Pearl 110 - Intelligent Interaction

Boolean events
Discriminant functions, classifiers and decision boundaries

Yesterday's lecture

Estimation of probability of events:

- Prior
- Evidence
- Class-conditional
- Posterior

How to use probabilities to make decisions (e.g. *class with highest posterior probability*).

Event

We have a bucket with 10 fishes (4 salmon and 6 sea bass).
We do an experiment, and obtain an outcome.

The **outcome of the experiment is an event,** and can be:

- We pick a salmon from the bucket
- We pick a fish of width 26cm
- We pick a sea bass of 25cm
- ...

To each of these events we can assign a probability

Different types of event

Terminology that we use in probabilistic decision theory.

- **Prior**: probability of an event to happen, before (prior) to measuring any characteristics (features) of the data
- **Evidence:** probability of *measuring a certain value of features* from all the samples
- **Class-conditional:** probability of *measuring a certain feature value* from samples that *belong to a certain class* (conditioned on the class)
- **Posterior:** probability of a sample to *belong to a certain class*, *given that* we have measured *a certain feature value*

Posterior probability with the ***Bayes formula***

$$\underbrace{\boxed{P(C|x)}}_{\text{posterior}} = \frac{\overbrace{\boxed{P(x|C)}}^{\text{class-conditional}}\ \overbrace{\boxed{P(C)}}^{\text{prior}}}{\underbrace{\boxed{P(x)}}_{\text{evidence}}}$$

The evidence is computed as: $P(X) = P(X|C_1)P(C_1) + P(X|C_2)P(C_2) + \ldots$

Naive Bayes

Independent Events

$$P(A, B) \longrightarrow P(A)P(B)$$

Naïve Bayes rule:

$$P(C|x_1, x_2, \ldots, x_n) = \frac{P(x_1|C)P(x_2|C)\ldots P(x_n|C)P(C)}{P(x_1, x_2, \ldots, x_n)}$$

What's for today?

Today's lecture

1. **Estimating the probability of repeated binary events**

   Bernoulli trial experiment

2. **How to build a classifier analytically?**

   Compute a decision boundary

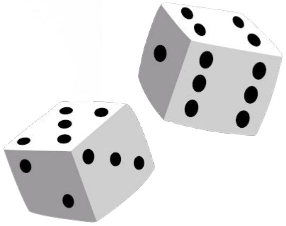
   Do it with Gaussian functions (the case of the Normal distribution)

A binary event

We make an experiment -> we have a certain outcome OR not (opposite outcome)

We toss a coin -> we get a tail or not a tail

We roll a dice -> we get a 5 or not a 5



Dice rolling/Coin tossing: Bernoulli Trial

If I toss a coin 5 times, what is the probability of getting 4 tails?

If I roll a dice 5 times, what is the probability of getting 3 four?

...or 3 four and 2 two?

Assumptions

The events are independent: one coin toss does not influence the next.

If I roll a dice and get a 4, the next dice roll does not depend on it.

The events are binary: If I toss a coin, I get a tail or I don't get a tail.

If I roll a dice, I can get a 5 or not.

Bernoulli Trial: probability estimation

If I toss a coin 5 times, what is the probability of getting 4 tails?

The probability of tails is equal to the probability of heads (balanced coin)

P(T) = P(H) = ½

TTTTH	½ * ½ *½ * ½ * ½ = $(½)^5$
TTTHT	½ * ½ *½ * ½ * ½ = $(½)^5$
TTHTT	½ * ½ *½ * ½ * ½ = $(½)^5$
THTTT	½ * ½ *½ * ½ * ½ = $(½)^5$
HTTTT	½ * ½ *½ * ½ * ½ = $(½)^5$

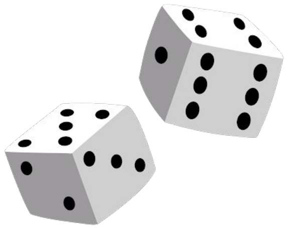
All possible combinations of the events (5)

Probability for each combination

Observation: the order of the events is not important for the experiment

$$P(t=4, \neg t=1) = 5 \cdot (\tfrac{1}{2})^5$$

Bernoulli Trial: probability estimation



If I roll a dice 5 times, what is the probability of getting 3 fours?

Event **F** = get a four
Event **N** = not get a four
P(F) = $\frac{1}{6}$ **P(N) = $\frac{5}{6}$**

We want to compute P(F = 3, N=2)
the probability of rolling 3 fours and 2 not four

$$\binom{5}{3} = \frac{5!}{3!(5-3)!} = \frac{5\cdot4\cdot3\cdot2\cdot1}{3\cdot2\cdot1\cdot(2\cdot1)} = 10$$

10 combinations
10 ways to obtain **3 F** out of **5 events**
(having a four in 3 out of 5 dice rolls)

FFFNN	$\frac{1}{6} * \frac{1}{6} * \frac{1}{6} * \frac{5}{6} * \frac{5}{6} = (\frac{1}{6})^3 (\frac{5}{6})^2$
FFNFN	$\frac{1}{6} * \frac{1}{6} * \frac{5}{6} * \frac{1}{6} * \frac{5}{6} = (\frac{1}{6})^3 (\frac{5}{6})^2$
FNFFN	$\frac{1}{6} * \frac{5}{6} * \frac{1}{6} * \frac{1}{6} * \frac{5}{6} = (\frac{1}{6})^3 (\frac{5}{6})^2$
NFFFN	$\frac{5}{6} * \frac{1}{6} * \frac{1}{6} * \frac{1}{6} * \frac{5}{6} = (\frac{1}{6})^3 (\frac{5}{6})^2$
NFFNF	...
NFNFF	
NNFFF	
FNNFF	
FFNNF	
FNFNF	

$$10 * [P(F)]^3 [P(N)]^2$$

Bernoulli Trial: formula

From N choose k

$$\binom{N}{k} = \frac{N!}{k!(N-k)!}$$

Binomial coefficient

Probability of ***x=k*** successes given **N** trials (where ***p*** is the probability of a single success)

