Data & Apprentissage

—

Mars 2023

—

Examen final

—

Le sujet est imprimé sur 3 pages – Durée 2 heures

—

LLes 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 sévérement sanctionnée.

Exercice 1. Restitution du cours *[10 points]*

Répondre aux questions suivantes en maximum 5-10 lignes par question.

1. *Quels sont les liens entre machine learning et simulation numérique ?*
2. *Préciser la différence entre* données *de classification et* problème *de classification.*
3. *Discuter les choix possibles des critères d'évaluation d'une méthode d'apprentissage dans le cas de données de classification.*
4. *Citer les trois grandes familles d'algorithmes d'apprentissage supervisé en donnant deux exemples d'algorithmes par famille.*
5. *En quoi consiste le compromis "biais-variance" en apprentissage ?*
6. *Décrire les principales caractéristiques d'un algorithme d'apprentissage supervisé de votre choix. Préciser les paramètres choisis par le data scientist et les paramètres "appris" par l'algorithme.*
7. *Quels avantages présentent les arbres de décision sur la majorité des autres algorithmes ?*
8. *Quel schéma d'optimisation générique est utilisé dans la calibration des architectures de type Deep Learning ?*
9. *Peut-on résoudre le problème de la calibration d'hyperparamètres par une technique d'optimisation ? Fournir une réponse argumentée en précisant la nature du problème d'un point de vue numérique, ses caractéristiques générales et le type de solution qui pourrait être envisagée.*
10. *Mentionner deux problématiques concrètes de mise en oeuvre dans les applications réelles. On prendra l'exemple de la détection d'anomalies (supervisée) sur une chaîne de production industrielle instrumentée par un réseaux de capteurs mesurant des grandeurs physiques en continu (qualité des flux entrants et des flux sortants, vibrations, etc.).*

Exercice 2. Regression and robust estimation *[4 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	-0.5	0.3	0.7	1.5
y_i	1.2	2.0	1.0	-1.0

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

c) *Describe in detail the standard RANSAC method and algorithm for linear regression. Give insights about the sensitivity to its parameters. What is the difference to a typical bootstrapping estimation procedure ?*

d) *Imagine that the data acquisition of the N data instances, $X = \{x_1, ..., x_N\}$, has a variable quality q_i for each instance x_i. A domain expert, who is responsible for the data, provides you a data matrix $D \in \mathbb{R}^{N\times(d+1)}$ where the last column contains non-negative real values that correspond to the quality indicators, $q_i \geq 0$, $\forall i = 1, ..., N$).*

Investigate the statistical properties of X by applying RANSAC for the robust estimation of $\delta = ||\mu - m||_2$, where μ is the center of the instances in the set, and m is the medoid instance of the set, which has the shortest average distance to the other objects.

Exercice 3. Clustering *[4 points]*

a) *Give a pseudocode for the k-means clustering algorithm and explain with simple arguments : why the procedure converges, and why multiple (re)runs are suggested.*

b) *Suppose we initialize the clusters by assigning each of the N data instances to one of the k clusters uniformly at random. How does this compare qualitatively to the typical initialization strategy ? Will it converge ; faster or slower ?*

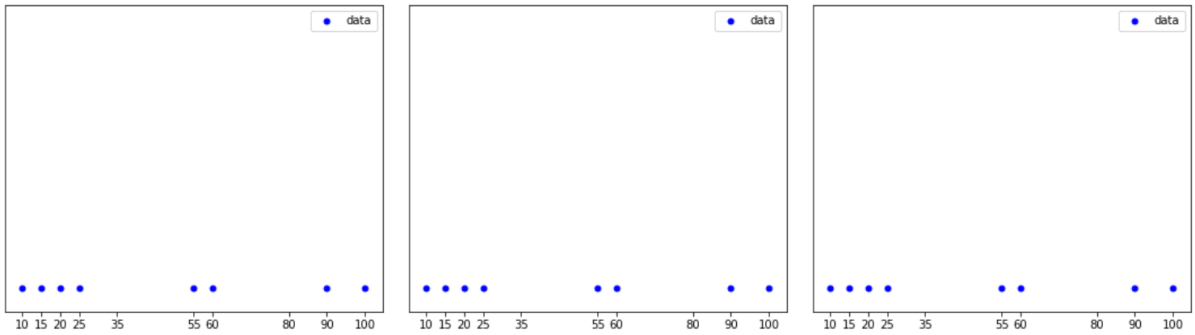
c) *Consider the 1D dataset $X = \{1, 1.5, 2, 2.5, 5.4, 6, 9, 10\}$. Simulate all the steps of a 2-means solution using the k-farthest initialization, and fill Table 2 below.*

d) *Given the clusters found in (c), summarize the dataset by providing two 'typical' data examples. Explain your answer and comment on how well each of them serves its role.*

e) *Give three or more cluster linkage criteria that are used in hierarchical agglomerative clustering, and provide the formulas to compute them. In the following figures, draw the cluster hierarchy that would be computed by each of these criteria.*

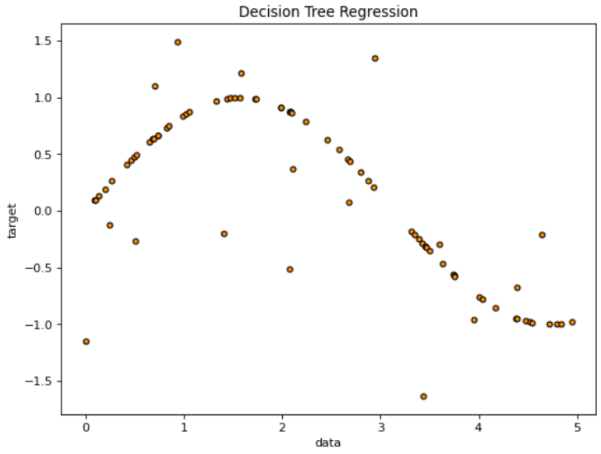
step	centers	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	MSE
1										
2										
3										
4										

Table 2 – k-means iterations. Compute the MSE only for the first and the final iterations.



Exercice 4. Decision Trees *[2 points]*

a) *Draw directly on the figure below, the way (roughly) a regression tree (RT) would approximate the data of the time-series, for depth* $= 1$ *and depth* $= 2$.



b) *Draw on the right of the figure, the RT for depth* $= 2$.

c) *You trained your RT to approximate the function* $f(\cdot)$ *using the above dataset. Now, other clients ask you to directly predict for their data, because they believe they are of the same nature but a bit transformed. Specifically, let* $f_{\#}(\cdot)$ *be clients' functions :*
client c1) $f_1(x) = 3f(x)$ *;*
client c2) $f_2(3x) = f(x)$ *;*
client c3) $f_3(x) = [f_1(x) + f_2(3x)]^2$.
What would you answer ? Is it feasible ? Which jobs would you accept and how would you manage in each of the cases ?