

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](https://twitter.com/cbouveyron)

What is model-based statistic learning?

TBSL techniques are a subset of the statistical learning methods that assumes a statistical model that is supposed to have generated the data.

TBSL techniques have all a data generation model!

Ex: the Linear model

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

What is model-based statistic learning?

Among those M.B statistical Learning techniques , we can cite:

- the linear model for regression
- Logistic regression
- LDA / QDA
- probabilistic PCA
- t-SNE
- ...

Its place in the statistical / machine learning field

MB techniques represent a large part of modern machine learning, in particular because the most popular methods are now quite old.

Stat learn	
1936	Fisher's LDA
1960	logistic regression
1965	k-means
1979	EM algorithm for clustering
:	

Machine learning	
1980	Neural Networks
1990	SVR
:	

Generative vs. discriminative techniques

We can encounter in the literature the terms "discriminative" and "generative" techniques for classification.

- Generative = there is a model that is supposed to have generated the data
- discriminative : the method aims to directly model the classification boundaries.

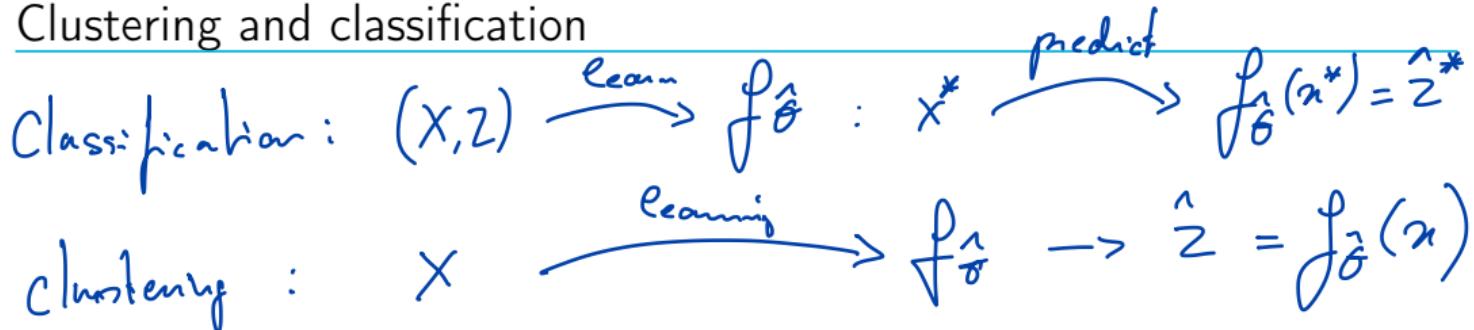
Why MB learning is interesting?

- the knowledge of the statistical model is usually comfortable for the analyst to interpret the studied phenomenon.
- the MB methods usually provide as output the knowledge of the prediction risk.
- the use of a MB method offers the possibility to rely on model selection to select the most appropriate model for the data.
- most of MB techniques can be extended to adapt to the complexity of the problem at hand.

Outline

1. Introduction
 2. Reminder on the learning process
 3. Model-based statistical learning
 4. Linear models for classification
 5. Mixture models and the EM algorithm *for clustering.*
- (...)

Clustering and classification



The task of clustering consists in forming groups only from the data X .

Definition: the goal of clustering is to form K homogeneous groups of data such that:

- the data of a group should be similar
- the data of different groups should be different.

Agglomerative hierarchical clustering

Even though the clustering task is simple, the general problem is combinatorial and it is possible to test in a reasonable time all possible configurations for $n \geq 10$ and $K \geq 2$.

↳ so, this is why we need algorithms to approximate the best configuration.

→ k-means

→ AdaBoost

- hierarchical
- spectral clustering

} not
NB
techniques

⇒ EM algo for mixture models.

The mixture model

The mixture model is a class of statistical model which assumes the following distribution:

$$p(x) = \sum_{h=1}^K \pi_h p_h(x)$$

where p_h is a certain probability density function and $\pi_h \in [0,1]$ and is such that $\sum_{h=1}^K \pi_h = 1$.

Then p is also a probability density function, a mixture distribution.

The mixture model

For instance, the following distributions belong to the mixture family:

$$p(x) = 0.4 N(x; 1, 1) + 0.6 N(x; 2, 1)$$

$$p(x) = 0.1 N(x; 1, 1) + 0.7 U_{[0,1]}(x) + 0.2 \text{Gamma}(x; 1, 1)$$

$$p(x) = 0.4 N(x; 1, 1) + 0.6 T(x; 3)$$

$$p(x) = \sum_{k=1}^{10} \pi_k p(x; k)$$