Machine Learning and apps in AI: practice session 00

Martin Palazzo

Universite de Technologie de Troyes
Universidad Tecnologica Nacional Buenos Aires

martin.palazzo@frba.utn.edu.ar

November 28, 2022

Overview

1. Supervised Learning
   - Data: a random variable
   - Labeled data
   - Empirical Risk Minimization
2. Bayes Decision Theory
   - Univariate cases
   - Multi-variate cases
3. Maximum Likelihood parameter estimation

Disclaimer

The following document has been created as supporting and guiding material during the practical lessons during the Pattern Recognition course of the Master OSS. The official bibliography and theory materials have been distributed previously by the organization committee of the Master OSS.

Data

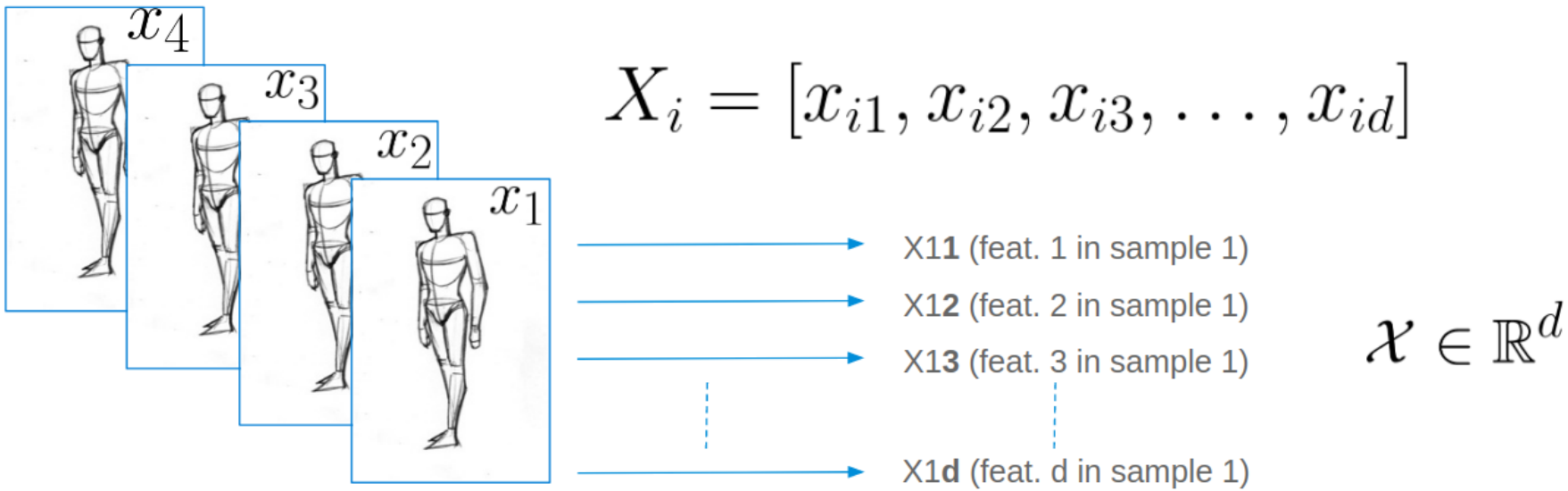


Figure: Samples and features.

Learning approaches

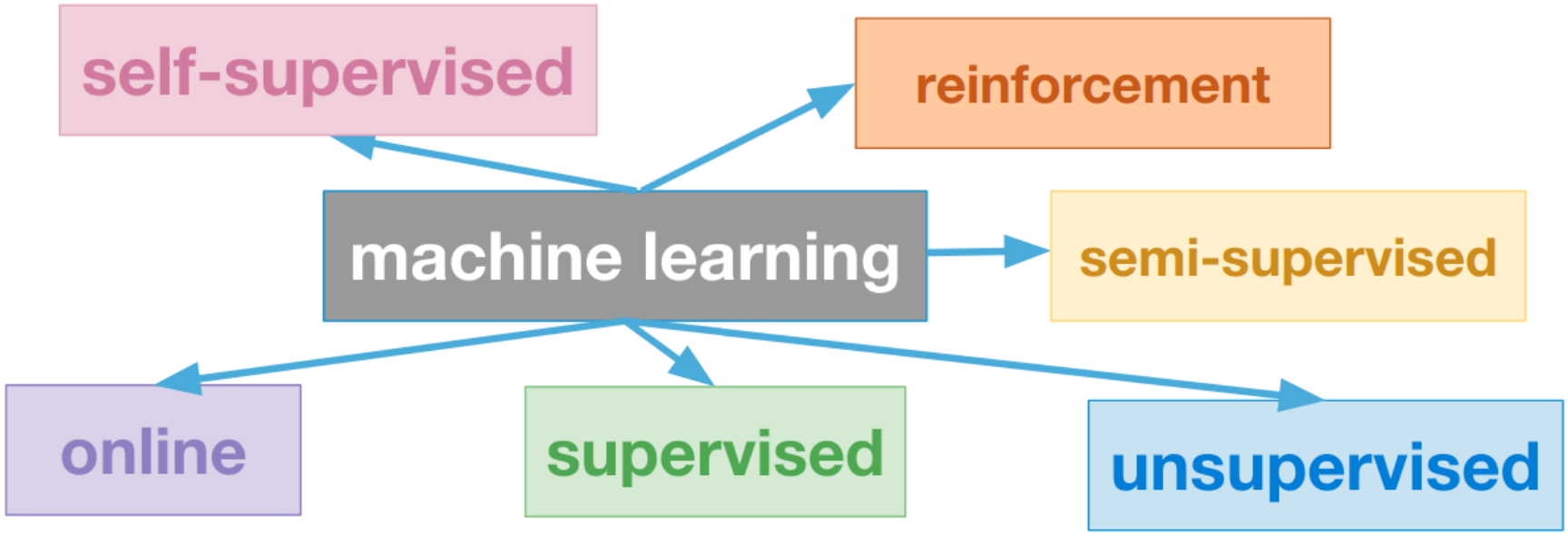
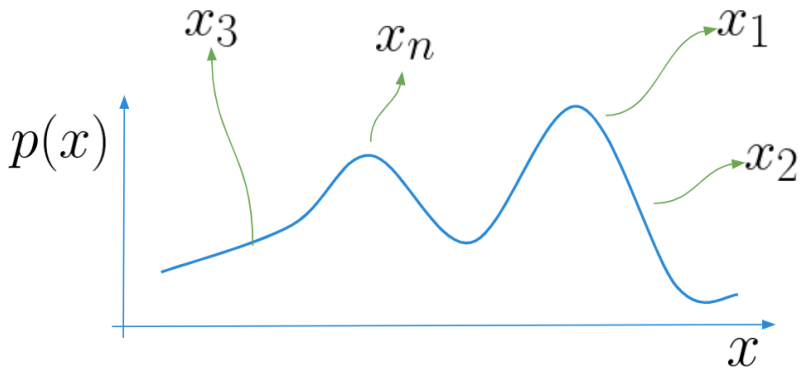


Figure: Learning approaches.