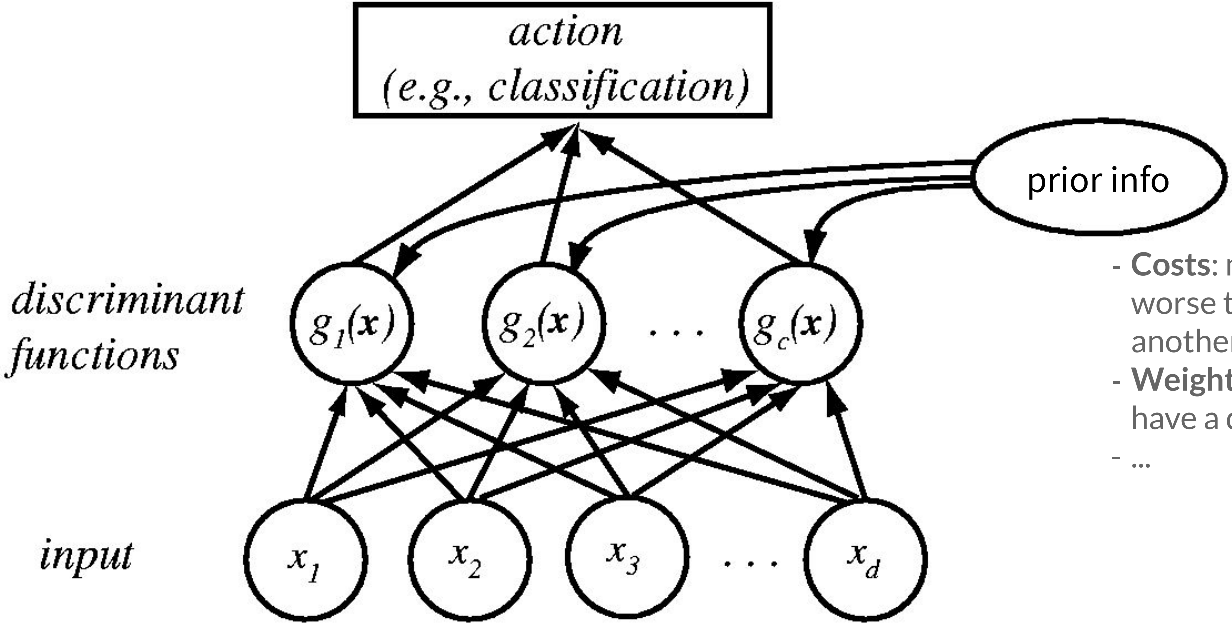
$$P(x=k)|_N = \binom{N}{k} p^k (1-p)^{N-k}$$

Binomial distribution

Determining the decision boundary

How to build classifiers using probability distribution functions

Structure of a classifier



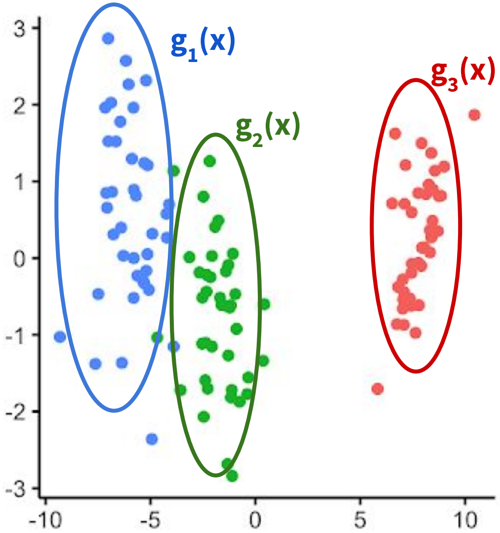
- **Costs**: making a mistake on one class is worse than making a mistake on another class (penalize certain errors)
- **Weighting**: different functions might have a different importance
- ...

(from Duda, Hart, Stork (2001) Pattern classification)

Discriminant function

A function that models a set of data points.

Instead of working with single data points, we summarize/approximate them with a function, that describes their distribution.



x: a vector of measured features

[length, lightness] in the fish example

$g_i(x)$ takes the feature vector as input

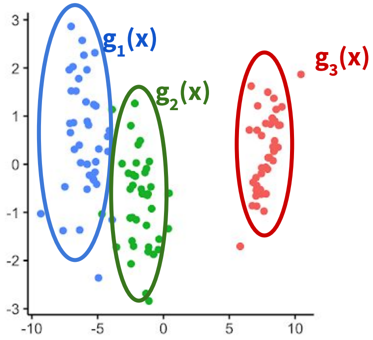
Discriminant functions

e.g. the function $\mathbf{g_1(x)}$ computes a score for the sample **x** for belonging to the class $\mathbf{C_1}$

A classifier computes the value of discriminant functions $\mathbf{g_i(x)}$ for several classes and assign to **x** the class corresponding to the discriminant function with the largest score.

The classifier assigns a feature vector **x** to class **i** if

$$g_i(\mathrm{x}) > g_j(\mathrm{x}), \quad \text{for all } j \neq i$$



$$g_i(x) = P(C_i|x)$$

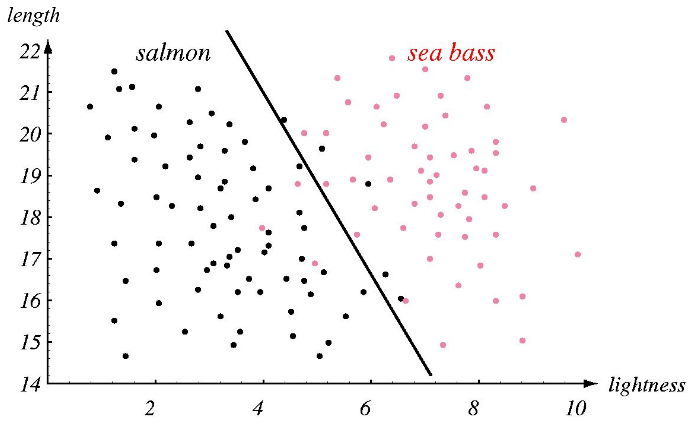
Posterior probability can be a discriminant function.

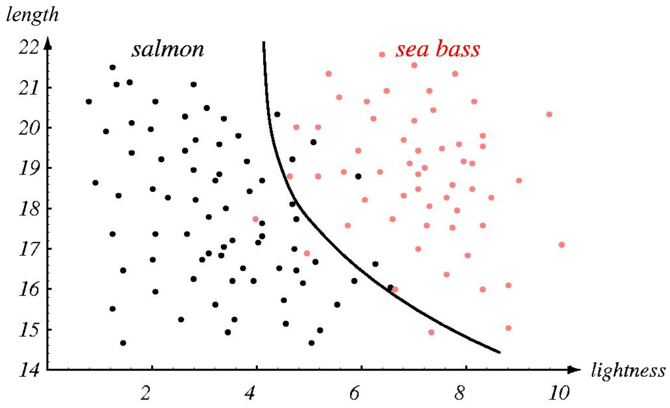
C_i is the Class i (i = 1, 2, …)

How to use discriminant functions?

Create a model of the samples in one class with a discriminant function *(the distribution of these samples is described by the discriminant function)*

Combine the discriminant functions of different classes to find the decision boundary **(HOW???)**





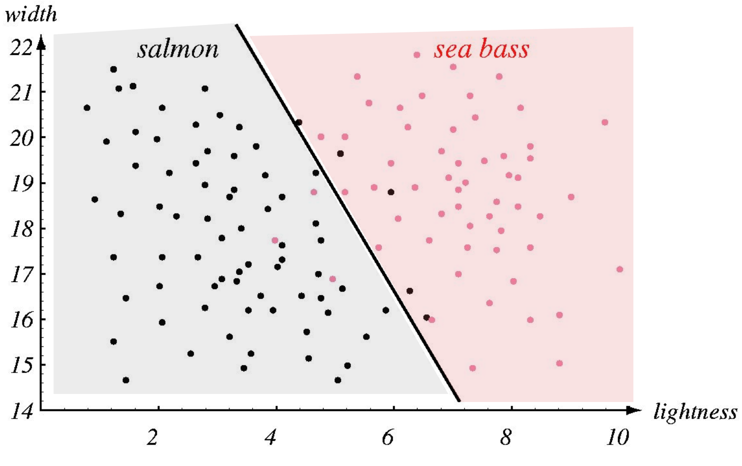
Dichotomizer: a two-class classifier

Classifies samples in two classes.

