} \left(\frac{1}{n+1} \cdot \frac{k+1}{n+2} \cdot \log \frac{n+2}{k+1} \right) + \left(\frac{1}{n+1} \cdot \frac{n-k+1}{n+2} \cdot \log \frac{n+2}{n-k+1} \right) =$$

$$\sum_{n=1}^{T} \sum_{k=0}^{n} \left(\frac{1}{n+1} \cdot H\left[\frac{k+1}{n+2} \right] \right) = T \int H[\theta] d\theta + \sum_{n=1}^{T} \frac{c}{n} = Bayes \ \ Risk \ + O(\log(T))$$

for some constant *c*.

Using Log Loss Function (3/3)

- In the above we used the fact that,

$$\sum_{i=1}^{n/2} \frac{1}{n} \cdot H(\frac{i-1}{n}) \leq \int_0^{1/2} H(\theta) d\theta \leq \sum_{i=1}^{n/2} \frac{1}{n} \cdot H(\frac{i}{n})$$

- The Difference between the upper and lower bound is:

$$\sum_{i=1}^{n/2} \frac{1}{n} \left(H(\frac{i}{n}) - H(\frac{i-1}{n}) \right) = \frac{1}{n} \left(H(\frac{1}{2}) - H(0) \right) = \frac{1}{n}$$

- Hence, we showed that by applying the Laplace Rule, we attained the optimal loss (the Bayes Risk) with an additional regret which is only logarithmic in the number of coin flips (T). Note that the Bayes Risk by itself grows linearly with T.

Naïve Bayes
Classification: Binary Domain (1/2)

Consider two binary classes +1 and -1, where each example is described with N attributes.

- X_n is a binary variable with possible values {0,1}. Example of dataset:

x_1	x_2	...	x_n	C
0	1		0	-1
1	1		0	+1
0	0		1	-1
0	1		1	+1
…	…		…	…
1	0		1	+1

We are looking for an hypothesis, *h*, to map $x_1, x_2, \ldots x_n$ to {-*1*,+*1*} such that:

$$\Pr(C = +1 \mid x_1, x_2, \ldots x_n) = \frac{\Pr(x_1, x_2, \ldots x_n \mid C = +1)\Pr(C = +1)}{\Pr(x_1, x_2, \ldots x_n)}$$

Naïve Bayes
Classification: Binary Domain (2/2)

The term $\Pr(C=+1)$ can be easily calculated from the data if it is not too large.

- Since Naive Bayes is based on independence assumption, therefore:

$$\Pr(x_1, x_2, \ldots x_n \mid C) = \prod^{N} \Pr(x_i \mid C)$$

- Each attribute x_i is independent on the other attributes once we know the value of C.
- For each $1 \leq i \leq n$ we have two parameters:

$$\theta_{i|+1} = \Pr(x_i = 1 \mid C = +1) \qquad \theta_{i|-1} = \Pr(x_i = 1 \mid C = -1)$$

- Assuming Simple Binomial estimation we can estimate these two parameters.
- Count the number of instances with $x_i = 1$ and with $x_i = 0$ among instances where $C = +1$ or $C = -1$, respectively. The convergence inequalities of Markov, Chebyshev and Chernoff can be used to bound deviations of the observed average from the mean (see the end of lecture 1 notes for more information.)

Naïve Bayes - Interpretation (1/2)

- According to Bayesian and MAP we need to compare two values:

$$\Pr(+1 \mid x_1, x_2, \ldots x_n) \text{ and } \Pr(-1 \mid x_1, x_2, \ldots x_n)$$

- We choose the most reasonable probability (maximum). By taking a Log of the fraction and comparing to 0.

$$\log \frac{\Pr(+1 \mid x_1, x_2, \ldots . x_n)}{\Pr(-1 \mid x_1, x_2, \ldots . x_n)} = \log \frac{\Pr(x_1, x_2, \ldots . x_n \mid +1) \Pr(+1)}{\Pr(x_1, x_2, \ldots . x_n \mid -1) \Pr(-1)}$$

$$= \log \frac{\Pr(+1)}{\Pr(-1)} + \log \prod_i^N \frac{\Pr(x_i \mid C = +1)}{\Pr(x_i \mid C = -1)}$$

$$= \log \frac{\Pr(+1)}{\Pr(-1)} + \sum_i \log \frac{\Pr(x_i \mid C = +1)}{\Pr(x_i \mid C = -1)}$$

- We conclude that:

$$\log \frac{\Pr(+1 \mid x_1, x_2, \ldots . x_n)}{\Pr(-1 \mid x_1, x_2, \ldots . x_n)} = \log \frac{\Pr(+1)}{\Pr(-1)} + \sum_i \log \frac{\Pr(x_i \mid C = +1)}{\Pr(x_i \mid C = -1)}$$

Naïve Bayes - Interpretation (2/2)

