

# Data & Apprentissage

—  
13 Novembre 2023

—  
EXAMEN FINAL

—  
Le sujet est imprimé sur 2 pages – Durée 2 heures

—  
*Les documents et notes de cours sont autorisées. L'utilisation de smartphones et accès à des ressources documentaires sur internet est interdit. L'examen est à réaliser individuellement et toute tentative de collaboration entre étudiants sera sévèrement sanctionnée.*

## Exercice 1. Understanding of main concepts of the course [10 points]

Answer the following question in 5 to 10 lines.

1. What is the so-called "bias-variance" trade-off in Machine Learning?
2. Define a theoretical criterion to assess the performance of a scoring algorithm (such as logistic regression for instance).
3. How to estimate a sparse linear model?
4. What are the hyperparameters of decision tree algorithms?
5. What are the two main ingredients of ensemble methods?
6. What is the mathematical space of functions for deep feedforward neural networks and what are the hyperparameters of such algorithms?
7. What is the role of penalization in machine learning?
8. How to achieve numerical optimization in machine learning methods? Give at least one specific example.

## Exercice 2. Clustering [3 points]

- a) Give the  $k$ -means objective function, the pseudocode of the clustering algorithm, explain with simple arguments : why the procedure converges and why multiple (re)runs are suggested.
- b) We are interested about a soft  $k$ -means that would use a probabilistic datapoint-to-cluster assignment. Propose a way to define this probabilistic assignment, give an updated pseudocode for this case, and argue whether or not this would converge.
- c) Is the clustering solution found by the soft  $k$ -means a local minima for the typical  $k$ -means? In other words, if we initialize  $k$ -means with the solution of soft  $k$ -means, would the clusters change or not and why?
- d) Give additional elements of comparison for the two aforementioned hard and soft  $k$ -means approaches? Which are the cases where the latter is better to be used?

**Exercice 3. Graphical models**

[4 points]

A security system relies on a battery, a sensor, and an alarm. Consider the three respective binary random variables  $a$ ,  $b$ ,  $c$ . When the alarm fires :  $a = 1$  ; when the battery functions properly :  $b = 1$  ; and when the sensor functions properly :  $c = 1$ . Otherwise, those variables take a value equal to 0. The system architecture is very simple : the battery powers the sensor, and the sensor feeds information to the controller of the alarm that decides to fire to signify a threat. The system manual informs us that :

- the probability for the battery to fail is 0.1
  - the probability for the sensor to fail is 0.01 when the battery is ok, and 0.3 when the battery is not ok.
  - the probability for the alarm to be silent is 0.99, when the sensor works fine.
- a) Give the graphical model that involves the variables of this problem, and translate all the known information given by the manual to proper writing with probabilities (it does not involve calculations).
  - b) Give the  $2 \times 2$  table of the probability  $p(a = 1|b, c)$ . Examine if  $a$  and  $b$  are independent.
  - c) Compute the probability for the alarm to be silent.
  - d) Suppose we don't hear any alarm and we have no information about whether the sensor functions correctly. What is the probability of the battery to be disfunctioning ?

**Exercice 4. Regression and robust estimation**

[3 points]

- a) Let  $P$  be a set of  $N$  2-dimensional points,  $p_i = (x_i, y_i)$ . Assume that we want to find the line that best fits the points of  $P$ . Present in detail how the Ordinary Least Squares (OLS) works when fitting a linear model by estimating its parameters.
- b) Considering the data of Table 1, apply the OLS by making the numerical calculations and report the fitted model and its Root Mean Squared Error (RMSE).

$i$	1	2	3	4
$x_i$	2.2	0.9	0.4	-0.6
$y_i$	1.2	2	1	-1

TABLE 1 – Practical example : data points  $P = \{(x_i, y_i), i = 1 \dots 4\}$ .

- c) Describe in detail the standard RANSAC method for linear regression. Explain how each of the several linear models produced internally can 'grade' datapoints for being outliers, and conversely how the datapoints can 'grade' each such linear model for being a good regressor.