It is defined as: $g(\mathrm{x}) = g_1(\mathrm{x}) - g_2(\mathrm{x})$

Discriminant function for class 1 (salmon)

Discriminant function for class 2 (seabass)



(from Duda, Hart, Stork (2001) Pattern classification)

Classify **x** into class C_1 if **g(x) > 0**

We can use the posterior probability

$$g(x) = P(C_1|x) - P(C_2|x)$$

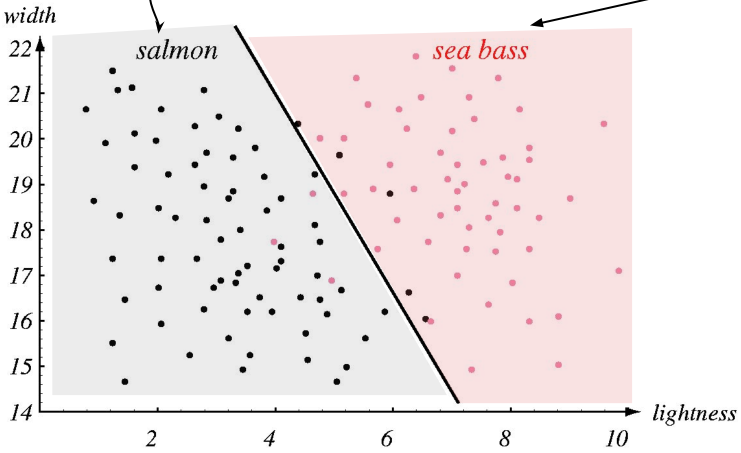
How to compute the decision boundary of a dichotomizer?

Decision regions

$$g(\mathrm{x}) = g_1(\mathrm{x}) - g_2(\mathrm{x})$$

$$g(x) > 0 \iff g_1(x) > g_2(x)$$

$$g(x) < 0 \iff g_1(x) < g_2(x)$$



The decision boundary is where:

$$g(x) = g_1(x) - g_2(x) = 0$$

Probabilistic dichotomizer

We use the Posterior probability as discriminant function.

We define the classifier as:

$$g(x) = g_1(x) - g_2(x) = P(C_1|x) - P(C_2|x)$$

And the decision regions become:

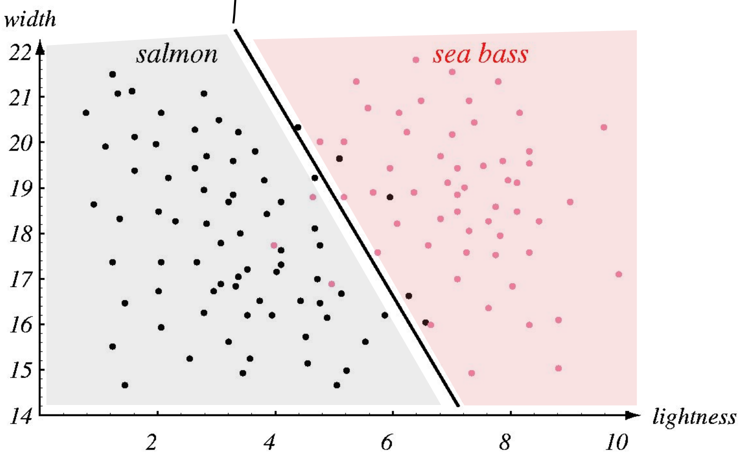
$$g(x) > 0 \iff P(C_1|x) > P(C_2|x)$$

We decide for class 1

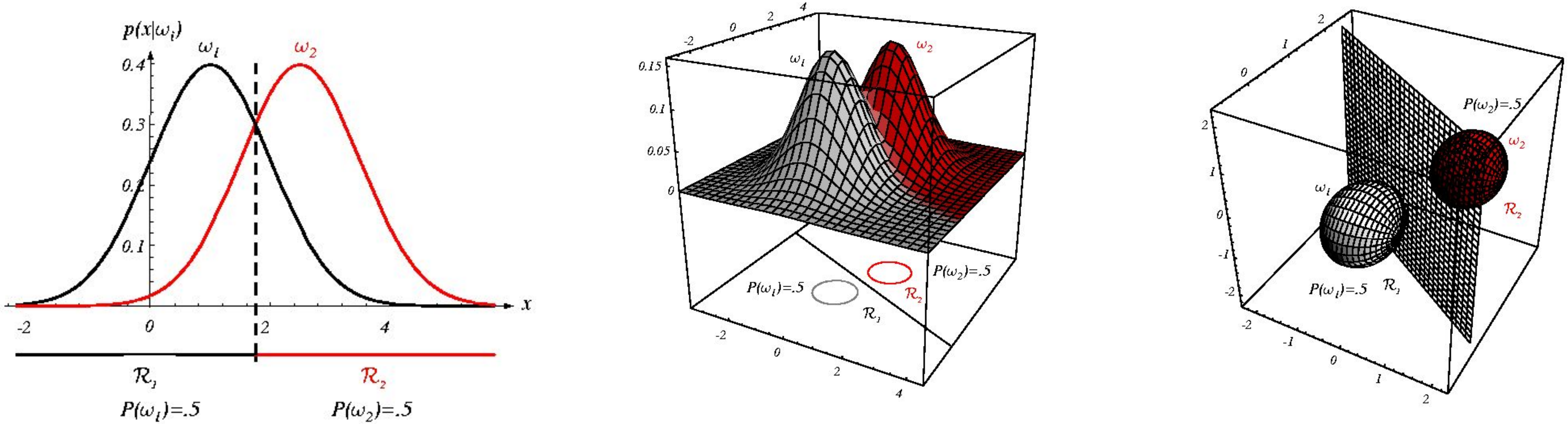
$$g(x) < 0 \iff P(C_1|x) < P(C_2|x)$$

We decide for class 2

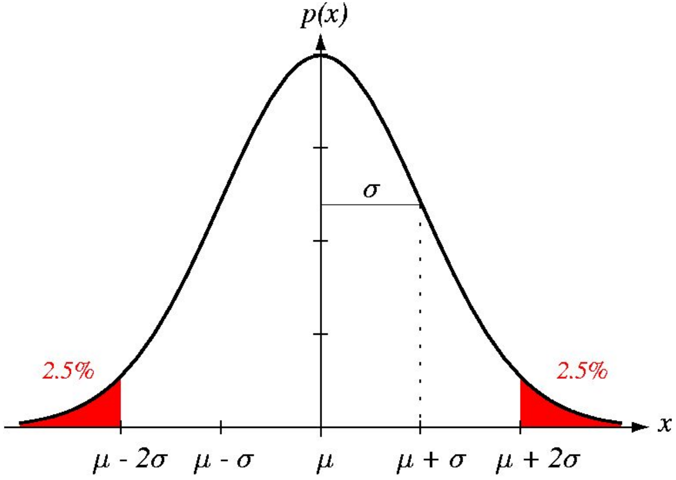
$$P(C_1|x) = P(C_2|x)$$



Dichotomizer: examples in higher dimensions



(from Duda, Hart, Stork (2001) Pattern classification)



