Model-based Statistical Learning

Pr. Charles BOUVEYRON

Professor of Statistics
Chair of the Institut 3IA Côte d'Azur
Université Côte d'Azur & Inria

charles.bouveyron@univ-cotedazur.fr
@cbouveyron

Preamble

"Ce qui est simple est toujours faux.
Ce qui ne l'est pas est inutilisable."

Paul Valéry

Outline

The AI revolution hasn't happened yet!

Artificial intelligence is a strategic field of research:

- with direct application in most scientific fields (Medicine, Biology, Astrophysics, Humanities)
- and with probably the most impact in innovation and transfer (health, transport, defense).

The AI revolution hasn't happened yet!

Artificial intelligence is a strategic field of research:

- with direct application in most scientific fields (Medicine, Biology, Astrophysics, Humanities)
- and with probably the most impact in innovation and transfer (health, transport, defense).

The recent and impressive NN results should not hide the remaining issues:

- deep learning has impressive results in a few specific cases and with a high-level supervision,
- use of DL techniques in various fields are promising but not well understood.

The AI revolution hasn't happened yet!

Artificial intelligence is a strategic field of research:

- with direct application in most scientific fields (Medicine, Biology, Astrophysics, Humanities)
- and with probably the most impact in innovation and transfer (health, transport, defense).

The recent and impressive NN results should not hide the remaining issues:

- deep learning has impressive results in a few specific cases and with a high-level supervision,
- use of DL techniques in various fields are promising but not well understood.

"*Artificial Intelligence: the revolution hasn't happened yet*"
M. Jordan (UC Berkley)

Open problems of AI

Some open problems are critical:

- reliability of models and algorithms,
- handling data heterogeneity (categorical, functional, networks, images, texts, ...),
- unsupervised learning (clustering, dimension reduction),
- learning from HD and small data (n small / p large),

Open problems of AI

Some open problems are critical:

- reliability of models and algorithms,
- handling data heterogeneity (categorical, functional, networks, images, texts, ...),
- unsupervised learning (clustering, dimension reduction),
- learning from HD and small data (n small / p large),

Combination of statistical theory with deep learning techniques is certainly the future of AI!

AI in France

French policy for AI:

- C. Villani presented in March a recommendation report for AI,
- President Macron announced the creation of a network of AI institutes.

The 3IA institutes:

- 12 french research centers applied for the 3IA call in Sept.,
- 4 projects have been selected in the Spring 2019:
  - Paris, Toulouse, Grenoble
  - and Nice!

A few examples: Cervical cancer detection

Cervical cancer detection:

- it is an important public health field which is currently treated mostly manually,
- screening by human experts is complicated by the amount of cells (20 000/smear),
- and by the very small proportion of cancer cells (less than 1%).

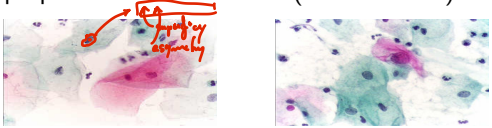


Figure: Normal (left) and abnormal (right) pap smears.

Classification is useful in this context:

- for building supervised classifiers which can select the most likely cancer cells,
- for helping experts in labeling the learning data through weakly-supervised classification,
- for selecting discriminative variables which can be used in a semi-automatic process.

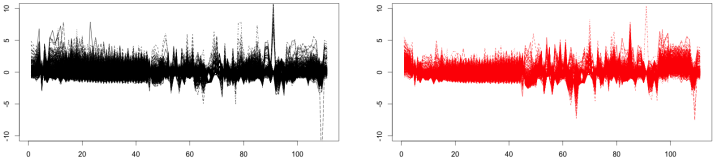


Figure: Control and (cervical) cancer data.

A few examples: Sparse models in Medicine (HEGP)

Problem:

- overcome the curse of dimensionality that occurs in Metabolomics,
- for disease diagnostic and early-stage marker identification,
- metabolomic data fall into the "ultra-high dimensional data" case.

