Lecture 2: March 21

Lecturer: Yishay Mansour *Scribe: Eli Zeitlin, Anat Rapoport*

2.1 Example - Classifying Oranges

Suppose we are required to build a controller that removes **bad oranges** from a packaging line. Decisions are made based on a sensor that reports the overall color of the orange.

2.1.1 Prior Probabilities:

Lets assume we know all the aspects of the problem:

Probability of good (+1) and bad (-1) oranges

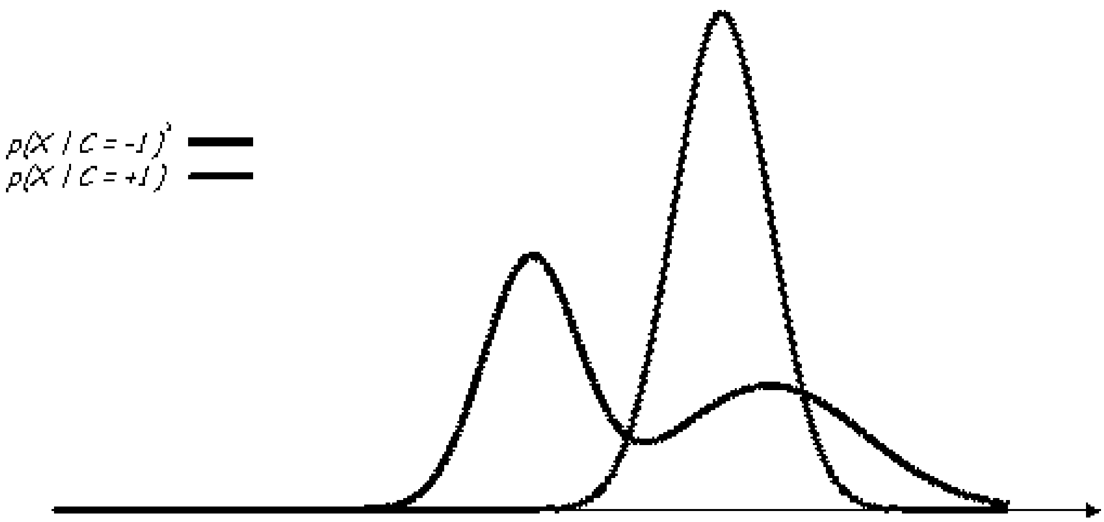
- **P(C=+1)** = Probability of a good orange
- **P(C= -1)**= Probability of a bad orange
  - note: $P(C = +1) + P(C = -1) = 1$

Assumption: The quality of the oranges on the packaging line is independent.
$\Longrightarrow$ the occurrence of a bad orange does not depend on previous.

Given a sensor's measurement we would like to determine whether the orange is good or bad. The result of the measurement is inconclusive.

Sensor performance:

- Let X denote sensor measurements from each type of oranges



2.1.2 Bayes Rule

Given this knowledge, we can compute the *posterior* probabilities.

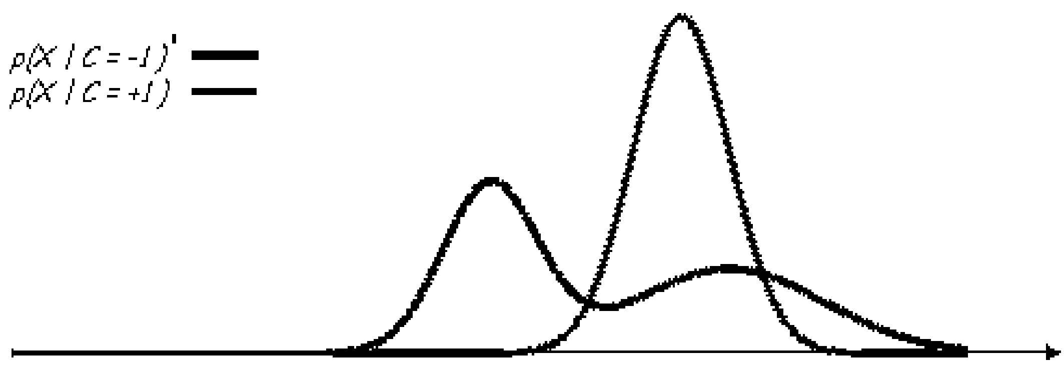
Bayes Rule:

