Machine Learning: Foundations **Fall Semester, 2012**

Lecture 2: October 28

Lecturer: Yishay Mansour *Scribe: Shahar Yifrah, Keren Yizhak, Hadas Zur*

2.1 Bayesian Inference - Overview

This lecture is going to describe the basic model of Bayesian Inference and its applications in Machine Learning. 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 being used in Bayesian inference:

1. **ML** - Maximum Likelihood rule
2. **MAP** - Maximum A Posteriori rule
3. **Bayes Posterior rule**

2.2 Bayes Rule

$$Pr\left[A|B\right] = \frac{Pr\left[B|A\right] \cdot Pr\left[A\right]}{Pr\left[B\right]} \tag{2.1}$$

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.

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

2.3 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%.

We wish to know how reliable the detection kit is. In other words, if a patient has a positive result, what is the probability that indeed he has cancer?
We want to compute $\mathbf{Pr}\left[\mathbf{cancer}|+\right]$

We know:

$$\begin{aligned} Pr\left[+|cancer\right] &= 0.98 \\ Pr\left[+|\neg cancer\right] &= 0.97 \\ Pr\left[cancer\right] &= 0.01 \end{aligned}$$

According to Bayes rule (2.1):

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

$$\begin{aligned} Pr\left[+\right] &= Pr\left[+|cancer\right] \cdot Pr\left[cancer\right] + Pr\left[+|\neg cancer\right] \cdot Pr\left[\neg cancer\right] \\ &= 0.01 \cdot 0.98 + 0.99 \cdot 0.03 = 0.0098 + 0.0297 \\ &= 0.0395 \end{aligned}$$

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

Surprisingly, the test, although it seems very accurate, with high detection probabilities of 97-98%, is almost useless. We have that 3 out of 4 patients found sick in the test, are actually not. If we want 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%.

2.4 Example 2: Normal Distribution

A random variable Z is distributed normally with mean μ and variance σ^2. I.e., $Z \sim N(\mu, \sigma^2)$, and $\mu, \sigma \sim N(0, 1)$. We have m i.i.d samples of a random variable Z.
Recall Normal distribution $N(\mu, \sigma^2)$:

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

Using Bayes rule:

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

where $p[z_1,z_2,\ldots,z_m]$ is a normalizing factor

We now discuss the three basic approached ML, MAP and general Bayesian Posterior:

2.4.1 ML: Maximum Likelihood

We aim to choose the hypothesis which best explains the sample, independent of the prior over the hypothesis space, i.e., the parameters which maximize the likelihood of the sample. Namely,

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

In our case

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

Take a Log (to simplify the computation):

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

Find the maximum for μ.

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

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

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

It's easy to see that the second derivative is positive, thus it's a maximum.
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 $\hat{\mu}$

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

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

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

where $\hat{\mu}$ was computed before.
Note, In this calculation we did not use the prior known distribution of μ or σ, only the Data.

2.4.2 MAP - Maximum A Posteriori

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.
We aim to maximize

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

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

$$MAP = \max_{\mu,\sigma} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{z_i - \mu}{\sigma})^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{\mu^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{\sigma^2}{2}}$$

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.

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

$$\frac{\partial}{\partial\mu}L_{MAP} = \sum_{i=1}^{m}\frac{z_i - \mu}{\sigma^2} - \mu = 0$$
$$\frac{\partial}{\partial\sigma}L_{MAP} = \sum_{i=1}^{m}\frac{z_i - \mu}{\sigma^3} - \frac{\mu}{\sigma} - \sigma = 0$$

Now we should maximize both equations simultaneously.

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

It can be easily seen that μ and σ will be closer to zero than in the ML approach, since $\frac{\hat{\sigma}^2}{m} > 0$.

2.4.3 Posterior (Bayes)

Assume $\mu \sim N(\eta, 1)$ and $Z \sim N(\mu, 1)$ (and the variance is known, $\sigma = 1$).
We see only one sample of Z. What is the new posterior distribution of μ?