Modeling using the normal distribution

An example

Given the sets (training data) S1 and S2

S_1 = [8.70, 3.31, -13.48, 15.48, -6.17, -6.99, -14.24, -1.10, -1.03, -3.23]
S_2 = [18.21, 1.79, 95.25, 65.02, 27.82, 32.70, 42.18, 34.76, 23.59, 53.68, 12.23, 74.15, -0.26, 28.53, 52.45]

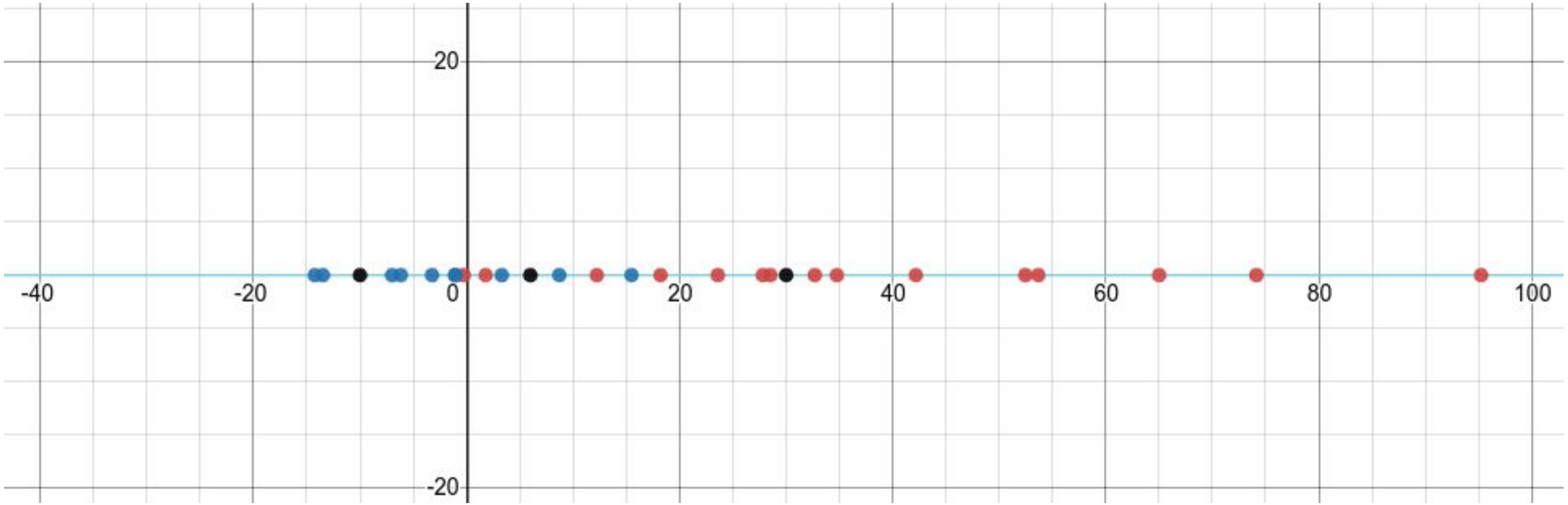
that come from two different classes.

Problem: we want to classify the following test points:

-10, 6, 30

…example…

Given the points in **Class 1** and **Class 2**, in which class should we classify the **black points**?



We look for the closest points

We count the **k=5 Nearest Neighbours.**

Test point **x = -10**.
The 5 nearest neighbors of this point are:

S_1 = [8.70, 3.31, **-13.48**, 15.48, **-6.17**, **-6.99**, **-14.24**, -1.10, -1.03, **-3.23**]
S_2 = [18.21, 1.79, 95.25, 65.02, 27.82, 32.70, 42.18, 34.76, 23.59, 53.68, 12.23, 74.15, -0.26, 28.53, 52.45]

All 5 nearest neighbors come from **S_1**.

Hence, point **-10** is decided to belong to **Class 1**.

...another sample...

Test point **x = 6**.

The 5 nearest neighbors of this point are

S_1 = [**8.70, 3.31**, -13.48, 15.48, -6.17, -6.99, -14.24, -1.10, -1.03, -3.23]
S_2 = [18.21, <u>**1.79**</u>, 95.25, 65.02, 27.82, 32.70, 42.18, 34.76, 23.59, 53.68, <u>**12.23**</u>, 74.15, <u>**-0.26**</u>, 28.53, 52.45]

Most (3) of the 5 nearest neighbors come from **S_2** (2 from **S_1**).

Hence, point **6** is decided to belong to **class 2**.

...and another...

Test point **x = 30**.
The 5 nearest neighbors of this point are:

S_1 = [8.70, 3.31, -13.48, 15.48, -6.17, -6.99, -14.24, -1.10, -1.03, -3.23]
S_2 = [18.21, 1.79, 95.25, 65.02, **27.82**, **32.70**, 42.18, **34.76**, **23.59**, 53.68, 12.23, 74.15, -0.26, **28.53**, 52.45]

All 5 nearest neighbors come from **S_2**.

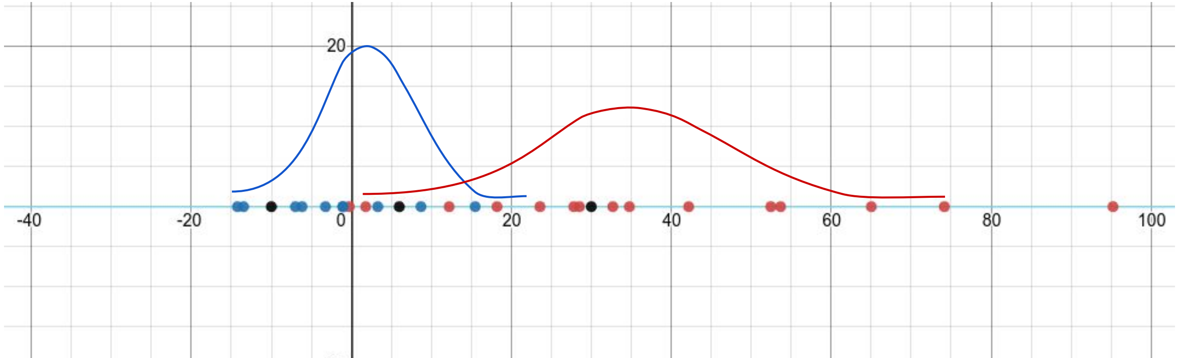
Hence, point **30** is decided to belong to **class 2**.