$$P(C|X=x) = \frac{P(C)P(X=x|C)}{P(X=x)}$$

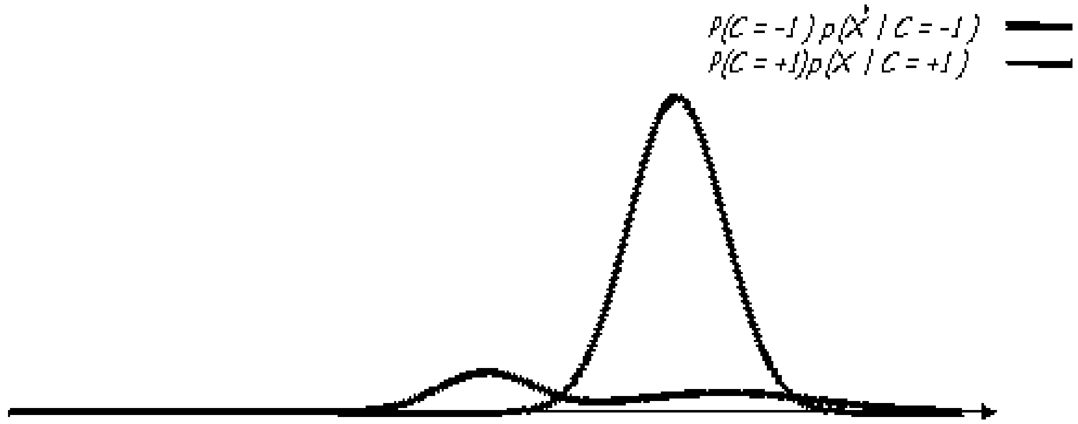
$$P(X=x) = P(C=+1)P(X=x|C=+1) + P(C=-1)P(X=x|C=-1)$$

2.1.3 Posterior of Oranges:

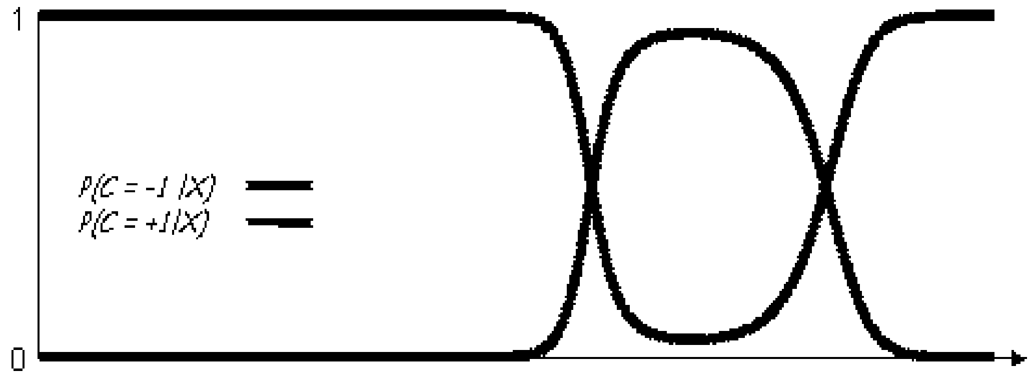
Data likelihood:



Combined with prior:



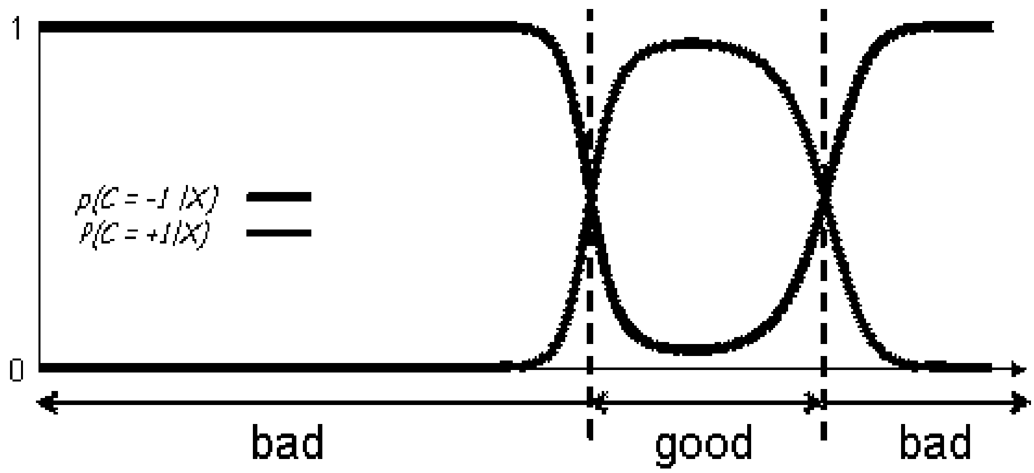
After normalization:



2.1.4 Decision making:

Intuition:

- Predict "Good" if $P(C = +1|X) > P(C = -1|X)$
- Predict "Bad", otherwise



2.1.5 Loss function:

Until now, we didn't take into consideration the price of making a wrong decision. The loss function will tell us what will be the price we'll pay as a function of what we guessed.
Lets assume we have classes of good and bad oranges: +1, -1.
Suppose we can make predictions a_1,..., a_k.
A **loss function** $L(a_j, c_j)$ describes the loss associated with making prediction a_j when the class real is c_j.
Example:

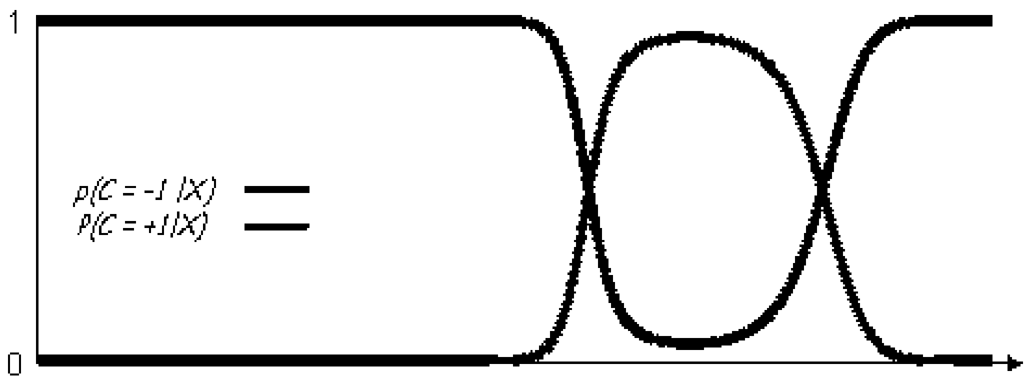
		Real Label	
		-1	+1
Prediction	Bad	1	5
	Good	10	0