Supervised Learning

Data: a random variable



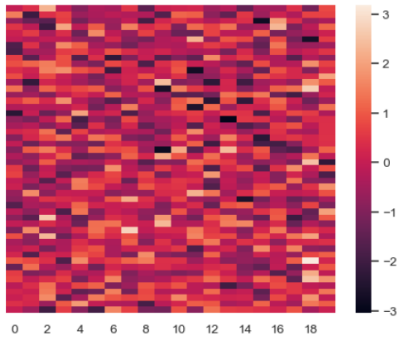
$$S : \{x_1, x_2, x_3, ..., x_n\}$$

Nature is assumed to be a random variable $\mathcal{X} \sim p(x)$ with an unknown probability density function $p(x)$. Data acquired by a sensor is gathered as a sample x from $p(x)$.

Data matrix

$$S : \{x_1, x_2, x_3, ..., x_n\}$$

$$\mathbf{X}_{(n,d)} = \begin{bmatrix} x_{00} & x_{01} & ... & x_{0d} \\ x_{10} & x_{11} & ... & x_{1d} \\ ... & ... & ... & ... \\ x_{n0} & x_{n1} & ... & x_{nd} \end{bmatrix}$$



If $x \in \mathbb{R}^d$ then by sampling n samples from $p(x)$ a $n \times d$ data matrix is built

Sampling a known PDF

$n = 10$ $n = 100$ $n = 1000$ $n = 10000$

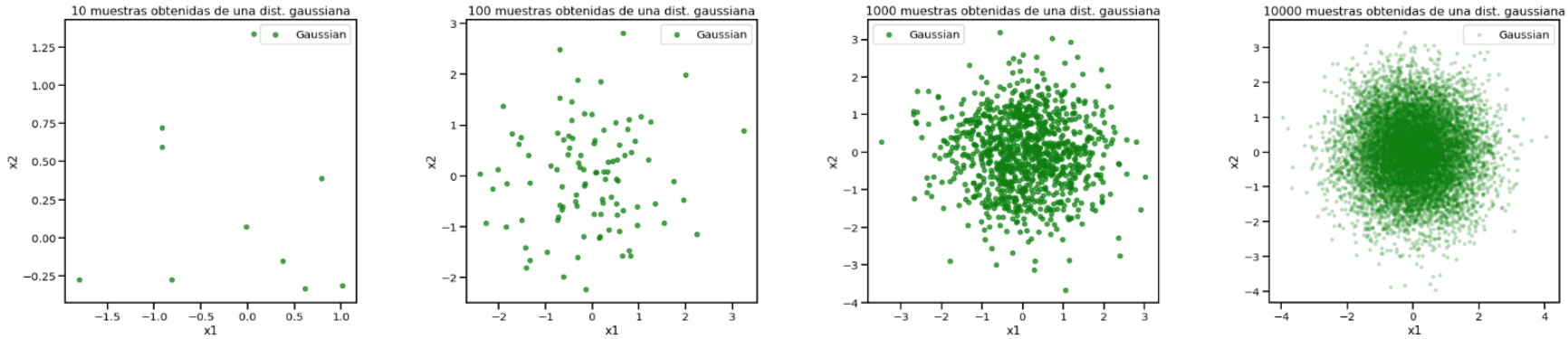


Figure: Synthetic samples from p(x) $\sim N(\mu, \sigma)$

The more samples we get from $p(x)$ the closer we are to the underlying structure (distribution) of $p(x)$.

Labeled data and supervised learning

Our data samples are expressed as vectors x of dimension d.

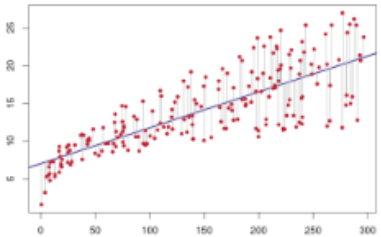
$$\begin{array}{rcl} X \in \mathcal{X} \subset \mathbb{R}^d & \rightarrow & \text{our data} \\ x_i \in X & \rightarrow & \text{our samples} \\ Y \in \mathcal{Y} \subset \mathbb{R} & \rightarrow & \text{labels} \\ y_i \in Y & \rightarrow & \text{sample labels} \end{array} \tag{1}$$

Sample set with labels

In a supervised learning approach, each sample x_i belongs to a class with label y_i and the sample set S is defined as $S = \{(x_1, y_1), \cdots, (x_n, y_n)\}$

Supervised Learning: regression and classification

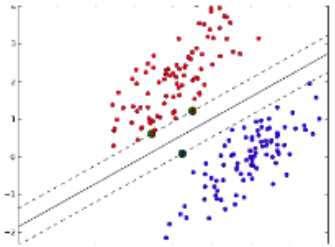
regression →



Y is a real number

$$y \subseteq \mathbb{R}$$

classification →



Y is categorical

$$y \in \{-1, 1\}$$

Figure: A labeled dataset.

Supervised Learning: classification

A supervised learning problem can present different types of labels and number of classes c. Particularly when the classes are categorical there are three approaches to consider

Binary Classification problem

There are two classes and the label vector Y has two values

$$Y = \begin{cases} +1 & \text{if } y_i = c_0 \\ -1 & \text{if } y_i = c_1 \end{cases} \tag{2}$$

Multi-class Classification problem

There are m classes and the labels dependant variable Y is represented as

$$Y \in 1, 2, 3, ..., m \quad \text{if} \quad y_i = c_k \tag{3}$$

One-class classification problem

There are only one class c_0.

Supervised Learning

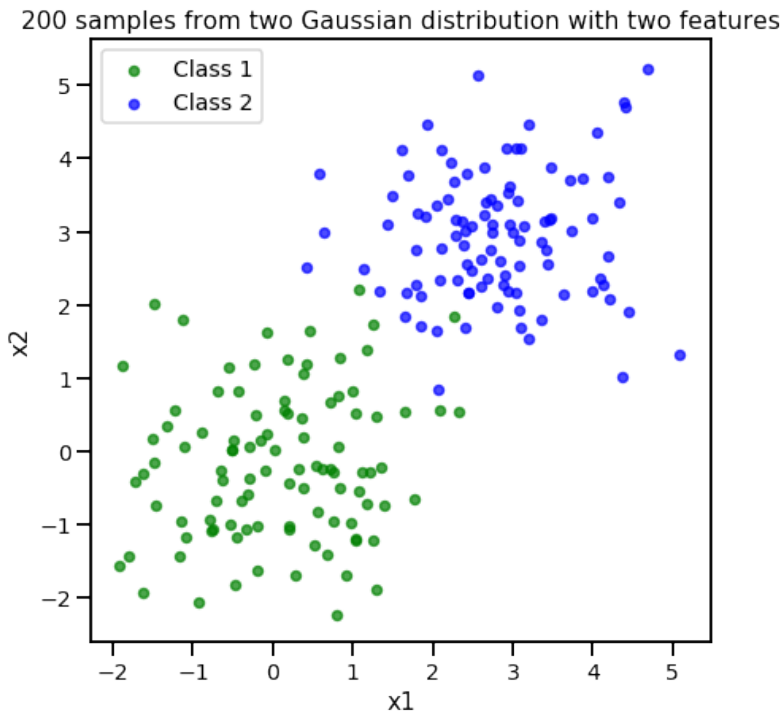


Figure: A labeled dataset.