Observations on k-nearest classification

This counting of nearest neighbors seems a little shaky – if the votes are 3:2 and I move the test point a little bit, they may become 2:3. **How reliable is this?**

We will model the probabilities by some smooth, reliable and predictably changing functions. For instance, a **Gaussian function**!

Why? Because we can use it to compute the probability of measuring a certain value of the features also in domains in which we do not have precise (training) data.



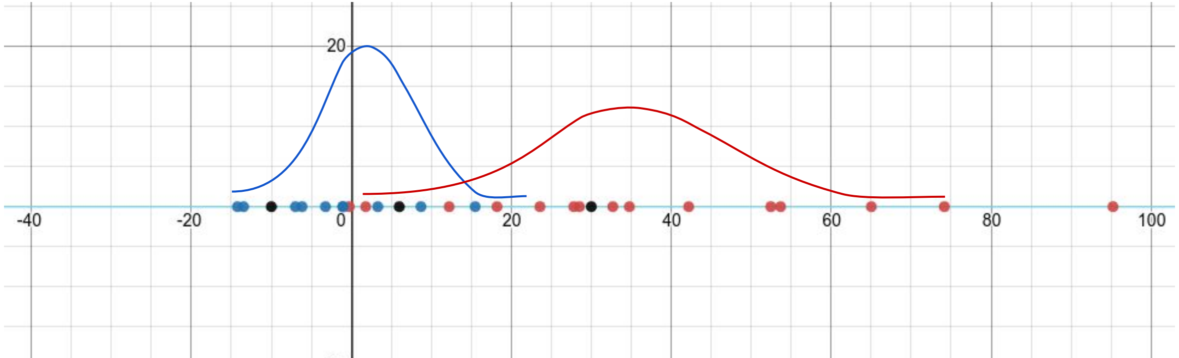
Where do we use the Gaussian function?

In the Bayes formula:

$$P(C_i|x) = \frac{P(x|C_i)P(C_i)}{P(x)}$$

We model the data points (i.e. ***x***) belonging to the class $\mathbf{C_i}$ with a function that describes the distribution of probability of their features.

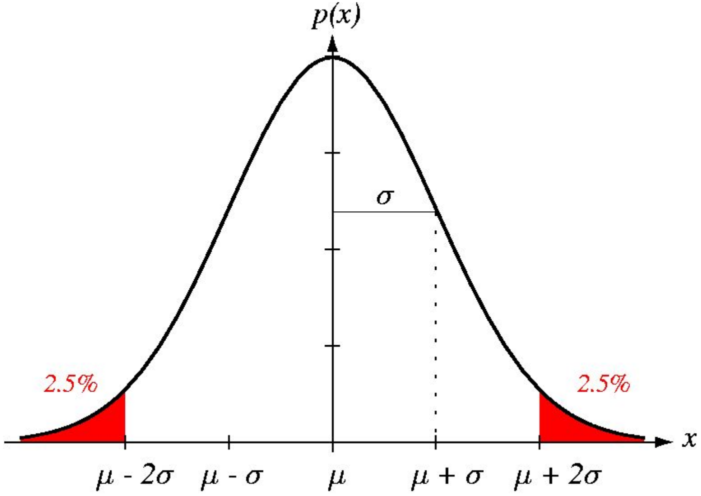
The class-conditional probability is a function (probability density function). Instead of working with the single data points, we work with this function.



Univariate normal density (Gaussian function)

Univariate: we consider only one dimension (on feature)

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$



The Gaussian function has two parameters that can be estimated:

mean: $\mu = \frac{1}{n}\sum_{k=1}^{n} x_k$

variance: $\sigma^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \mu)^2$

Standard deviation: $\sigma = \sqrt{\sigma^2}$

Back to the example

$\mathbf{S_1}$ = [8.70, 3.31, -13.48, 15.48, -6.17, -6.99, -14.24, -1.10, -1.03, -3.23]
$\mathbf{S_2}$ = [18.21, 1.79, 95.25, 65.02, 27.82, 32.70, 42.18, 34.76, 23.59, 53.68, 12.23, 74.15, -0.26, 28.53, 52.45]

We decide to model the two classes that generated this data by two normal distributions. *What are the parameters of these normal distributions?*

Estimation of the mean and variance

$$\mu = \frac{1}{n}\sum_{k=1}^{n} x_k$$

$$\sigma^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \mu)^2$$

We compute: $\mu_1 = 0$, $\sigma_1^{\,2} = 10^2$
$\mu_2 = 35$, $\sigma_2^{\,2} = 20^2$

Example: the class-conditional probability

We use Gaussian functions to model the class-conditional probability.

We have mean and standard deviation of the two classes!!!

$\mu_1 = 0, \quad \sigma_1^2 = 10^2$

$\mu_2 = 35, \sigma_2^2 = 20^2$

$$P(x|C_1) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} = \frac{1}{10\sqrt{2\pi}} e^{-\frac{x^2}{2\cdot 10^2}}$$

$$P(x|C_2) = \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}} = \frac{1}{20\sqrt{2\pi}} e^{-\frac{x^2-35}{2\cdot 20^2}}$$

Example: estimating the prior probabilities

From class 1, we observe 10 values: $|S_1|=10$.
From class 2, we observe 15 values: $|S_2|=15$.

We estimate the prior probabilities of **class 1** and **class 2** by the their frequency of occurrence:

$$P(C_1) = \frac{|S_1|}{|S_1|+|S_2|} = \frac{10}{10+15} = 0.4$$

$$P(C_2) = \frac{|S_2|}{|S_1|+|S_2|} = \frac{15}{10+15} = 0.6$$

Example: the posterior probability

Estimation of the posterior probability using the Bayes rule (and the quantities we have computed earlier):

$$P(C_1|x) = \frac{P(C_1)P(x|C_1)}{P(x)} = \frac{0.4\frac{1}{10\sqrt{2\pi}}e^{-\frac{x^2}{2\cdot 10^2}}}{P(x)}$$

$$P(C_2|x) = \frac{P(C_2)P(x|C_2)}{P(x)} = \frac{0.6\frac{1}{20\sqrt{2\pi}}e^{-\frac{(x-35)^2}{2\cdot 20^2}}}{P(x)}$$

Note: for a new data point **x**’ we can compute the posterior probability of it belonging to class 1 and the posterior probability of belonging to class 2. We select the class with highest posterior probability.