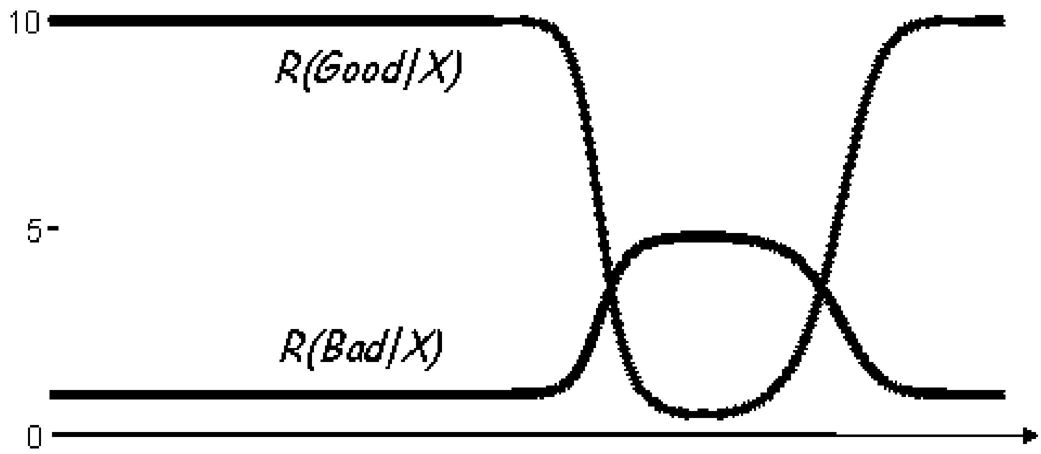
2.1.6 Expected Risk:

Given the estimates of $P(C|X)$ we can compute the **expected conditional risk** of each decision:

$$R(a|X) = \sum_c L(a, c)P(C = c|X)$$

2.1.7 The Risk in Oranges:





Note: If we say that the orange is good when actually it's not, we pay a very high price. Therefor the part of the graph which before was marked as "good", became smaller.

2.2 Optimal Decisions:

Our goal is to minimize the risk. Optimal decision rule: "Given $X = x$, predict a_j if $R((a_j|X = x) = \min_a(R(a|X = x)$". Meaning - choose the prediction which minimizes the risk.
Note: randomized decisions do not help.

- (break ties arbitrarily)

2.2.1 0-1 Loss:

If we don't have prior knowledge, it is common to use the **0-1 loss.**

- $L(a, c) = 0$ if $a = c$
  If your prediction equals the real label, the cost is 0.
- $L(a, c) = 1$ otherwise.
  Your prediction differs from the real label thus there is a cost.

Consequence: $R(a|X) = P(a \neq c|X)$
The risk of your prediction equals the probability that the real label differs from your prediction.

Decision Rule: "choose a_j if $P(C = a_j|X) = \max_a P(C = a|X)$". Meaning: Given a sensor's measurement choose the prediction with the highest probability.

2.2.2 Bayesian Decision: Summary

Decisions are based on two components:

- Conditional distribution $P(C|X)$ - the probability of each decision given a sensor's measurement.
- Loss function $L(A, C)$ - the price we pay for our prediction.

Pros:

- Specifies optimal actions in presence of noisy signals
- Can deal with skewed loss functions

Cons:

- Requires $P(C|X)$

2.3 Simple Statistics: Binomial Experiments

Suppose we toss thumbtack. When tossed, it can land in one of two positions: Head or Tail. We denote by θ the (unknown) probability $P(H)$.

Estimation task:
Given a sequence of toss samples $x[1], x[2], ..., x[M]$ we want to estimate the probabilities $P(H) = \theta$, and $P(T) = 1 - \theta$

2.3.1 Why Learning is Possible?

Suppose we perform M independent flips of the thumbtack.
The number of head we see is binomial distribution

$$P(\#Heads = k) = \left(\begin{array}{c} M \\ k \end{array} \right) \theta^k (1 - \theta)^{M-k}$$

and thus $E[\#Heads] = M\theta$

This suggests, that we can estimate θ by $\frac{\#Heads}{M}$

2.3.2 Maximum Likelihood Estimation

MLE Principle:

Learn parameters that maximize the likelihood function

MLE is one of the most commonly used estimators in statistics. The reason for this is that it's intuitively appealing, and has well studied properties.

2.3.3 Computing the Likelihood Functions

Given a series of tosses D: To compute the likelihood in the thumbtack example we only require N_H and N_T (the number of heads and the number of tails)

$$L(\theta : D) = \theta^{N_H}(1-\theta)^{N_T}$$

Applying the MLE principle we get

$$\widehat{\theta} = \frac{N_H}{N_H + N_T}$$

N_H and N_T are **sufficient statistics** for the binomial distribution.