Supervised approaches have the condition that every sample is associated to a label and the problem generally is to find the rule that explain the relation between each sample x and its label y. An unsupervised approach only consider the set of samples S.

Supervised Learning

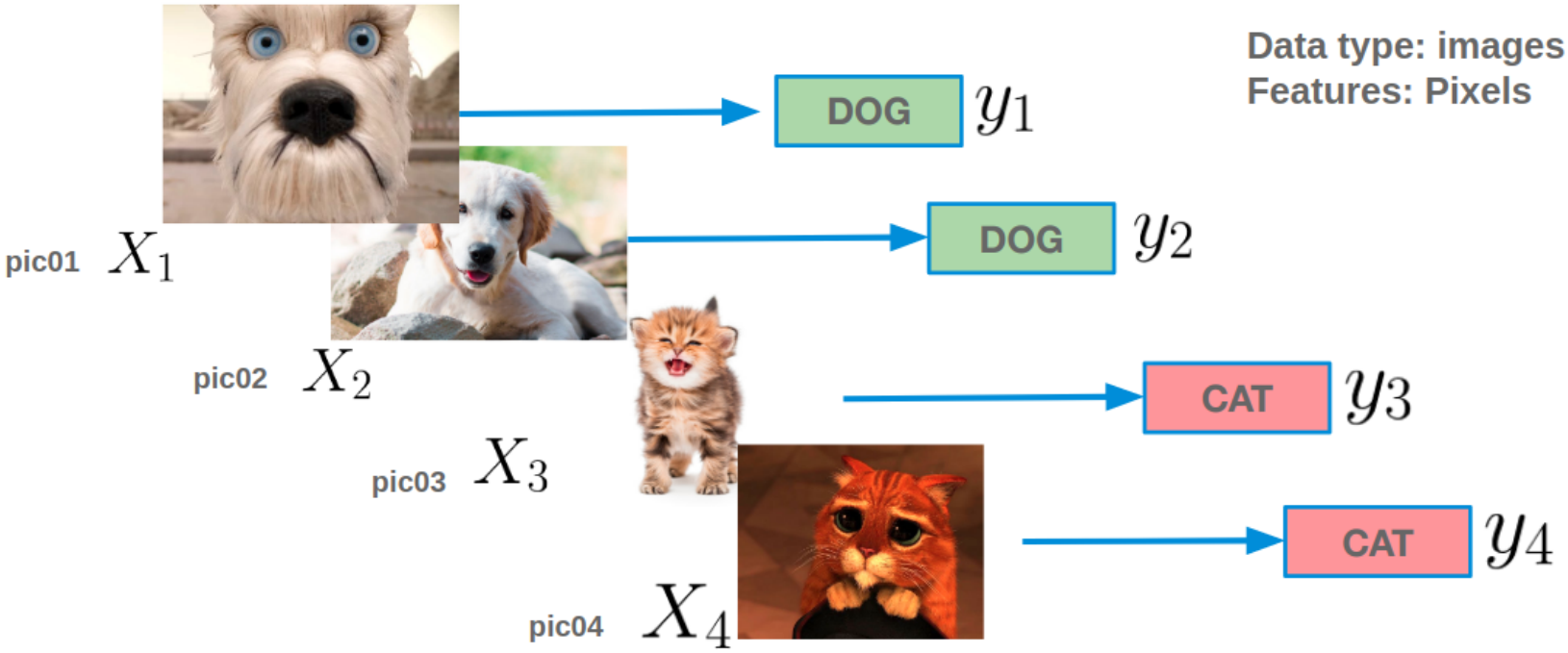


Figure: A labeled dataset.

Supervised Learning

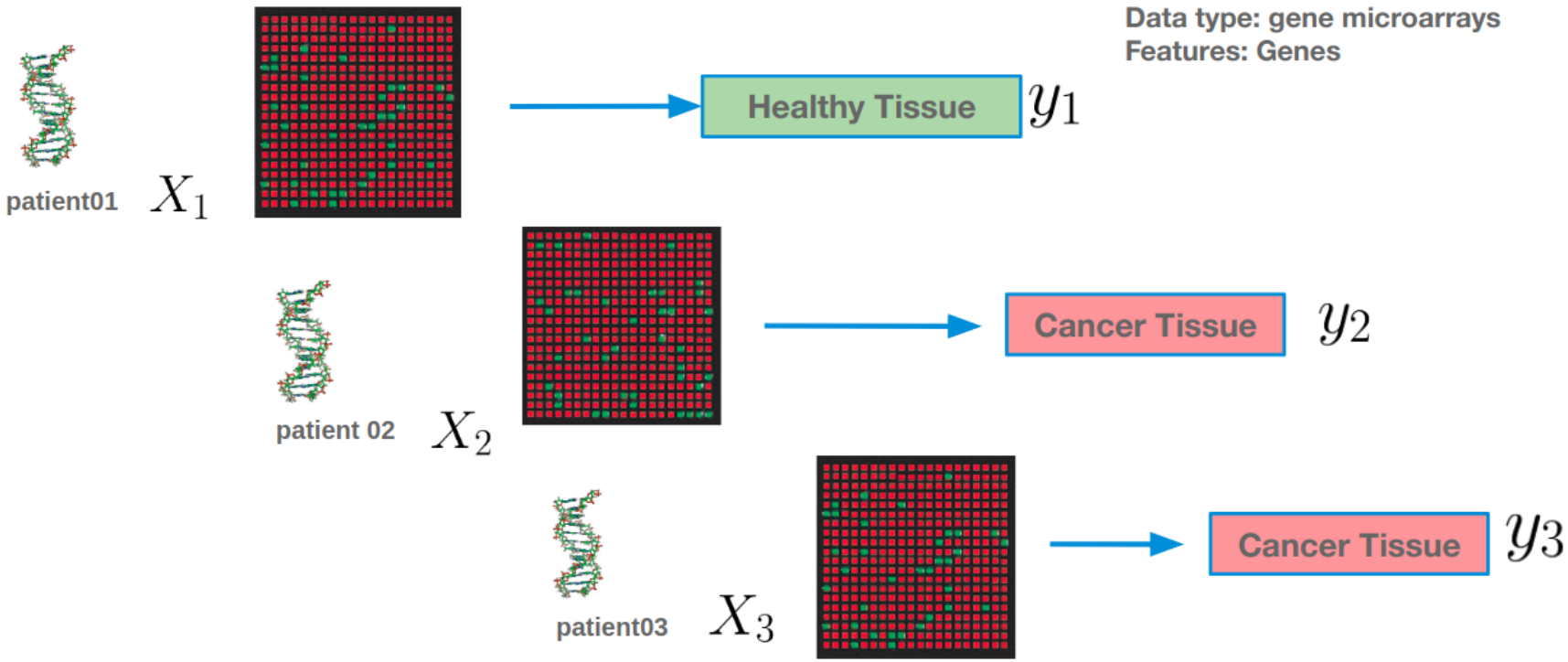


Figure: A labeled dataset.