- Each *xi* "votes" about the prediction:

$$If\ \Pr(x_i \mid C=-1) = \Pr(x_i \mid C=+1) \text{ then } x_i \text{ has "no say" in classification}$$

$$If\ \Pr(x_i \mid C=-1) = 0 \text{ then } x_i \text{ overrides all other votes ("veto")}$$

- Let us denote:

$$\omega_i = \log \frac{\Pr(x_i = 1 \mid C=+1)}{\Pr(x_i = 1 \mid C=-1)} - \log \frac{\Pr(x_i = 0 \mid C=+1)}{\Pr(x_i = 0 \mid C=-1)}$$

$$b = \log \frac{\Pr(C=+1)}{\Pr(C=-1)} + \sum_i \log \frac{\Pr(x_i = 0 \mid C=+1)}{\Pr(x_i = 0 \mid C=-1)}$$

- The classification rule becomes a separating hyperplane:

$$h(x) = sign(b + \sum_i \omega_i x_i) = \text{if } \begin{cases} < 0 \text{ say -1 Class} \\ = 0 \text{ say either -1 Class or +1 Class} \\ > 0 \text{ say +1 Class} \end{cases}$$

Naïve Bayes – Practical Considerations

- Easy to estimate the parameters (each one has many samples)
- A relatively naive model
- Very simple to implement
- Reasonable performance (pretty often)

Naïve Bayes – Normal Distribution (1/3)

- Usually we also say Gaussian distribution

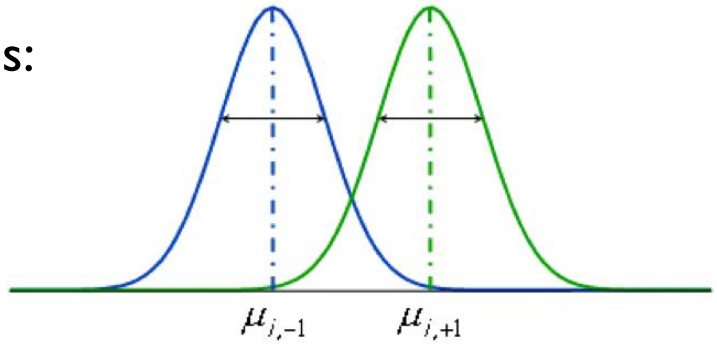
$$X \sim N(\mu,\sigma^2) \text{ if } p(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \qquad \Pr[a \le x \le b] = \int_a^b p(x)dx$$

$$E[x] = \mu \text{ and } Var[x] = E[(x - E(x))^2] = E[x^2] - E[x]^2 = \sigma^2$$

- We recall the independence assumption:

$$\Pr(x_1, x_2, \ldots x_n \mid C) = \prod_i^N \Pr(x_i \mid C)$$

- In addition, we make the following assumptions:
  - $\Pr(x_i \mid C) \sim N(\mu_{i,C}, \sigma^2{}_i)$
  - Mean of x_i depends on class, i.e., $\mu_{i,c}$.
  - Variance of x_i does not depend on class, i.e., σ_i.



Naïve Bayes – Normal Distribution (2/3)

- As before,

$$\log \frac{\Pr(+1 \mid x_1, x_2, \ldots . x_n)}{\Pr(-1 \mid x_1, x_2, \ldots . x_n)} = \log \frac{\Pr(+1)}{\Pr(-1)} + \sum_i \log \frac{\Pr(x_i \mid C = +1)}{\Pr(x_i \mid C = -1)}$$

- Using the Gaussian parameters,

$$\log \frac{\Pr(x_i \mid C = +1)}{\Pr(x_i \mid C = -1)} = \log \frac{\frac{1}{\sqrt{2\pi\sigma_i}} e^{-\frac{1}{2}\left(\frac{x_i - \mu_{i,+1}}{\sigma_i}\right)^2}}{\frac{1}{\sqrt{2\pi\sigma_i}} e^{-\frac{1}{2}\left(\frac{x_i - \mu_{i,-1}}{\sigma_i}\right)^2}}$$

$$= -\frac{1}{2}\left(\frac{x_i - \mu_{i,+1}}{\sigma_i}\right)^2 + \frac{1}{2}\left(\frac{x_i - \mu_{i,-1}}{\sigma_i}\right)^2$$

$$= -\frac{\mu_{i,+1} + \mu_{i,-1}}{\sigma_i} \frac{1}{\sigma_i}\left(\frac{\mu_{i,+1} + \mu_{i,-1}}{2} - x_i\right)$$

Distance between means

Distance of x_i to midway point

Naïve Bayes – Normal Distribution (3/3)

- If we allow different variances, the classification rule is more complex.

- The term $\log \frac{\Pr(x_i \mid C = +1)}{\Pr(x_i \mid C = -1)}$ is quadratic in x_i.