2.4 Sufficient Statistics

A *sufficient statistic* is a real function of the data that summarizes the relevant information for the likelihood.

Definition Function $T : R^n \to R$ is a *sufficient statistic* for parameter θ if:

$$P(\theta \mid T(X^n), X^n) = P(\theta \mid T(X^n))$$

If we have two data sets X^n, Y^n this definition equivalent to:

$$T(X^n) = T(Y^n) \Rightarrow L(\theta \mid X^n) = L(\theta \mid Y^n)$$

For example: We are not interesting in all the sequence $x_1, \ldots, x_n$ in our binomial experiment, but just the number of instances with H and the number of instances with T.

$$\begin{array}{ll} X^{(4)} = H, H, T, T & \#H = \#T = 2 \\ \hat{X}^{(4)} = H, T, H, T & \#H = \#T = 2 \end{array}$$

2.5 Maximum A Posterior (MAP)

Suppose we flip a coin n times and end up with k heads and $n-k$ tails. We would like to estimate the probability that the next flip will come up head. The probability that out of n flips of coin there are exactly k heads is

$$P(k|\ \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

With the Maximum Likelihood approach, one would choose θ as to maximize $P(k|\ \theta)$. This yields

$$P(H) = \theta = \frac{k}{n} \qquad P(T) = 1 - \theta = \frac{n-k}{n}$$

Yet this result seems unreasonable when n is small. For example if you flip the coin only twice and observe the sequence:

$$T,\ T$$

MLE estimates the probabilities as $P(H) = 0,\ P(T) = 1$. Should we believe that it is *impossible* to get head as the outcome of the next toss? Such an estimate can have disastrous effect. If we assume that $P(H) = 0$, then we are willing to act as though this outcome is impossible.

2.5.1 Laplace Correction

Instead of Maximum Likelihood estimation probabilities to head and tail:

$$P(H) = \theta = \frac{k}{n} \qquad P(T) = 1 - \theta = \frac{n-k}{n}$$

We can use Laplace correction to estimate the probabilities to head and tail:

$$P(H) = \theta = \frac{k+1}{n+2} \qquad P(T) = 1 - \theta = \frac{n-k+1}{n+2}$$

As though we observed one additional H and one T.
Can we justify this estimate? Yes, let us suppose a *uniform* prior distribution of θ. That is, the prior distribution on all the possible coins is uniform, i.e.

$$P(\theta \leq x) = \int_0^x d\theta = x$$

2.5.2 Bayesian Reasoning

- In Bayesian reasoning we represent our **uncertainty** about the unknown parameter θ by a probability distribution

- This probability distribution can be viewed as **subjective** probability. This is a personal judgment of uncertainty.

2.5.3 Bayesian Interface

We start with:

- $P(\theta)$ - *prior distribution* about the values of θ

- $P(x_1, \dots, x_n \mid \theta)$ - we can compute *posterior distribution* on θ

$$P(\theta \mid x_1, \dots, x_n) = \frac{P(x_1, \dots, x_n \mid \theta)P(\theta)}{P(x_1,\ \dots,\ x_n)}$$

Where the *marginal likelihood* is

$$P(x_1, \dots, x_n) = \int_\Theta P(x_1, \dots, x_n \mid \theta)P(\theta)d\theta$$

2.5.4 Binomial Distribution: Laplace Estimation

In this case:

- The unknown parameter is $\theta = P(H)$

- Simplest prior $P(\theta) = 1$ for $0 < \theta < 1$

- Likelihood (k - number of heads in the sequence)

$$P(x_1, \dots, x_n \mid \theta) = \theta^k(1-\theta)^{n-k}$$

- Marginal likelihood

$$P(x_1, \dots, x_n) = \int_0^1 \theta^k(1-\theta)^{n-k}d\theta$$

Using integration by parts we have:

$$\begin{aligned} P(x_1,\ldots,x_n) &= \int_0^1 \theta^k(1-\theta)^{n-k}d\theta \\ &= \frac{1}{k+1}\theta^{k+1}(1-\theta)^{n-k}\Big|_0^1 + \frac{n-k}{k+1}\int_0^1 \theta^{k+1}(1-\theta)^{n-k-1}d\theta \\ &= \frac{n-k}{k+1}\int_0^1 \theta^{k+1}(1-\theta)^{n-k-1}d\theta \end{aligned}$$