Our solution:

- a Bayesian variable selection technique for PCA,
- that identify the relevant variable for each stage of the disease.

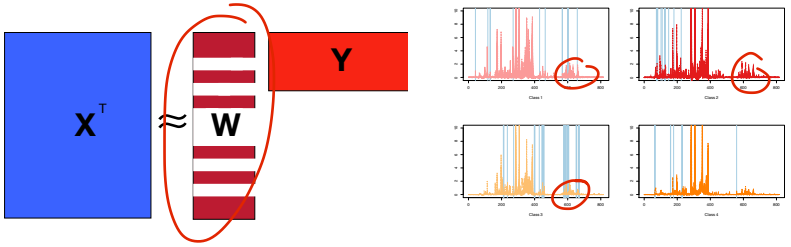


Figure: gsPPCA and variable selection on MNR specrta for CKD diagnosis.

Analysis of massive functional data (Linky / EDF)

Problem:

- Linky meters will allow EDF to have access to 27 million of Linky data,
- data are functional data and are measured every 30 minutes -> 17 520 obs./year,
- necessity to summarize those massive data before exploitation.

Our solution:

- a statistical co-clustering technique for functional data,
- that form homogeneous groups of both individuals and days.

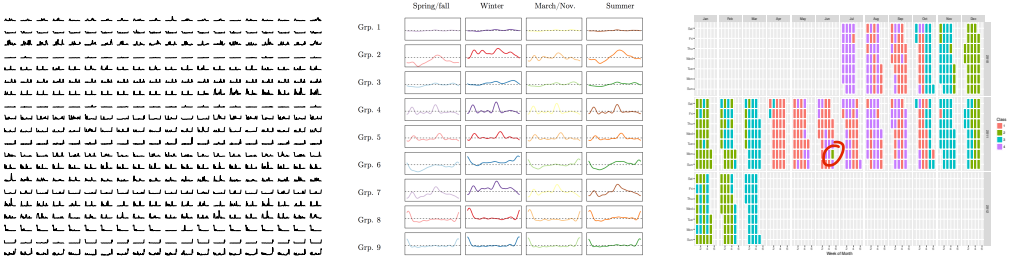


Figure: Functional co-clustering of Linky data (EDF).

Outline of the course

The course will be organized as follows:

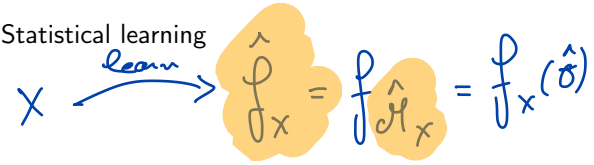
1. Introduction to model-based statistical learning (CB)
2. Linear models for classification (PAM)
3. Mixture models and the EM algorithm (CB)
4. Another view on the EM algorithm (PAM)
5. Practical work (1st evaluation, PAM)
6. Between supervised and unsupervised classification (PAM)
7. Practical work (2nd evaluation, CB)
8. Missing values (PAM)
9. Model-based image analysis (CB)
10. Co-clustering (CB)

Outline

Learning from data...

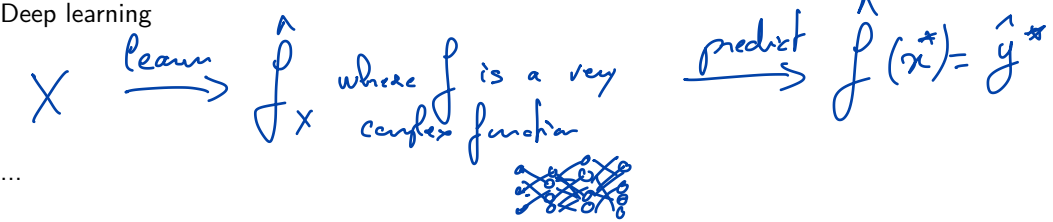
One task, several families of approaches:

- Statistical learning



- Machine learning

- Deep learning



- ...