Supervised Learning

A supervised problem invites to propose a rule $f(x)$ that explain the relation between x and y. The rule can be determined by the following approaches depending of the availability of the data.

- Parametric Learning (learn parameters of known density distributions)
- learn a rule from a set of data S
- a human to define the rule (like pathologists)

Initially we will explore rules defined by parametric learning assuming we know the probability distribution of our data. In the following chapters we will estimate the rules from a set of data without having any knowledge about the underlying distribution.

Supervised Learning

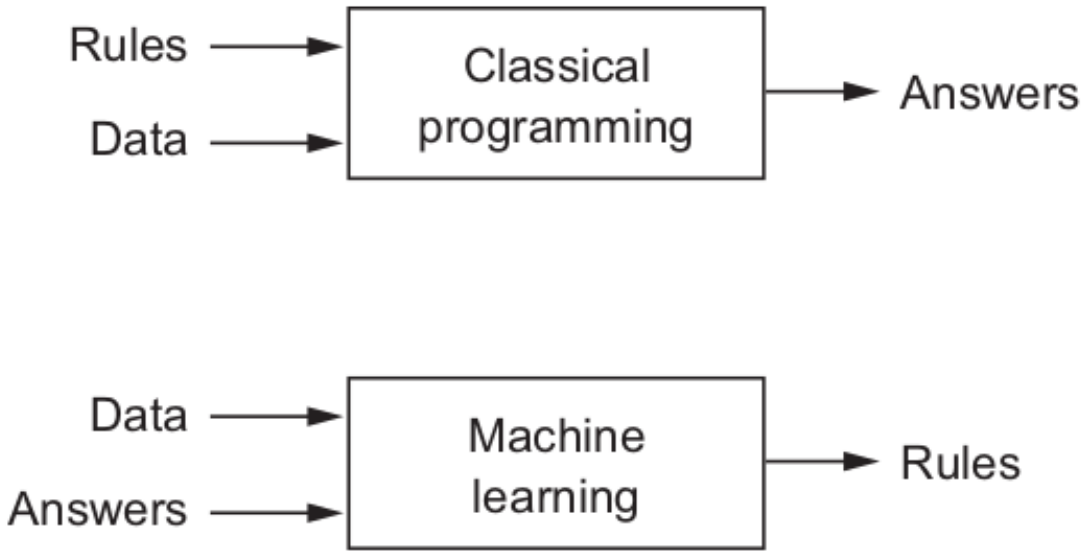


Figure: Machine Learning programming paradigm (Deep Learning with Python, Francois Chollet, 2018) [1]

The learning algorithms build rules based on data.

Classification: decision function

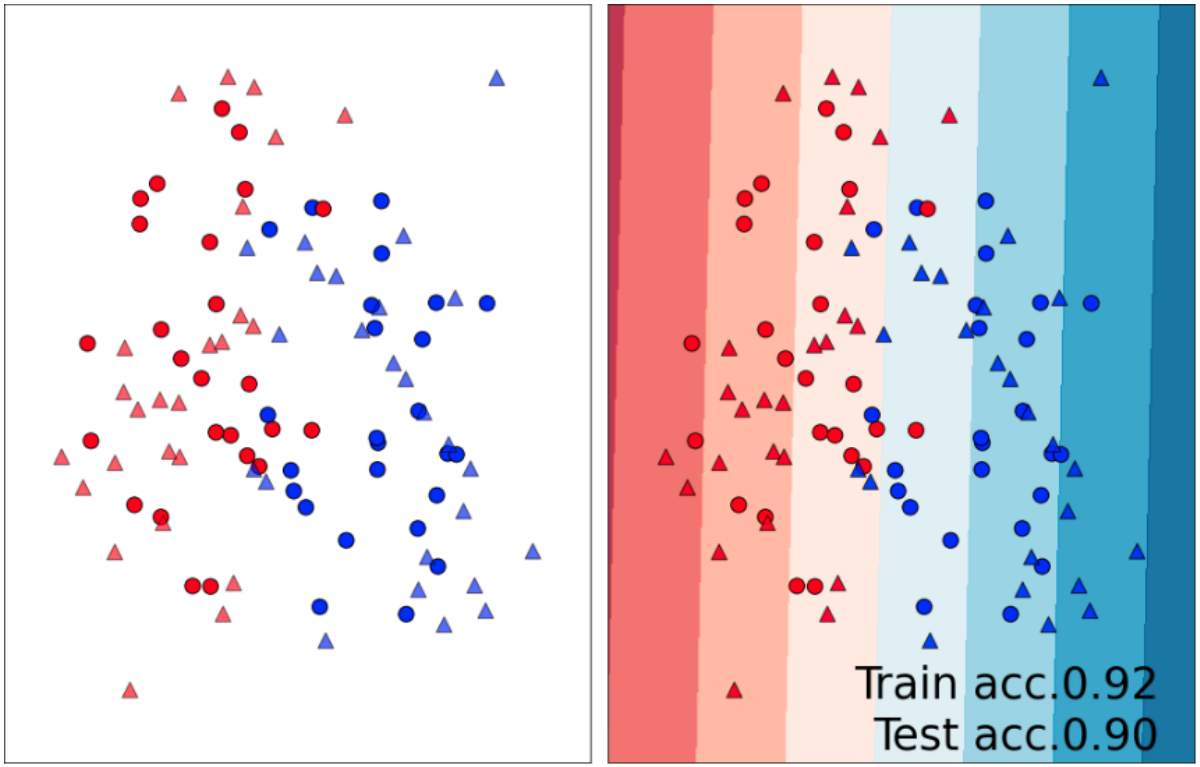


Figure: Linear decision boundary) [1]

The learning algorithms build rules based on data.