$p[z]$ is a normalizing factor, so we can drop it for the calculations.

$$\begin{aligned}
p[\mu] &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\mu-\eta)^2} \\
p[z|\mu] &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-\mu)^2} \\
p[\mu|z] &= p[\mu] \cdot p[z|\mu] \\
&\propto \exp\{-\frac{1}{2}(\mu^2 - 2\eta\mu + \eta^2) - \frac{1}{2}(z^2 - 2z\mu + \mu^2)\} \\
&= \exp\{-\frac{1}{2}(2\mu^2 - 2\mu(\eta + z) + \eta^2 + z^2)\} \\
&= \exp\{-(\mu - \frac{\eta + z}{2})^2 + (\frac{\eta + z}{2})^2 + \eta^2 + z^2\}
\end{aligned}$$

$(\frac{\eta + z}{2})^2 + \eta^2 + z^2 =$ normalizing factor, does not depend on μ

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

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

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

In general, for:

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

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

$$\begin{aligned}
\hat{\mu} &= \frac{\frac{1}{S^2}\eta + \frac{m}{\sigma^2}\overline{y}}{\frac{1}{S^2} + \frac{m}{\sigma^2}} \\
\hat{\sigma^2} &= (\frac{1}{S^2} + \frac{m}{\sigma^2})^{-1}
\end{aligned}$$

If we assume $S = \sigma$ then:

$$\hat{\mu} = \frac{\eta + \sum_{i=0}^{m} y_i}{m+1}$$
$$\hat{\sigma^2} = \frac{\sigma^2}{m+1}$$

Which is like starting with an additional sample of value μ, i.e., $y_0 = \mu$.

2.5 Learning a Concept Family

We are given a Concept Family H. Our information consist of examples $\langle x, f(x)\rangle$, $f \in H$ unknown target function that all classifies all samples.
We assume that the functions in H are deterministic function, i.e., $Pr[h(x) = 1] = \{1, 0\}$. We will also assume that the process that generates the input is independent of the target function f. That means that the chosen points (x_i) alone contain no information on f (the target function).
For each $h \in H$ we will calculate $Pr[S|h]$ where $S = \{\langle x_i, b_i\rangle, 1 \le i \le n\}, b_i = f(x_i)$. We have the case first,

$$\exists_i : b_i \neq h(x_i) \Rightarrow Pr[\langle x_i, b_i\rangle |h] = 0 \Rightarrow Pr[S|h] = 0$$

Alternatively,

$$\forall_i : b_i = h(x_i) \Rightarrow Pr[\langle x_i, b_i\rangle |h] = Pr[x_i] \cdot Pr[b_i|h, x_i] = Pr[x_i]$$

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

A **consistent function** $h \in H$ classifies all the samples S correctly, i.e., $\forall_{\langle x_i, b_i\rangle \in S} h(x_i) = b_i$. Let $H' \subseteq H$ - all the functions consistent with S.

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.

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

2.6 Example 3: Biased Coins

In n coin tosses, a 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:

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

With the Maximum Likelihood approach, one would choose p that maximizes $Pr[(k,n)|p]$, 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?

2.6.1 Laplace Rule

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

$$\Pr[p \leq \theta] = \int_0^\theta dp = \theta.$$

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

$$\begin{aligned}
Pr[(k,n)] &= \int_0^1 \binom{n}{k} x^k (1-x)^{n-k} dx \\
&= \left[\binom{n}{k}\frac{x^{k+1}}{k+1}\cdot(1-x)^{n-k}\right]_0^1 + \int_0^1 \binom{n}{k}\frac{x^{k+1}}{k+1}(n-k)(1-x)^{n-k-1}dx \\
&= \int_0^1 \binom{n}{k+1} x^{k+1}(1-x)^{n-k-1}dx \\
&= Pr[(k+1|,n),
\end{aligned}$$

where the transition from the second to the third expression is due to the identity

$$\binom{n}{k}\frac{(n-k)}{k+1} = \binom{n}{k+1}.$$

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* $E[p|s(k,n)]$, where $s(k,n)$ is a specific sequence with k heads out of n. We have,

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

Hence:

$$\begin{aligned} E[p|(k,n)] &= \int_0^1 p \cdot \frac{Pr[s(k,n)|p] \cdot Pr[p]}{Pr[s(k,n)]} dp \\ &= \frac{\int_0^1 p \cdot p^k(1-p)^{n-k}dp}{\frac{1}{n+1} \cdot \frac{1}{\binom{n}{k}}} \\ &= \frac{\frac{1}{n+2} \cdot \frac{1}{\binom{n+1}{k+1}}}{\frac{1}{n+1} \cdot \frac{1}{\binom{n}{k}}} \\ &= \frac{k+1}{n+2}. \end{aligned}$$

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.

2.6.2 Loss Function

In the previous lecture we defined a few loss functions. We will now use one of them - the **logarithmic loss function** - 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. For example, consider the bias coin problem - 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.)