Learning from data...

Learning is a two-head problem:

Supervised

(X, Y) ← target variable
↑ explanatory variables

in this supervised context, we need examples of both variables to learn the predictor f

$(X, Y) \longrightarrow f_{X,Y}$

Unsupervised

In this situation, we only observe X and we would like to infer Y from it.

semi-supervised learning.

Learning from data...

Methods are specific to each task:

Supervised

- classification
  Y is categorical
- regression
  Y is continuous.
- time series forecasting
- ...

Unsupervised

- clustering
  $X \xrightarrow{\text{predict}} Y$ is categorical
- dimension reduction / representation learning
  $X \xrightarrow{\text{predict}} Y$ is (multivariate) continuous
- image denoising

Supervised learning

Supervised learning is also a field with different sub-tasks:

- classification:

(X, y) ← y is categorical

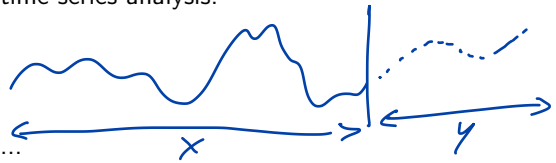
- Logistic regression
- decision trees
- SVM
- Naive Bayes
- LDA
- k-NN

- regression:

(X, y) ← y is continuous

- Linear regression
- SVM
- d.trees
- k.NN.

- time series analysis:

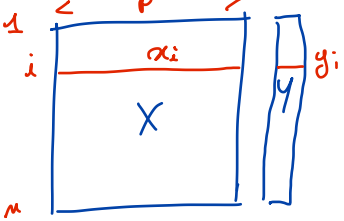


- ARIMA
- ....

- ...

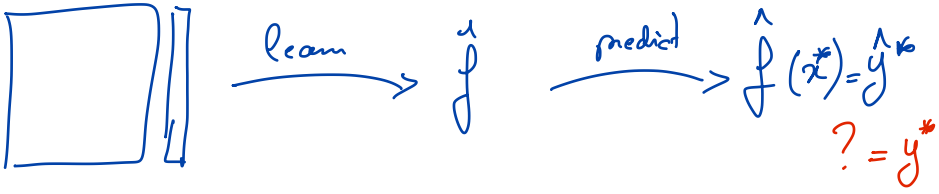
The supervised learning process

The material: a set of (complete) data



(X, y) is called the learning data set.

The goal: learn a predictor $f(.)$ from the (complete) data



Measuring the learning performance

One comfortable thing of working in the supervised context is:

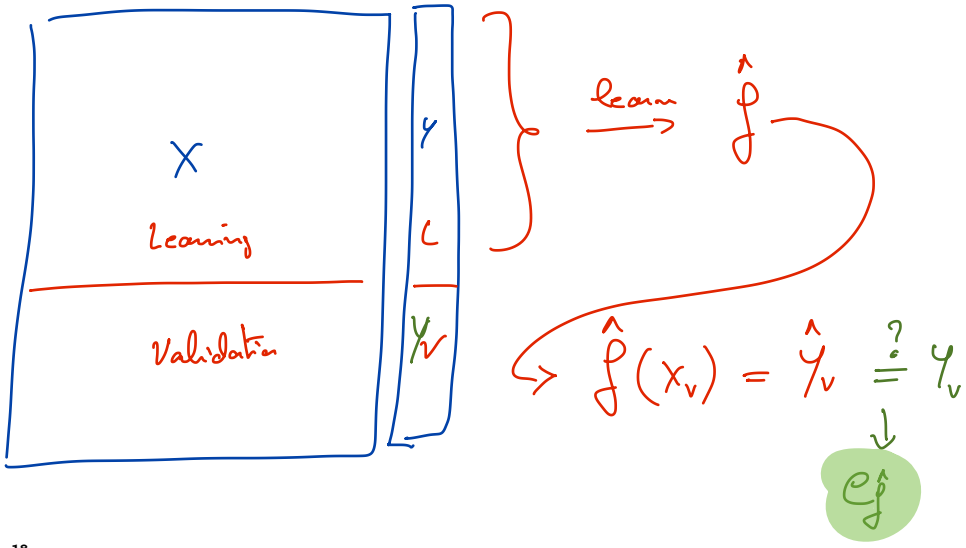
- to be able to measure the performance of the learned predictor,
  - classification error : $e_{\hat{f}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\hat{y}_i \neq y_i\} \in [0,1]$
  - regression error : $MSE = \frac{1}{n} \sum_{i=1}^{n} \| \hat{y}_i - y_i \|^2 \geq 0$
- compare several predictors and pick the most efficient one.