Supervised Learning: decision function

Learning f(x)

From the dataset S a function $\hat{f}_w$ defined by w parameters is learned as a decision boundary (in classification) such as $\hat{f}(x) = \hat{y}$ in order to minimize the empirical error function L

$$\min_w \ L(y, \hat{y}) = \min_w \ L(y, \hat{f}(x))$$

where w is considered as a decision parameter to find in the optimization problem.

Empirical error rate

Expected Loss

$$\mathcal{E}(f) = \mathbb{E}\left[L(y, f(x))\right]$$

Empirical error

$$\mathcal{E}(f) = \mathbb{E}\left[L(y, f(x))\right] = \frac{1}{n}\sum_{i=1}^{n} \delta(f(x_i) - y_i)$$

where

$$\delta(x) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

Bayes Decision Theory

Bayes Rule

Bayes Rule

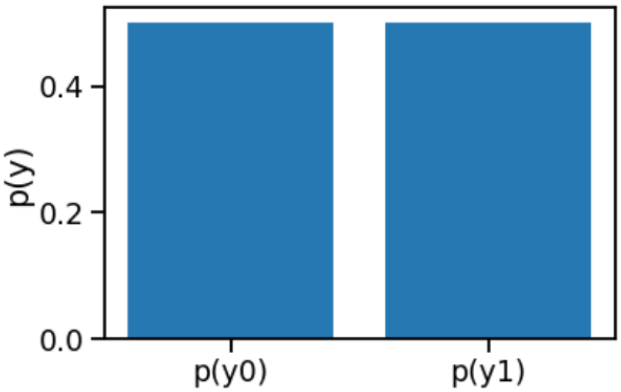
$$P(c_j \mid x) = \frac{p(x \mid c_j)P(c_j)}{p(x)}$$

Where $P(c_j)$ is the prior probability, $p(x \mid c_j)$ is the likelihood, $P(c_j \mid x)$ is the posterior and $p(x)$ is an scaling factor and

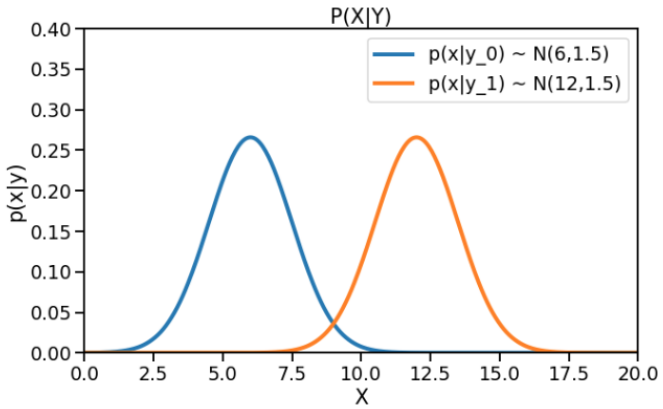
$$x \in \mathbb{R}^d$$

with d as the dimensionality of the random variable $\mathcal{X}$

Bayes Rule



$$p(Y)$$

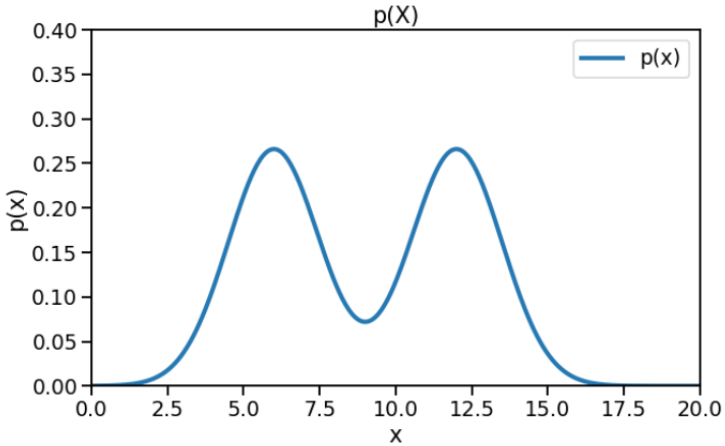


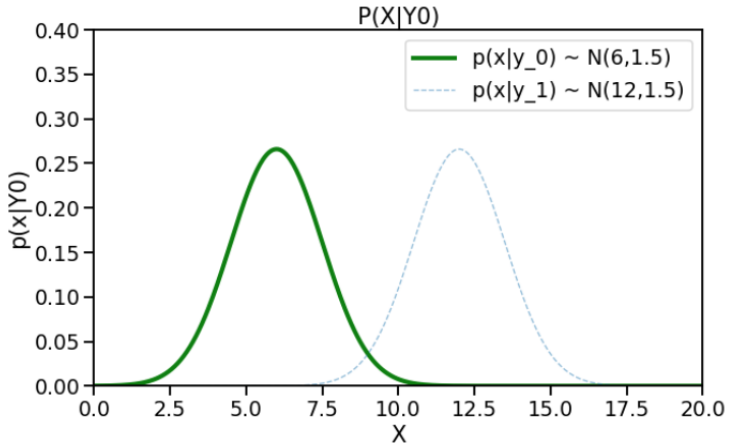
$$p(X \mid Y)$$

$$p\left(x \mid Y_i\right) \sim N(\mu, \sigma)$$

Figure: Left: class priors. Right: Class Likelihoods

Bayes Rule



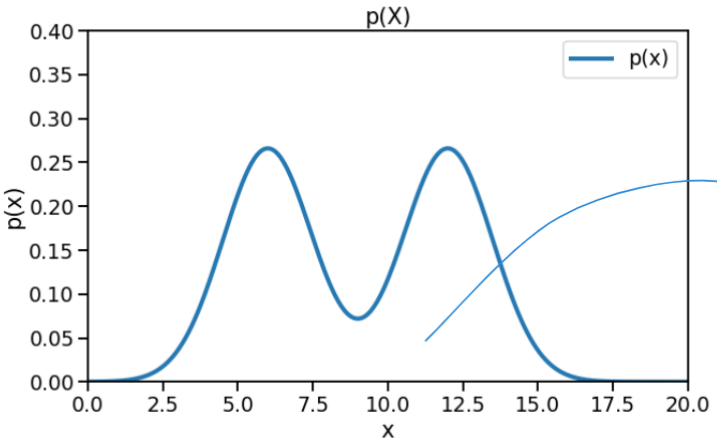


$$p(X) = \sum_{Y} p(X, Y)$$

$$p(X \mid Y)$$

Figure: Left: Data distribution and joint class probability. Right: class Likelihood.