LogLoss Function - Reminder

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

$$\begin{array}{c}\log\frac{1}{p} \text{ when the outcome is 1 (head)}\\ \log\frac{1}{1-p} \text{ when the outcome is 0 (tail)}\end{array}$$

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 $p=\theta$ (as required). Consider the loss at $p=\theta$,

$$H\left[\theta\right]=\theta\cdot\log\frac{1}{\theta}+(1-\theta)\cdot\log\frac{1}{1-\theta},$$

which is known in the *Information Theory* literature as the *Entropy of* θ, this is essentially the **Bayes Risk**.

How far are we from the Bayes Risk when using the guess of p according to the Laplace Rule ? (We cannot do any better then $H\left[\theta\right]$ since Bayes Risk is the loss we cant avoid)

$$\begin{aligned}
E\left[\text{LogLoss}\right] &= \int_0^1\sum_{n=1}^{T}\sum_{k=0}^{n}\left[\theta\cdot\log\frac{n+2}{k+1}+(1-\theta)\cdot\log\frac{n+2}{n-k+1}\right]\cdot\left(\begin{array}{c}n\\k\end{array}\right)\theta^k(1-\theta)^{n-k}d\theta\\
&= \sum_{n=1}^{T}\sum_{k=0}^{n}\left(\begin{array}{c}n\\k\end{array}\right)\log\frac{n+2}{k+1}\int_0^1\theta\cdot\theta^k(1-\theta)^{n-k}d\theta+\\
&\quad\sum_{n=1}^{T}\sum_{k=0}^{n}\left(\begin{array}{c}n\\k\end{array}\right)\log\frac{n+2}{n-k+1}\int_0^1\theta^k(1-\theta)^{n-k}d\theta\\
&= \sum_{n=1}^{T}\sum_{k=0}^{n}\frac{1}{n+1}\frac{k+1}{n+2}\log\frac{n+2}{k+1}+\frac{1}{n+1}\frac{n-k+1}{n+2}\log\frac{n+2}{n-k+1}\\
&= \sum_{n=1}^{T}\sum_{k=0}^{n}\frac{1}{n+1}H\left[\frac{k+1}{n+2}\right]\\
&= T\int H\left[\theta\right]d\theta+\sum_{n=1}^{T}\frac{c}{n}\\
&= \text{Bayes Risk }+O\left(\log T\right)
\end{aligned}$$

for some constant c.

In the above we used the fact that,

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

and 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 grows linearly with T.

2.7 Naïve Bayes

2.7.1 Bayesian Classification: Binary Domain

Consider the following situation: We have two classes $+1$, -1 and each example is described by N attributes. X_n is binary variable with value $\{0, 1\}$. Example dataset:

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

We want to build a hypothesis, h, which is a mapping from $x_1, ..., x_n$ to $\{+1, -1\}$.

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

$Pr(C = +1)$ is easy to estimate from the data (if it's not too large).

How do we estimate $Pr(x_1, \ldots, x_n|\ C = +1)$?
Naive Bayes is based on the *independence assumption*:

$$Pr(x_1, \ldots, x_n|\ C) = \prod_i Pr(x_i|\ C)$$

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:

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

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

How do we estimate $\theta_{i|+1}$ or $\theta_{i|-1}$? We use again Simple Binomial estimation. 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.)

2.7.2 Interpretation of Naïve Bayes

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

$$Pr(+1|\ x_1, \ldots, x_n) \ \ and \ \ Pr(-1|\ x_1, \ldots, x_n)$$

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

$$\begin{aligned}
\log \frac{Pr(+1|\ x_1, \ldots, x_n)}{Pr(-1|\ x_1, \ldots, x_n)} &= \log \frac{Pr(x_1, \ldots, x_n|\ +1)Pr(+1)}{Pr(x_1, \ldots, x_n|\ -1)Pr(-1)} \\
&= \log \frac{Pr(+1)}{Pr(-1)} + \log \prod_i \frac{Pr(x_i|C = +1)}{Pr(x_i|C = -1)} \\
&= \log \frac{Pr(+1)}{Pr(-1)} + \sum_i \log \frac{Pr(x_i|C = +1)}{Pr(x_i|C = -1)},
\end{aligned}$$

Thus, we conclude that

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

Each x_i "votes" about the prediction

- If $Pr(x_i|\ C = -1) = Pr(x_i|\ C = +1)$ then x_i has no say in classification
- If $Pr(x_i|\ C = -1) = 0$ then x_i overrides all other votes ("veto")

Let us denote:

$$\begin{aligned} w_i &= \log \frac{Pr(x_i = 1|C = +1)}{Pr(x_i = 1|C = -1)} - \log \frac{Pr(x_i = 0|C = +1)}{Pr(x_i = 0|C = -1)} \\ b &= \log \frac{Pr(+1)}{Pr(-1)} + \sum_i \log \frac{Pr(x_i = 0|C = +1)}{Pr(x_i = 0|C = -1)} \end{aligned}$$

The classification rule becomes a separating hyperplane:

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

2.7.3 Practical considerations

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

2.8 Normal Distribution

Usually we also say Gaussian distribution.

