Bases de l'Apprentissage statistique

Examen - Contrôle de Connaissances (duration 1h30)

Les notes de cours ne sont pas autorisées, l'usage d'ordinateurs ou tablettes est prohibé. Les réponses doivent être précises et concises. No document is authorized. No computers no laptop. The answers must be precise and short.

1 - Introduction to supervised classification

Notations. We consider the probabilistic and statistical framework of supervised classification where X is a random vector on $\mathbb{R}^d$, $d \geq 1$ and Y is a binary random variable with values in $\{-1,\ +1\}$. A random sample $\mathcal{S}_n = \{(X_1,\ Y_1),\ \ldots,\ (X_n, Y_n)\}$, with n independent copies of the pair (X, Y) of joint probability distribution P. $\langle .,\ . \rangle$ (resp. $||.||$) denotes the inner product and the euclidean norm on $\mathbb{R}^d$.

1. What is a binary classifier ?
2. Give a definition of the theoretical problem of supervised binary classification relying upon the definition of risk.
3. Define the best classifier in the case of the prediction loss (the so-called 0-1 loss)
4. Define the empirical risk of a classifier calculated using $\mathcal{S}_n$. Explain the principe of *Empirical risk minimization.*
5. Briefly justify when the minimization of empirical risk is relevant.
6. Conversely, why the empirical risk minimization may rise issues ? Which approach do you propose to address this issue ?

2-Supervised learning Algorithms

2.a Support Vector Machines

We consider the framework of binary supervised classification.

1. Give the definition of a positive definite kernel and explain its key property used in SVM to deal with data non linearly separable.
2. How would you solve a multi-class classification problem with this method ?

2.b Trees and Ensemble methods

1. Explain why it might relevant to combine many decision trees

2. Explain the difference between Bagging and Random Forest

3 - Unsupervised Learning

Let X be a random vector on $\mathbb{R}^d$ of probability law $P(.)$ and of density probability $p(\cdot)$ and $\mathcal{S}_n = \{X_1, X_2, \ldots, X_n\}$ a random sample of size n drawn from this probability law. Let K be the number of clusters to be determined.

(a) Define the K-means algorithm

(b) Define the goal of Gaussian Mixture Modeling

(c) Give the principle of the Expectation-Maximization algorithm without entering into the computing details

(d) What is K-means compared to GMM ?

4 - Bonus question (outside the lectures)

In a problem called Novelty Detection, a method has been proposed to model a set of one-class data in a hypersphere in some feature space induced by a kernel. The algorithm defines a spherical boundary around the training data except very far data. The volume of this hypersphere is minimized, in order to limit the effect of incorporating outliers in the solution.

(a) Write this problem as a primal optimization one where the parameters to find are the center of the hypersphere, its radius and slack variables

(b) Bonus++ : write the dual problem by using Lagrangian relaxation.