Bayes Rule

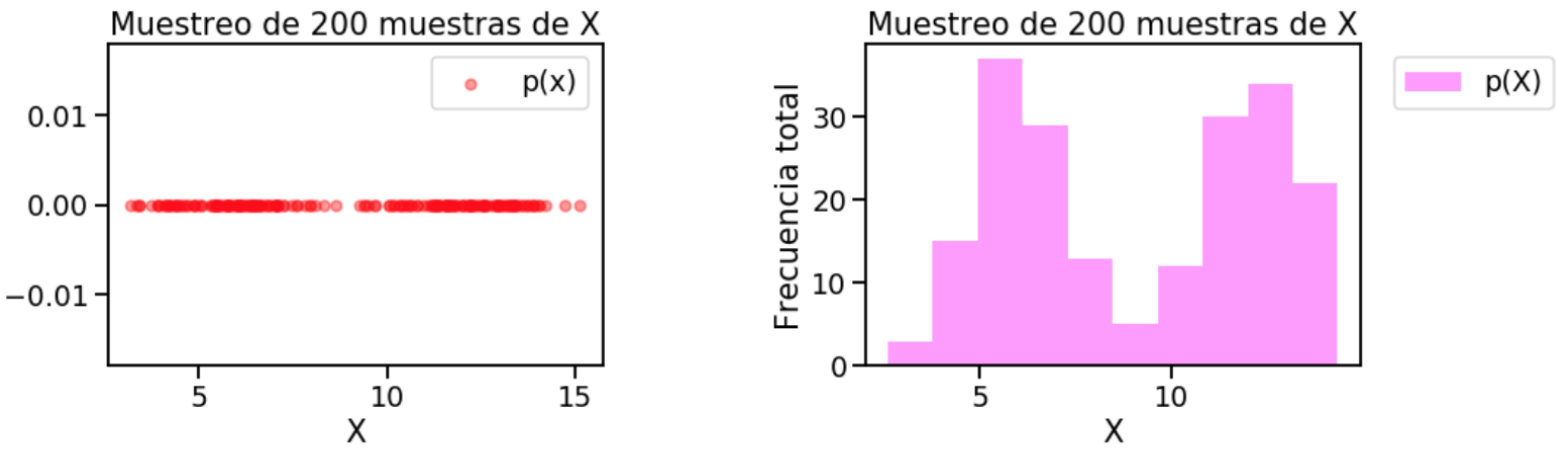


???

$$p(Y \mid X) = \frac{p(X \mid Y)p(Y)}{p(X)}$$

Figure: Posterior probability is an unknown.

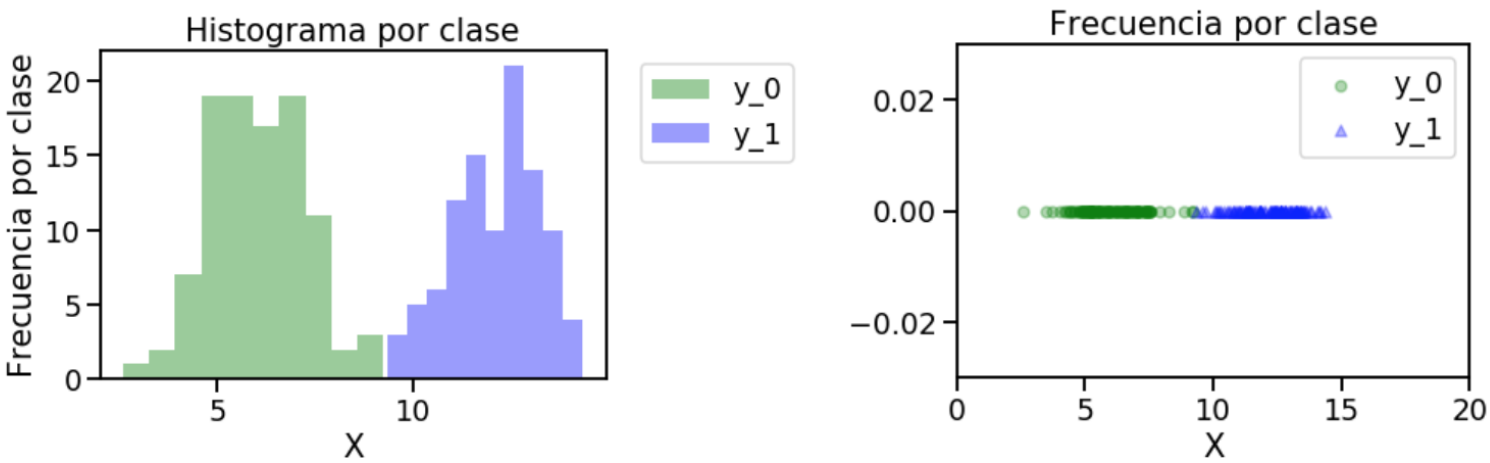
Bayes Rule



$$p(X)$$

Figure: Data distribution $p(x)$.

Bayes Rule



$$p(x \mid Y_0) \sim N(\mu = 6, \sigma = 1.5)$$
$$p(x \mid Y_1) \sim N(\mu = 12, \sigma = 1.5)$$

Figure: Likelihood of each class

Decision boundary by Bayes Rule

Suppose that we want to classify samples generated by a random variable of known distribution and parameters. We want to build a decision boundary given a set of classes $C = \{C_0, C_1\}$ using Bayes rule. We have a set of data generated by a Gaussian distribution for each class with same variance and different mean which are known parameters:

$$p(x|c_j) \sim N(\mu, \sigma)$$

with

$$p(x|c_j) \sim \frac{1}{\sqrt{2\pi\sigma^2}} \mathbf{exp}\left[\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

and the prior probabilities $P(c_0) = P(c_1) = 0.5$, the variance of the distribution as $\sigma_0^2 = \sigma_1^2 = 1$ and the means $\mu_0 = 6$ and $\mu_1 = 12$.

Decision boundary by Bayes Rule

From the Bayes expression we want to decide c_0 if $P(c_0 \mid x) > P(c_1 \mid x)$, else we decide c_1. Then

$$P(c_j \mid x) = \frac{p(x \mid c_j)P(c_j)}{p(x)} \tag{4}$$

$$\frac{p(x \mid c_0)P(c_0)}{p(x)} > \frac{p(x \mid c_1)P(c_1)}{p(x)} \tag{5}$$

$$p(x \mid c_0)P(c_0) > p(x \mid c_1)P(c_1) \tag{6}$$

Decision boundary by Bayes Rule

$$\frac{p(x \mid c_0)}{p(x \mid c_1)} > \frac{P(c_1)}{P(c_0)} \tag{7}$$

$$\frac{p(x \mid c_0)}{p(x \mid c_1)} > \frac{0.5}{0.5} \tag{8}$$

$$\frac{p(x \mid c_0)}{p(x \mid c_1)} > 1 \tag{9}$$

$$p(x \mid c_0) > p(x \mid c_1) \tag{10}$$

We can compute the equation 10 using the input values of the distribution given above :)

Decision boundary by Bayes Rule

$$\frac{1}{\sqrt{2\pi\sigma^2}}\textbf{exp}\left[\frac{-1}{2}\left(\frac{x-\mu_0}{\sigma}\right)^2\right] > \frac{1}{\sqrt{2\pi\sigma^2}}\textbf{exp}\left[\frac{-1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2\right] \quad (11)$$

Remember we have different mean and the same variance, then the first term is not necessary

$$\textbf{exp}\left[\frac{-1}{2}\left(\frac{x-\mu_0}{\sigma}\right)^2\right] > \textbf{exp}\left[\frac{-1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2\right] \quad (12)$$

Then compute the decision function $f(x)$ from the input parameters μ and σ of each class.