$f_1 \longrightarrow e_{\hat{f}_1} = 0.05$

$f_2 \longrightarrow e_{\hat{f}_2} = 0.04$

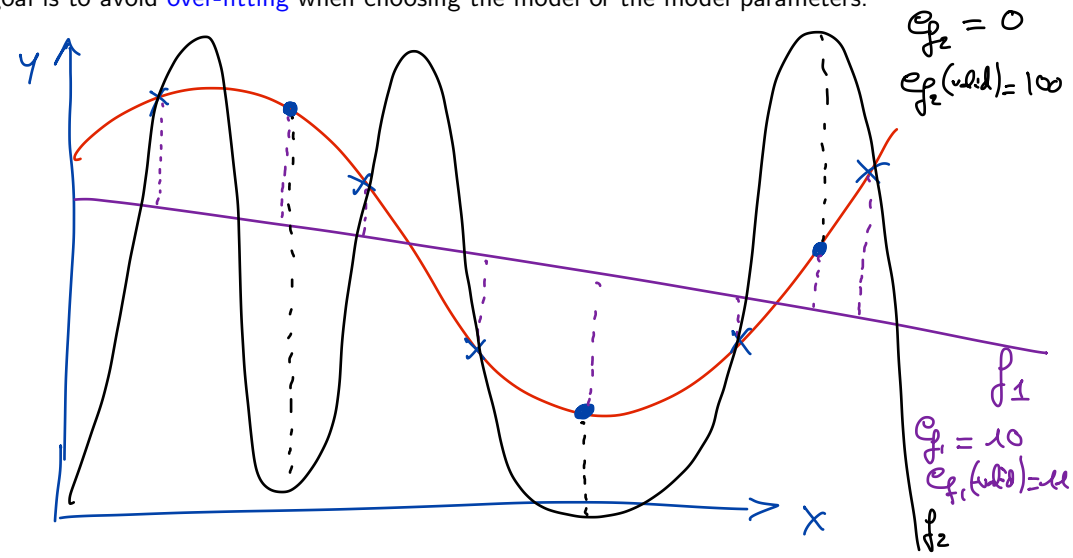
A minimal setup for supervised learning

The minimal setup for building a supervised predictor $f()$ from data is as follows:



Why such a minimal setup?

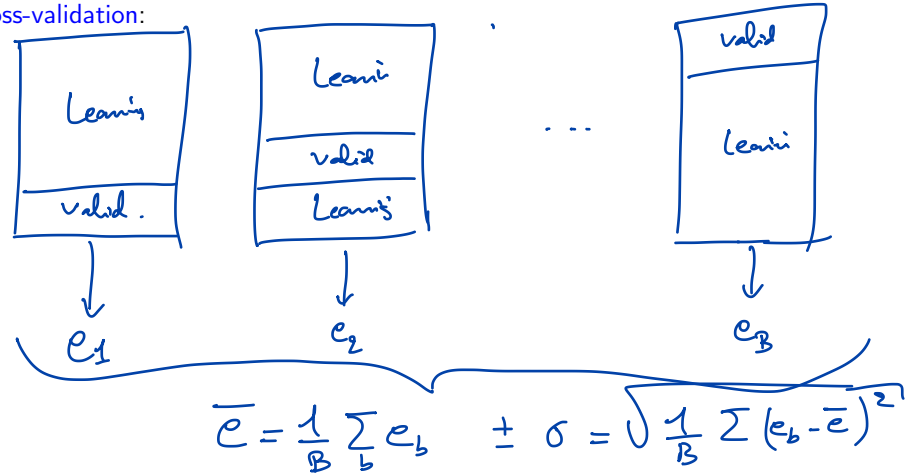
The goal is to avoid over-fitting when choosing the model or the model parameters:



An advanced setup for supervised learning

Resampling techniques:

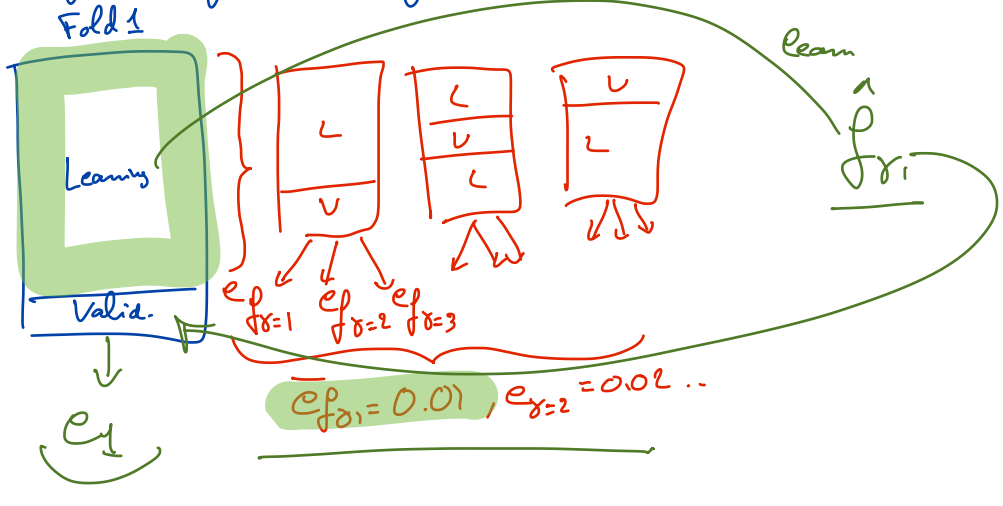
- there are several methods (leave-one-out, V-fold cross-validation, bootstrap) depending on the context (sample size, computing time, ...),
- V-fold cross-validation:



$e_{f_1} = 0.05 \pm 0.02$

$e_{f_2} = 0.04 \pm 0.08$

In the case of comparing methods with tuning parameters, we have to use double-CV to evaluate correctly the average performance of the method.



Outline

What is model-based statistic learning?

Model-based stat. learning methods are the methods that assume a model supposed to have generated the data at hand.

All these methods assume that it exists a statistical data generation process.

Ex: the linear model for regression.

$$\mathcal{M} \quad \begin{cases} Y = \beta X + \varepsilon \\ \varepsilon \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

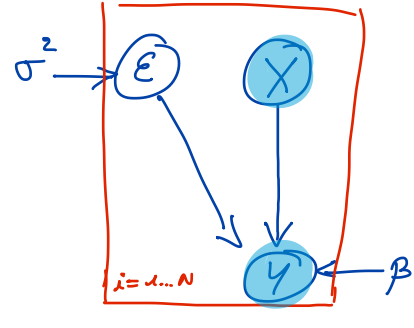
Graphical models to describe stat. models:

In graphical model, we represent each random var. by a node of a network, and the relationships between the variables are represented by (directed) edges.

We usually highlight the variables that are observed (here in blue).