Computing the decision boundary

For the decision criterion (boundary), it holds **$P(C_1|x)$** = **$P(C_2|x)$**
This leads to the following equation:

$$\frac{0.4\frac{1}{10\sqrt{2\pi}}e^{-\frac{x^2}{2\cdot10^2}}}{P(x)}=\frac{0.6\frac{1}{20\sqrt{2\pi}}e^{-\frac{(x-35)^2}{2\cdot20^2}}}{P(x)}$$

After simplifying [P(x) is on both sides, and the factors at the beginning can be manipulated], we have:

$$8e^{-\frac{x^2}{2\cdot10^2}}=6e^{-\frac{(x-35)^2}{2\cdot20^2}}$$

Still somehow complicate, due to the exponential function

…we can make it better to work with!

The choice of discriminant functions is not unique!

...we can make it better to work with!

The choice of discriminant functions is not unique!

In general, if every discriminant function is replaced by a monotonically increasing function $f(g_i(\mathrm{x}))$ the classification result does not change.

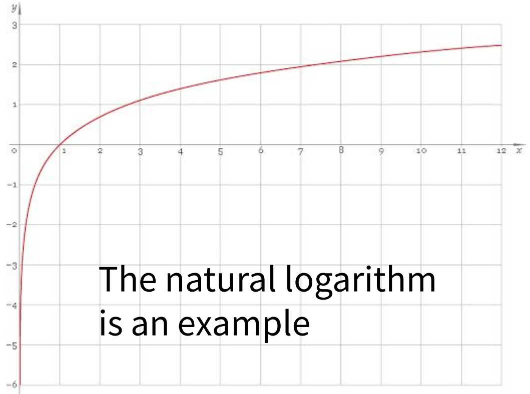
Motivation: other quantities, simpler to understand or to compute, lead to identical classification results.

…we can make it better to work with!

The choice of discriminant functions is not unique!

In general, if every discriminant function is replaced by a monotonically increasing function $f(g_i(\mathrm{x}))$ the classification result does not change.

Motivation: other quantities, simpler to understand or to compute, lead to identical classification results.



…we can make it better to work with!

The choice of discriminant functions is not unique!

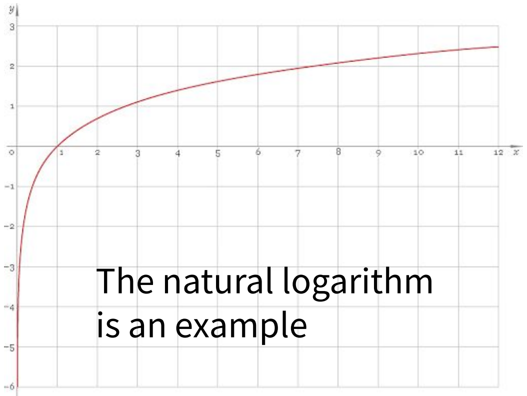
In general, if every discriminant function is replaced by a monotonically increasing function $f(g_i(\mathrm{x}))$ the classification result does not change.

Motivation: other quantities, simpler to understand or to compute, lead to identical classification results.

$$g_i(x) = P(x|C_i)P(C_i)$$

$$f(g_i(x)) = ln\left[P(x|C_i)P(C_i)\right]$$

$$f(g_i(x)) = ln\left[P(x|C_i)\right] + ln\left[P(C_i)\right]$$



(It is useful in the important case of normal distribution)

One step to simplify the decision boundary

We have stopped at this point, when computing the decision boundary:

$$8e^{-\frac{x^2}{2\cdot 10^2}} = 6e^{-\frac{(x-35)^2}{2\cdot 20^2}}$$

If we take the natural logarithm on both sides, we can have benefits:

$$ln\left[8e^{-\frac{x^2}{2\cdot 10^2}}\right] = ln\left[6e^{-\frac{(x-35)^2}{2\cdot 20^2}}\right]$$

Computing the decision boundary (2)

We can use the *natural logarithm* function ln(x) to further simplify the equation:

$$ln\left[8e^{-\frac{x^2}{2\cdot10^2}}\right] = ln\left[6e^{-\frac{(x-35)^2}{2\cdot20^2}}\right]$$

Computing the decision boundary (2)

We can use the *natural logarithm* function ln(x) to further simplify the equation:

$$ln\left[8e^{-\frac{x^2}{2\cdot 10^2}}\right] = ln\left[6e^{-\frac{(x-35)^2}{2\cdot 20^2}}\right]$$

$$ln\left[8\right] + ln\left[e^{-\frac{x^2}{2\cdot 10^2}}\right] = ln\left[6\right] + ln\left[e^{-\frac{(x-35)^2}{2\cdot 20^2}}\right]$$

Computing the decision boundary (2)

We can use the *natural logarithm* function ln(x) to further simplify the equation:

$$ln\left[8e^{-\frac{x^2}{2\cdot10^2}}\right] = ln\left[6e^{-\frac{(x-35)^2}{2\cdot20^2}}\right]$$

$$ln\,[8] + ln\left[e^{-\frac{x^2}{2\cdot10^2}}\right] = ln\,[6] + ln\left[e^{-\frac{(x-35)^2}{2\cdot20^2}}\right]$$

$$ln8 - \frac{x^2}{2\cdot10^2} = ln6 - \frac{(x-35)^2}{2\cdot20^2}$$

Computing the decision boundary (2)

We can use the *natural logarithm* function ln(x) to further simplify the equation:

$$ln\left[8e^{-\frac{x^2}{2\cdot 10^2}}\right] = ln\left[6e^{-\frac{(x-35)^2}{2\cdot 20^2}}\right]$$

$$ln\,[8] + ln\left[e^{-\frac{x^2}{2\cdot 10^2}}\right] = ln\,[6] + ln\left[e^{-\frac{(x-35)^2}{2\cdot 20^2}}\right]$$