Decision boundary by Bayes Rule

Find the Bayes decision rule using the known parameters of the probability distributions and the prior probabilities.

Exercise 01

Same variance, different mean, same prior and 1 feature:

- $P(c_0) = P(c_1) = 1/2$
- $\sigma_0^2 = \sigma_1^2 = 1$
- $\mu_0 = 6$ and $\mu_1 = 12$

How is the obtained decision function?

Decision boundary by Bayes Rule

Find the Bayes decision rule using the known parameters of the probability distributions and the prior probabilities.

Exercise 02

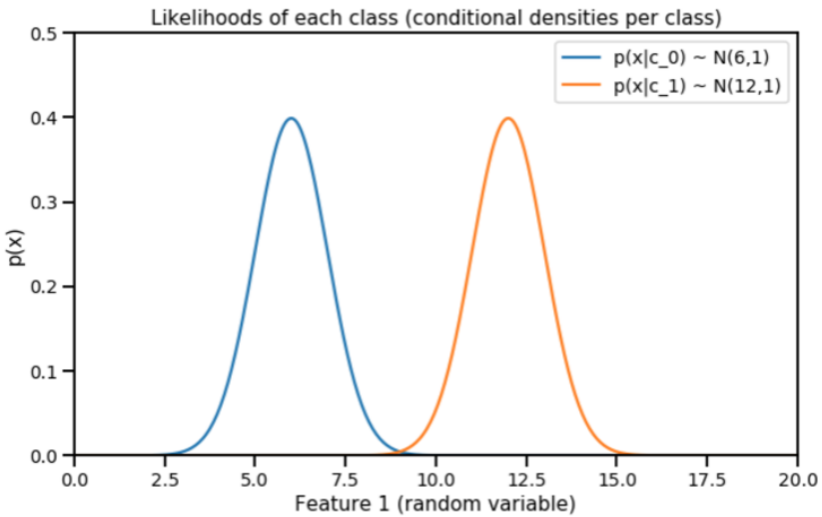
Same variance, different mean, different prior and 1 feature:

- $P(c_0) = 3/4$ and $P(c_1) = 1/4$
- $\sigma_0^2 = \sigma_1^2 = 1$
- $\mu_0 = 6$ and $\mu_1 = 12$

How is the obtained decision function?

Decision boundary by Bayes Rule

$$P(C_0) = P(C_1) = 0.5$$



$$P(C_0) = 0.75,\ P(C_1) = 0.25$$

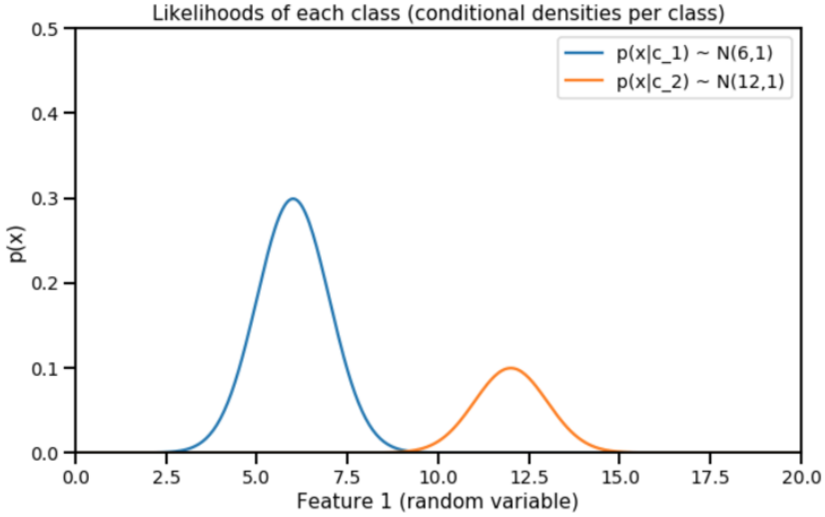


Figure: Exercise 01 and 02 probability density functions of each class.

If we know the family of PDF, its parameters and its prior probabilities then we can determine analytically the decision function.

Decision boundary by Bayes Rule

Exercise 03

For each of the two previous exercises use the probability density function of each class to generate 100 new samples. Then by using the previously obtained decision boundaries compute the empirical error rate for the 100 new samples per clas per exercise.

Decision boundary by Bayes Rule

Find the Bayes decision rule using the known parameters of the probability distributions and the prior probabilities.

Exercise 04

Same variance, different mean, same prior and 1 feature:

- $P(c_0) = P(c_1) = 1/2$
- $\sigma_0^2 = \sigma_1^2 = 2$
- $\mu_0 = 6$ and $\mu_1 = 12$

Exercise 05

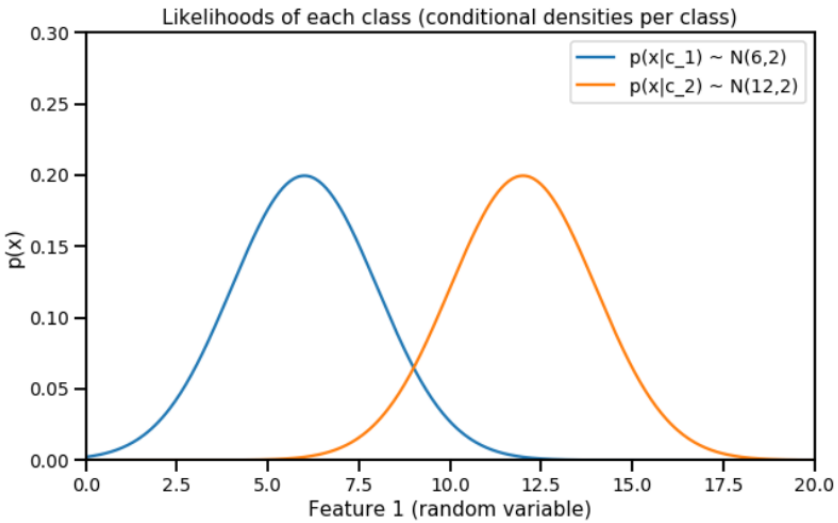
Same variance, different mean, different prior and 1 feature:

- $P(c_0) = 3/4$ and $P(c_1) = 1/4$
- $\sigma_0^2 = \sigma_1^2 = 2$
- $\mu_0 = 6$ and $\mu_1 = 12$

How is the obtained decision function?