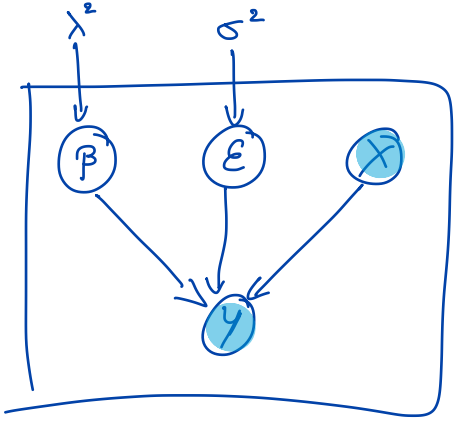
For instance, for the linear model:

$$\begin{cases} Y = \beta X + \varepsilon \\ \varepsilon \sim \mathcal{N}(0, \sigma^2) \end{cases}$$



For example, the Bayesian linear model is as follows:

$$\begin{cases} Y = \beta X + \varepsilon \\ \varepsilon \sim N(0, \sigma^2) \\ \beta \sim N(0, \lambda^2) \end{cases}$$



What is model-based statistic learning?

Among the MB stat. learning methods:

- the linear model
- the logistic regression model $\text{logit}(Y) = \beta X + \epsilon$
- Gaussian mixtures $p(X) = 0.2\, \mathcal{N}(x; 1, 1) + 0.8\, \mathcal{N}(x; 2, 1)$
- Probabilistic PCA: $Y = WX + \epsilon$ with $X \sim \mathcal{N}(0, I_d)$, $\epsilon \sim \mathcal{N}(0, \sigma^2 I_p)$
- (...)

Its place in the statistical / machine learning field

MB stat. techniques are covering most of the learning field : supervised and unsupervised learning.

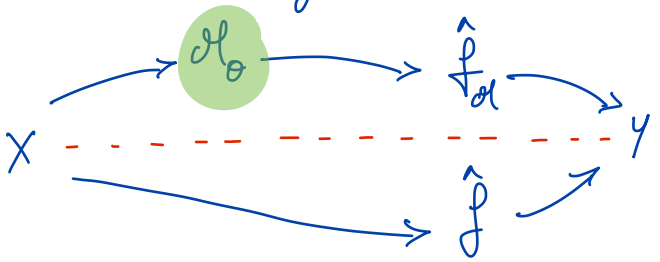
In supervised learning, MB techniques are usually the old ones, not necessarily the most performant but usually quite ok with a computational cost which is reasonnable.

In unsupervised learning, the situation is very diff. and MB techniques are the most performant for the moment.

Generative vs. discriminative techniques

The terms "generative" and "discriminative" techniques are used in supervised learning to discriminate model-based and non-model-based techniques.

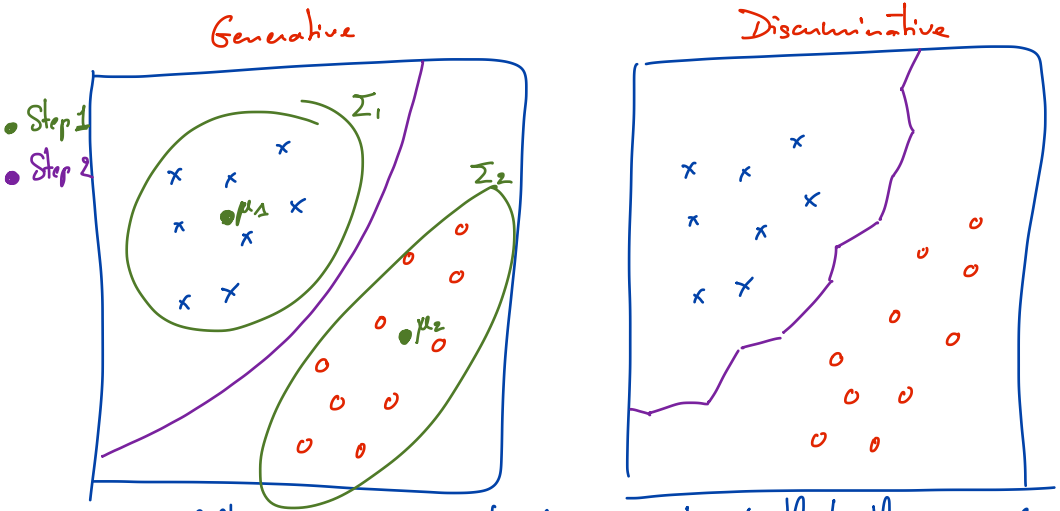
The two families are discriminated by the way they build the classification rule:



Generative

Discriminative

Generative vs. discriminative techniques



Rmq: On the left, it is important to understand that the purple boundary has been build only from the (green) model and it is not directly related with the learning data.

Why MB learning is interesting?

- First, MB learning is the best solution in unsupervised learning.
- In sup. learning, MB techniques may also be interesting:
  - MB techniques are easier to understand and to interpret the results
  - MB tech. output probabilities, allowing to have a measure of the prediction risk.
  - MB tech may avoid CV for parameter tuning.

Why MB learning could be difficult?

- MB techniques usually behave badly with high-dimensional data (p large)

- another difficult situations is the case of small data (n small, p moderate or large).

- dealing with different types of data is not easy for MB techniques.

Model-based techniques for supervised learning.

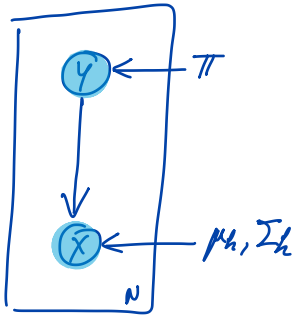
Among the numerous MB techniques for classification, one of the most simple is Quadratic discriminant analysis (QDA):

QDA assumes that each class has a pdf which is Gaussian with a specific mean μ_k and a specific covariance matrix Σ_k

$$\begin{cases} Y \sim \mathcal{M}(\pi) \quad \text{with } \pi = (\pi_1, \dots, \pi_K) \\ X \mid Y = k \sim \mathcal{N}(x; \mu_k, \Sigma_k) \quad \forall k = 1 \dots K \end{cases}$$

QDA

The graphical model of QDA:



We can see here that this model is quite simple and that the parameters to estimate from the data are $\theta = \{\pi_1, \dots, \pi_K, \mu_1 \dots \mu_K, \Sigma_1, \dots, \Sigma_K\}$.

The learning process of QDA is two parts:

1) The first step of the learning process is to estimate the model (=> estimate the parameters from the data. => we will use the maximum likelihood technique

$$(X, Y) \xrightarrow{ML} \hat{\theta}$$

2) Find a classification rule from this model. The MAP (maximum a posteriori) rule is very popular in this case. It is known to minimize the classification risk.

The MAP rule is as follows:

$$\hat{y}^* = \underset{k=1\dots K}{\operatorname{argmax}} \; P(Y=k \mid X=x^*)$$

$$P(Y=k \mid X=x^*) \overset{\text{Bayes}}{=} \frac{P(Y=k)\,P(X=x^* \mid Y=k)}{p(X)} \propto P(Y=k)\,P(X=x^* \mid Y=k)$$

$$\propto \hat{\pi}_k \cdot \mathcal{N}(x^*; \hat{\mu}_k, \hat{\Sigma}_k)$$

$$\hat{y}^* = \underset{k=1\dots K}{\operatorname{argmax}} \; \hat{\pi}_k \times \mathcal{N}(x^*; \hat{\mu}_k, \hat{\Sigma}_k)$$

$$= \underset{k=1\dots K}{\operatorname{argmax}} \; \log(\hat{\pi}_k) + \log\left(\mathcal{N}(x^*; \hat{\mu}_k, \hat{\Sigma}_k)\right)$$

In practice, on some (fake) data:

1st step: estimate

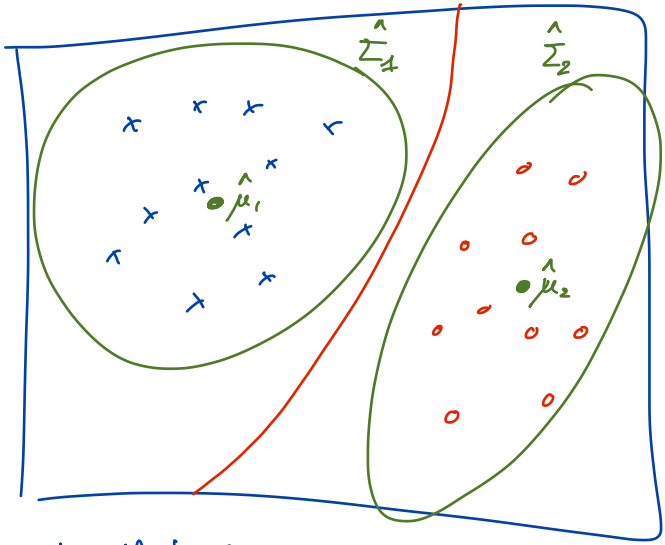
$\hat{\pi}_1 = 0.45$

$\hat{\pi}_2 = 0.55$

$\mu_1, \mu_2, \Sigma_1, \Sigma_2$

2nd step:

Compute for each possible x^* the MAP rule



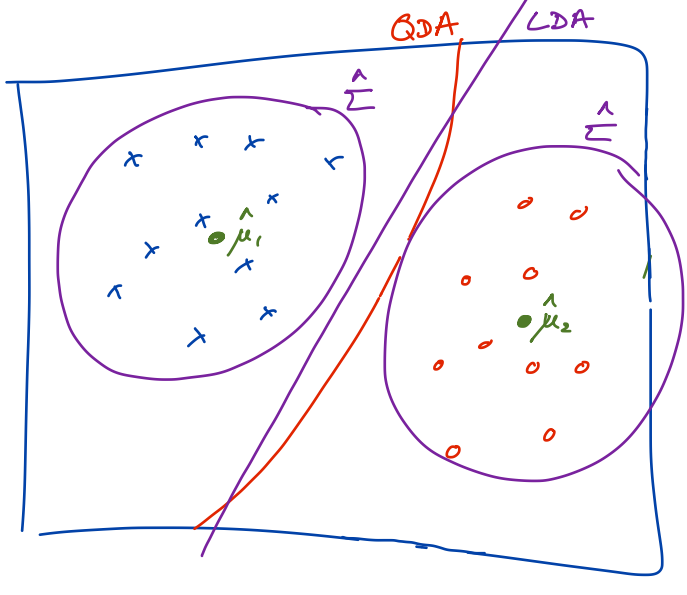
Exercice: demonstrate that the classification rule (MAP) is quadratic according to x^*

In the line of QDA, there is also the famous Linear Discriminant Analysis (LDA) proposed by Fisher in 1936.

The model of LDA is as follows:

$$\begin{cases} Y \sim \mathcal{M}(\pi) \text{ with } \pi = \pi_1, \dots, \pi_K \\ X \mid Y = k \sim \mathcal{N}(x; \mu_k, \Sigma) \quad \forall k = 1 \dots K. \end{cases}$$

Rmk: the model of LDA is exactly the same as QDA except that $\Sigma_k = \Sigma \ \forall k$.


QDA
LDA
$\hat{\Sigma}$
$\hat{\Sigma}$
$\hat{\mu}_1$
$\hat{\mu}_2$

Summary on LDA / QDA:

- Both methods assume a generative model based on the Gaussian distribution
- LDA is assuming also that $\Sigma_k = \Sigma \ \forall k$.
- QDA is unfortunately not performing well when p is moderate or large.
- LDA is clearly outperforming QDA in most situations and is part of the reference methods in classification.