Multiply both side by n choose k, we have

$$\binom{n}{k}\int_0^1 \theta^k(1-\theta)^{n-k}d\theta = \binom{n}{k+1}\int_0^1 \theta^{k+1}(1-\theta)^{n-k-1}d\theta$$

The recursion terminates when $k = n$

$$\binom{n}{n}\int_0^1 \theta^n(1-\theta)^{n-n}d\theta = \int_0^1 \theta^n d\theta = \frac{1}{n+1}$$

Thus

$$P(x_1,\ldots,x_n) = \int_0^1 \theta^k(1-\theta)^{n-k}d\theta = \frac{1}{n+1}\binom{n}{k}^{-1}$$

We conclude that the posterior is

$$P(\theta \mid x_1,\ldots,x_n) = (n+1)\binom{n}{k}\theta^k(1-\theta)^{n-k}$$

2.5.5 Bayesian Prediction

How do we predict using the posterior? We can think of this as computing the probability of the next element in the sequence

$$\begin{aligned} P(x_{n+1}\mid x_1,\ldots,x_n) &= \int_\Theta P(x_{n+1},\theta\mid x_1,\ldots,x_n)d\theta \\ &= \int_\Theta P(x_{n+1}\mid \theta,x_1,\ldots,x_n)P(\theta\mid x_1,\ldots,x_n)d\theta \\ &= \int_\Theta P(x_{n+1}\mid \theta)P(\theta\mid x_1,\ldots,x_n)d\theta \end{aligned}$$

Assumption: If we know θ, the probability of x_{n+1} is independent of $x_1,\ldots,x_n$

$$P(x_{n+1} \mid \theta, x_1,\ldots,x_n) = P(x_{n+1}\mid \theta)$$

Thus, we conclude that

$$\begin{aligned} P(x_{n+1} = H \mid x_1, \ldots, x_n) &= \int_\Theta P(x_{n+1}|\ \theta) P(\theta|\ x_1, \ldots, x_n) d\theta \\ &= \int_\Theta \theta P(\theta|\ x_1, \ldots, x_n) d\theta \\ &= (n+1) \binom{n}{k} \int_\Theta \theta^{k+1} (1-\theta)^{n-k} d\theta \\ &= (n+1) \binom{n}{k} \frac{1}{n+2} \binom{n+1}{k+1}^{-1} \\ &= \frac{k+1}{n+2} \end{aligned}$$

2.6 Naïve Bayes

2.6.1 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	$+1$

How do we estimate $P(C)$? We can use Simple Binomial estimation. Count number of instances with $C = -1$ and with $C = +1$

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	$+1$

How do we estimate $P(x_1, \ldots, x_n|\ C)$?
We can think on two sub-problems:

- Training set for $P(x_1, \ldots, x_n|\ C = +1)$

- Training set for $P(x_1, \ldots, x_n|\ C = -1)$

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	$+1$

In Naive Bayes we based on the *independence assumption*:

$$P(x_1, \ldots, x_n|\ C) = P(x_1|\ C) \cdot P(x_2|\ C) \cdots P(x_n|\ C)$$

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:

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

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

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.

2.6.2 Interpretation of Naïve Bayes

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

$$P(+1|\ x_1, \ldots, x_n) \ \ \textit{and} \ \ P(-1|\ x_1, \ldots, x_n)$$

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

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

Thus, we conclude that

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

Each x_i "votes" about the prediction

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

Let us denote:

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

The classification rule becomes:

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

2.7 Normal Distribution

Usually we also say Gaussian distribution.