$$ln8 - \frac{x^2}{2\cdot 10^2} = ln6 - \frac{(x-35)^2}{2\cdot 20^2}$$

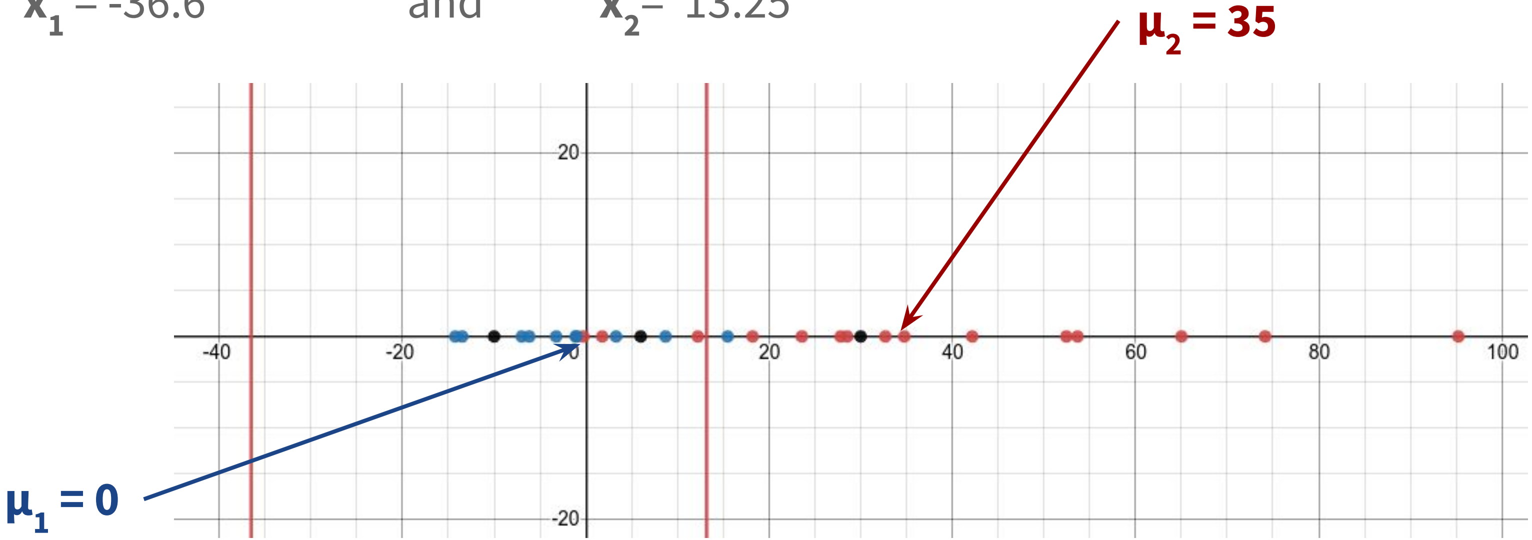
And reduce it to: $3x^2 + 70x - 1445.14 = 0$

Computing the decision boundary (3)

The quadratic equation $$3x^2 + 70x - 1445.14 = 0$$

Has the following solutions:

x_1 = -36.6 and x_2 = 13.25



Computing the decision boundary (3)

The quadratic equation $$3x^2 + 70x - 1445.14 = 0$$

Has the following solutions:

x_1 = -36.6 and **x_2** = 13.25

We use them to compute the decision regions: $$g_1(x) > g_2(x)$$

If **-36.6 < x < 13.25** then **x** belongs to **Class 1**

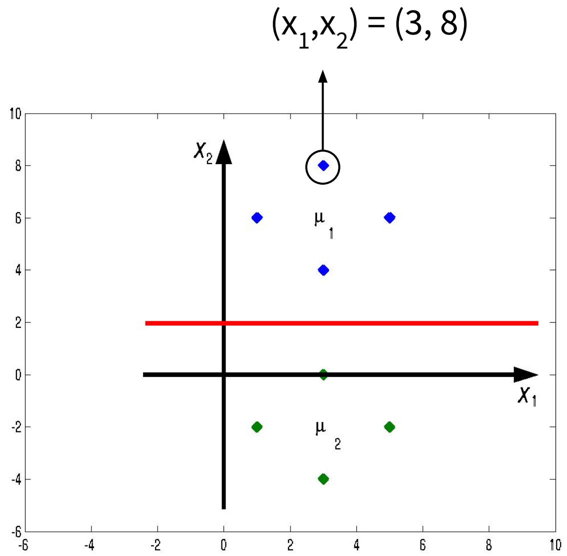
If **x < -36.6 or x > 13.25**, x belongs to **Class 2**

Our problem: **-10** and **6** are **Class 1**, and **30** is **Class 2**

More advanced cases

When we have more dimension...

In the 2D case, a data point (sample) is described 2 features (arranged in a vector).



We need to work with vectors, and know how to multiply two vectors.

$$\begin{pmatrix} a & b \end{pmatrix} \times \begin{pmatrix} e \\ g \end{pmatrix} = a \cdot e + b \cdot g$$

equivalent to: $\begin{pmatrix} a & b \end{pmatrix} \cdot \begin{pmatrix} e & g \end{pmatrix} = a \cdot e + b \cdot g$

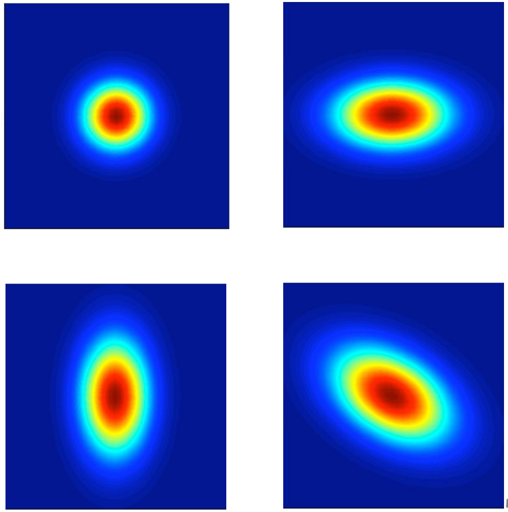
not for the exam

Many features: multivariate Gaussian

When you have ***d*** features, the Gaussian function becomes:

$$p(\mathrm{x}) = \frac{1}{(2\pi)^{d/2}\,|\sum|^{1/2}} \exp\left[-\frac{1}{2}(\mathrm{x}-\mu)^{t}\boxed{\sum}^{-1}(\mathrm{x}-\mu)\right]$$

It is important because it describes the shape of the distribution



$\mathrm{x} \in \mathrm{R}^{d}$ is a *d*-dimensional vector

μ is the mean (a *d*-dimensional vector itself)

Σ is the covariance matrix

($|\Sigma|$ its determinant and Σ^{-1} its inverse)

Decision boundary for a simple 2D case

$$\begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$$

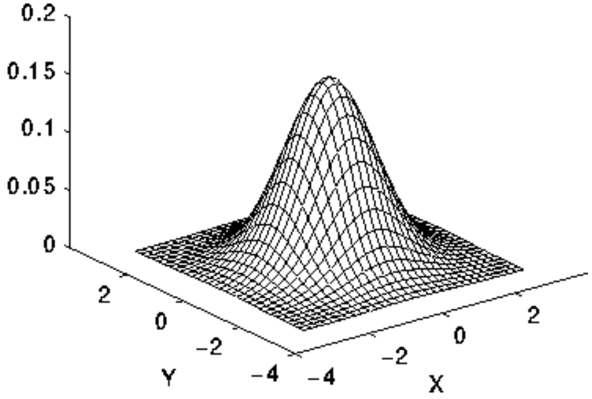
We consider statistical independence between features, i.e. $\Sigma_i = \sigma^2 I$