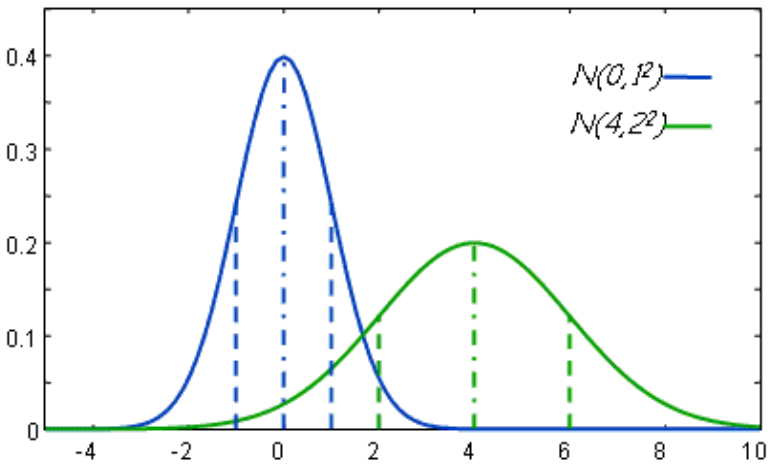
2.8.1 Short reminder

$$X \sim N(\mu, \sigma^2) \text{ if } p(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$Pr[a \leq X \leq b] = \int_a^b p(x)dx$$

$$E[x] = \mu$$

$$Var[x] = E(x - E[x])^2 = E[x^2] - E^2[x] = \sigma^2$$



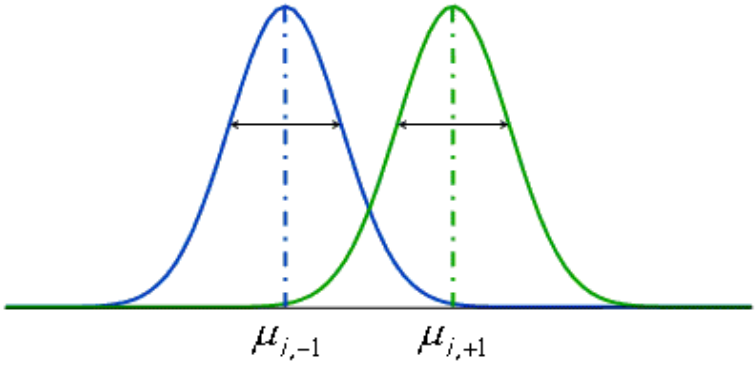
2.8.2 Naïve Bayes with Gaussian Distributions

We recall the *independence assumption*:

$$Pr(x_1, \ldots, x_n | \ C) = \prod_i Pr(x_i | \ C)$$

In addition, we make the following assumptions:

- $Pr(x_i | \ C) \sim N(\mu_{i,c}, \sigma_i^2)$
- 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.



As before,

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

Using the gaussian parameters:

$$\begin{aligned}
\log \frac{Pr(x_i|C=+1)}{Pr(x_i|C=-1)} &= \log \frac{\frac{1}{\sqrt{2\pi}\sigma_i} \cdot e^{-\frac{1}{2}\left(\frac{x_i-\mu_{i,+1}}{\sigma_i}\right)^2}}{\frac{1}{\sqrt{2\pi}\sigma_i} \cdot e^{-\frac{1}{2}\left(\frac{x_i-\mu_{i,-1}}{\sigma_i}\right)^2}} \\
&= -\frac{1}{2}\left(\frac{x_i-\mu_{i,+1}}{\sigma_i}\right)^2 + \frac{1}{2}\left(\frac{x_i-\mu_{i,-1}}{\sigma_i}\right)^2 \\
&= -\underbrace{\frac{\mu_{i,+1}-\mu_{i,-1}}{\sigma_i}}_{\substack{\text{Distance} \\ \text{between} \\ \text{means}}} \underbrace{\frac{1}{\sigma_i}\left(\frac{\mu_{i,+1}+\mu_{i,-1}}{2} - x_i\right)}_{\substack{\text{Distance of} \\ x_i \text{ to midway} \\ \text{point}}}
\end{aligned}$$

If we allow different variances, the classification rule is more complex. The term $\log \frac{Pr(x_i|\ +1)}{Pr(x_i|\ -1)}$ is quadratic in x_i.