Decision boundary by Bayes Rule

$$P(C_0) = P(C_1) = 0.5$$



$$P(C_0) = 0.75,\ P(C_1) = 0.25$$

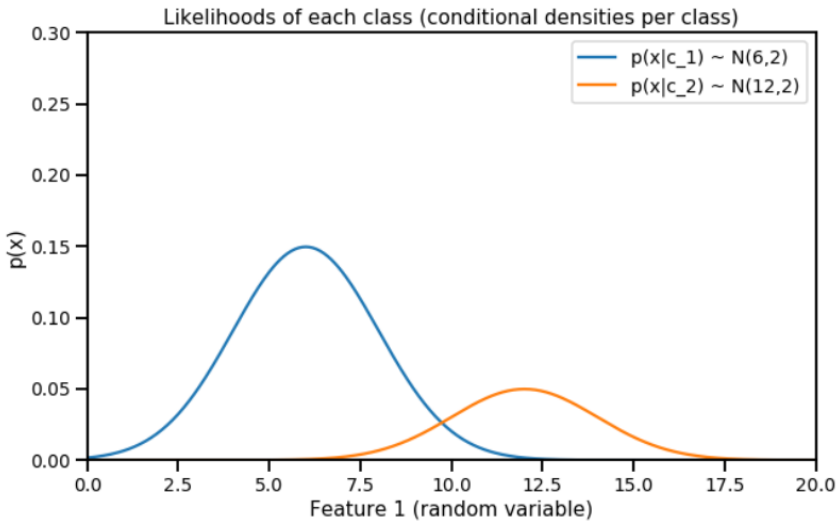


Figure: Exercise 03 and 04 probability density functions of each class.

If we know the family of PDF, its parameters and its prior probabilities then we can determine analytically the decision function.

Decision boundary by Bayes Rule

Exercise 06

For each of the two previous exercises use the probability density function of each class and generate 100 new samples. Then compute the empirical error rate considering the obtained decision rule.

Decision boundary by Bayes Rule

Find the Bayes decision rule using the known parameters of the probability distributions and the prior probabilities.

Exercise 07

Different variance, different mean, same prior and 1 feature:

- $P(c_0) = P(c_1) = 1/2$
- $\sigma_0^2 = 3$
- $\sigma_1^2 = 1$
- $\mu_0 = 6$ and $\mu_1 = 12$

How is the obtained decision function?

Exercise 08

For the exercises 07 use the probability density function of each class and generate 100 new samples. Then compute the empirical error rate considering the obtained decision rule.

Decision boundary by Bayes Rule

We can extend the Bayes decision rule to a multi-variate context. This time using two features x_1 and x_2.

$$p(X|c_j) \sim (2\pi)^{-d/2}|\Sigma^{-1/2}|\mathbf{exp}\left[-\frac{1}{2}(X-M)^t\,\Sigma^{-1}\,(X-M)\right]$$

Then the problem is defined as

$$X = [x_1, x_2]; M_1 = [0,0]; M_2 = [2,2]; P(C_1) = P(C_2) = 0.5$$

with a co-variance matrix

$$\Sigma_1 = \Sigma_2 = \begin{bmatrix} \sigma_{00}^2 & \sigma_{01}^2 \\ \sigma_{10}^2 & \sigma_{11}^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I$$

So the problem presents two classes generated by normal distributions of equal variance, equal prior probabilities and different means.

Decision boundary by Bayes Rule

We want to decide c_0 if $P(c_0 \mid x) > P(c_1 \mid x)$, else we decide c_1. Using the equation 10 obtained from the problem with equal variance, equal priors and different mean we can express the problem as

$$q(x) = \frac{p(x \mid c_0)}{p(x \mid w_1)} > 1$$

$$log(q(x)) = [-1/2(X-M_0)^T\Sigma^{-1}(X-M_0)]-[-1/2(X-M_1)^T\Sigma^{-1}(X-M_1)]$$

$$(M_0 - M_1)^T\Sigma^{-1}X + 1/2M_1^T\Sigma^{-1}M_1 - 1/2M_0^T\Sigma^{-1}M_0 > 1$$

Decision boundary by Bayes Rule

Exercise 09

Given the multi-variate context, find the decision function. Once it is defined, generate 200 random samples from the probability density function of each class and compute the empirical error of the obtained boundary.

About the model

What assumption taken into account in the model do you think is naive?

About the decision boundary

How is the obtained decision function?

Maximum Likelihood parameter estimation

Maximum Likelihood parameter estimation

Figure: We assume that the conditional probability density function of each class is a Gaussian distribution.

There are cases where we know the family of the probability density function of each class but the parameters are unknown. Via MLE we can estimate the parameters using data coming from these distributions. Finally with the obtained parameters we can define a decision function via Bayes rule.

Maximum Likelihood parameter estimation

The conditional gaussian distribution for a given class c is

$$p(X|M,\Sigma) \sim (2\pi)^{-d/2}|\Sigma^{-1/2}|\mathbf{exp}\left[-\frac{1}{2}(X-M)^t\,\Sigma^{-1}\,(X-M)\right]$$

given a set of data $X = (x_1, x_2, x_3, ..., x_n)^t$ drawn from a gaussian distribution, its log-likelihood is [2]

$$\ln[p(X|M,\sigma)] = -\frac{Nd}{2}ln(2\pi) - \frac{N}{2}|\Sigma| - \frac{1}{2}\sum_{n=1}^{N}(X_n - M)^t\Sigma^{-1}(X_n - M)$$

Then applying the derivative of the log likelihood with respect to M

$$\frac{\partial}{\partial M}ln[p(x|M,\Sigma)] = \sum_{n=1}^{N}\Sigma^{-1}(X_n - M)$$

Maximum Likelihood parameter estimation

Finally, making the derivative equal to zero let us estimate the value of $\hat{\boldsymbol{M}}$ as

$$\hat{\boldsymbol{M}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

and we obtain the maximum likelihood estimate of the mean. For the covariance parameter

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{n=1}^{n}(\boldsymbol{x}_n - \hat{M})(\boldsymbol{x}_n - \hat{M})^t$$

Maximum Likelihood parameter estimation

Exercise 10

Given 100 random samples of each class generated by the conditional density functions of exercise 09 estimate the mean vector $\hat{M}$ and its co-variance matrix $\hat{\Sigma}$ for both distributions. Compare these estimations with the true parameters. Finally plug the new parameters in the Bayes framework and find an estimated decision boundary between the two classes.

Conclusions

Decision functions by parametric approaches to density modelling

We have considered Gaussian PDF with different mean, equal or different variance, equal or different prior probabilities and with known and unknown parameters. What can we conclude about the obtained classification boundaries?

Chollet, F. (2018). Keras: The python deep learning library. Astrophysics Source Code Library.

Bishop, C. M. (2006). Pattern recognition and machine learning. springer.