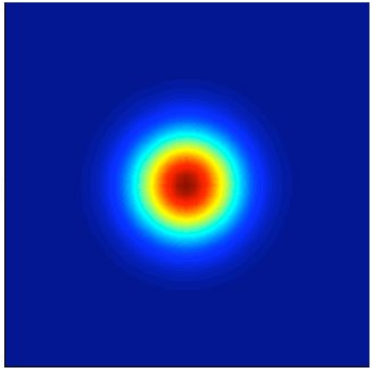
The form of the discriminant functions simplifies to:

$$g_i(x) = -\frac{||x-\mu_i||^2}{2\sigma^2} + lnP(C_i)$$

where $\| \mathrm{x} - \mu_i \|^2 = (\mathrm{x} - \mu_i)^t (\mathrm{x} - \mu_i)$

t means the *transpose* operation

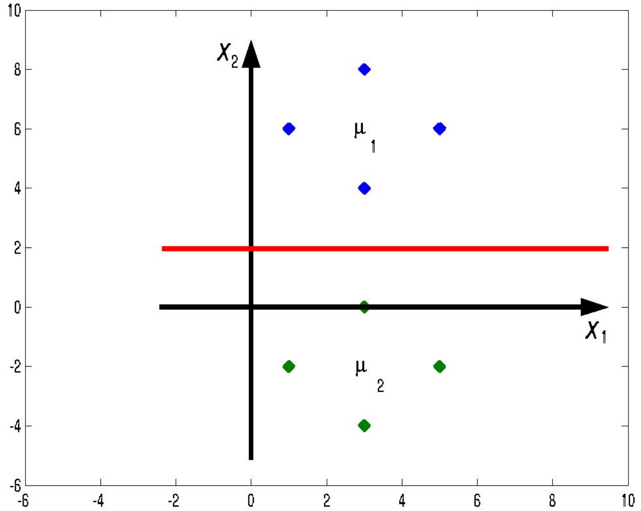




Let us see it with an example

$$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix} \quad \mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix} \quad \sigma_1 = \sigma_2 = \sqrt{2} \quad P(C_1) = P(C_2) = 0.5$$

Class 1: {(3,8), (1,6), (5,6), (3,4)}
Class 2: {(3,0), (1,-2), (3,-4), (5,-2) }



To determine the decision boundary we have to compute:

$$g_1(\mathrm{x}) = g_2(\mathrm{x})$$

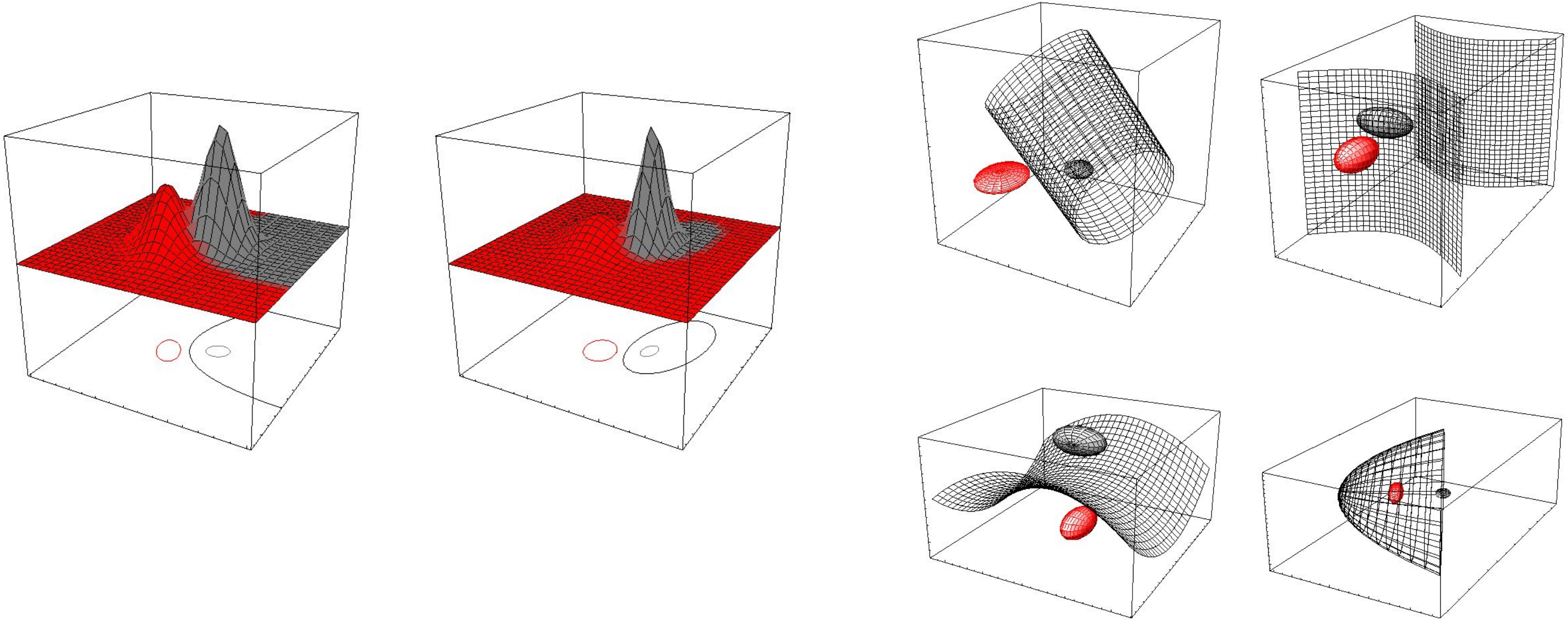
$$\Rightarrow \quad -\frac{\|x-\mu_1\|^2}{2\sigma^2} + lnP(C_1) = -\frac{\|x-\mu_2\|^2}{2\sigma^2} + lnP(C_2)$$

$$\Rightarrow \quad (\mathrm{x} - \mu_1)^t(\mathrm{x} - \mu_1) = (\mathrm{x} - \mu_2)^t(\mathrm{x} - \mu_2)$$

$$\Rightarrow \quad \begin{bmatrix} x_1 - 3 & x_2 - 6 \end{bmatrix} \begin{bmatrix} x_1 - 3 \\ x_2 - 6 \end{bmatrix} = \begin{bmatrix} x_1 - 3 & x_2 + 2 \end{bmatrix} \begin{bmatrix} x_1 - 3 \\ x_2 + 2 \end{bmatrix}$$

$$\Rightarrow \quad (x_1 - 3)^2 + (x_2 - 6)^2 = (x_1 - 3)^2 + (x_2 + 2)^2 \quad \Rightarrow \quad \boxed{x_2 = 2}$$

Possible boundary determined from data



(from Duda, Hart, Stork (2001) Pattern classification)

Summary

How to estimate the probability of repeated boolean events (Bernoulli trials)

Discriminant functions, and how to use them to build a classifier

> *With accurate modeling of the class samples distributions you can define the analytical form of a classifier*

How to construct a classifier (determining the decision boundary of a two-class problem) using the Gaussian distribution

Assignments

This afternoon, focus on the **preparative assignments 7.2 and 7.3**

You should also start working on the **pearl assignment 7.6**

Look also in the extra preparatory exercises and exam questions