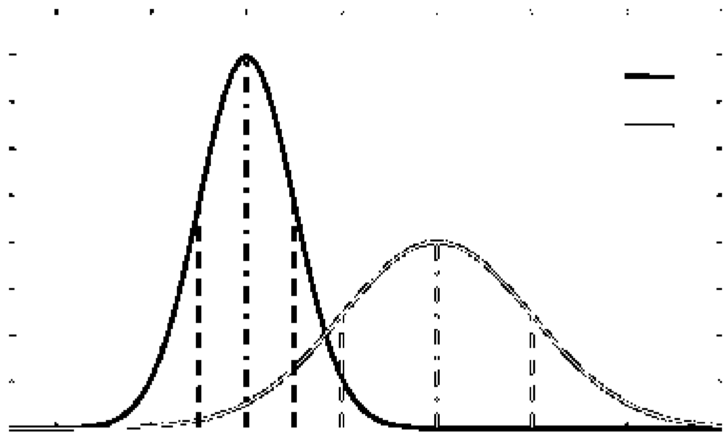
2.7.1 Short reminder

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

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

$$E[x] = \mu$$

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



2.7.2 Maximum Likelihood Estimate

Suppose we observe the sequence $x_1, \ldots, x_m$. Simple calculations show that the MLE is:

$$\begin{aligned} \mu &= \frac{1}{m}\sum_i x_i \\ \sigma^2 &= \frac{1}{m}\sum_i (x_i - \mu)^2 = \frac{1}{m}\sum_i x_i^2 - 2\frac{\mu}{m}\sum_i x_i + \mu^2 \end{aligned}$$

Sufficient statistics are:

$$\frac{1}{m}\sum_i x_i \qquad \frac{1}{m}\sum_i x_i^2$$

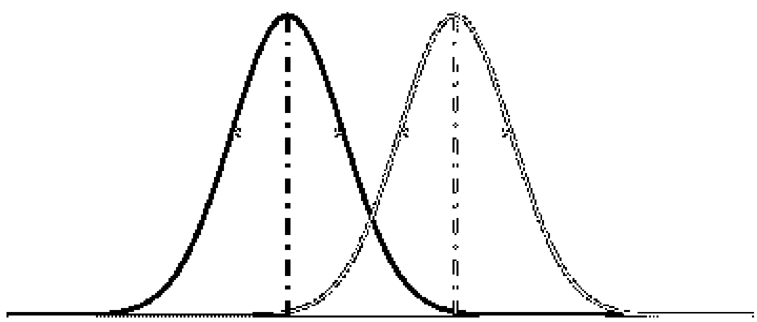
2.7.3 Naïve Bayes with Gaussian Distributions

We recall the *independence assumption*:

$$P(x_1, \ldots, x_n | \, C) = P(x_1 | \, C) \cdot P(x_2 | \, C) \cdots P(x_n | \, C)$$

In addition, we make the following assumptions:

- $P(x_i | \, C) \sim N(\mu, \sigma^2)$
- Mean of x_i depends on class
- Variance of x_i does not depend on class



Recall:

$$\begin{aligned}
\log \frac{P(+1 | \, x_1, \ldots, x_n)}{P(-1 | \, x_1, \ldots, x_n)} &= \log \frac{P(+1)}{P(-1)} + \sum_i \log \frac{P(x_i | \, +1)}{P(x_i | \, -1)} \\
\log \frac{P(x_i | \, +1)}{P(x_i | \, -1)} &= \underbrace{\frac{\mu_{i,+1} - \mu_{i,-1}}{\sigma_i}}_{\substack{\text{Distance}\\ \text{between}\\ \text{means}}} \underbrace{\frac{1}{\sigma_i}\left(\frac{\mu_{i,+1} + \mu_{i,-1}}{2} - x_i\right)}_{\substack{\text{Distance of}\\ x_i \text{ to midway}\\ \text{point}}}
\end{aligned}$$

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