Data & Apprentissage

—

8 novembre 2021

—

Examen

—

Le sujet est imprimé sur 4 pages – Durée 2 heures

—

Aucun document n'est autorisé. L'utilisation de smartphones et l'accès à des ressources documentaires sur internet sont rigoureusement interdits. Toute collaboration entre étudiants pendant l'épreuve sera sévérement sanctionnée.

Exercice 1. Restitution du cours

Répondre aux questions suivantes en 5 lignes maximum par question.

1. *Expliquer la différence entre* données *de classification et* problème *de classification.*
2. *Donner la définition de la courbe ROC. Quel est son intérêt ?*
3. *Décrire le modèle de régression logistique et expliquer comment l'utiliser pour un problème de classification ? Pour quel autre problème d'apprentissage ce modèle peut-il être utilisé ?*
4. *Décrire le principe des méthodes d'apprentissage dites locales. Donner un exemple en indiquant les hyperparamètres de la méthode.*
5. *Décrire le principe des méthodes d'apprentissage dites globales. Pourquoi utilise-t-on parfois des pertes convexes en apprentissage ? Quels autres choix de pertes conduisent à des méthodes pratiques ?*
6. *Quels sont les termes d'erreur permettant de rendre compte du phénomène de surapprentissage (overfitting) ? Discuter l'impact du temps de calcul (à ressources fixées) quand la taille d'échantillon augmente ?*
7. *Evoquer les enjeux et la problématique sous-jacente à la calibration des hyperparamètres d'une méthode d'apprentissage.*
8. *Donner deux exemples d'applications du machine learning en lien avec des sujets de simulation numérique.*

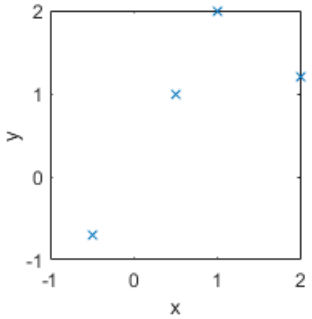
Exercice 2. Regression

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 2, apply the OLS by making the numerical calculations and report the fitted model and its Root Mean Squared Error (RMSE). c) You are asked

i	1	2	3	4
x_i	2.0	1.0	0.5	-0.5
y_i	1.2	2	1	-0.7

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

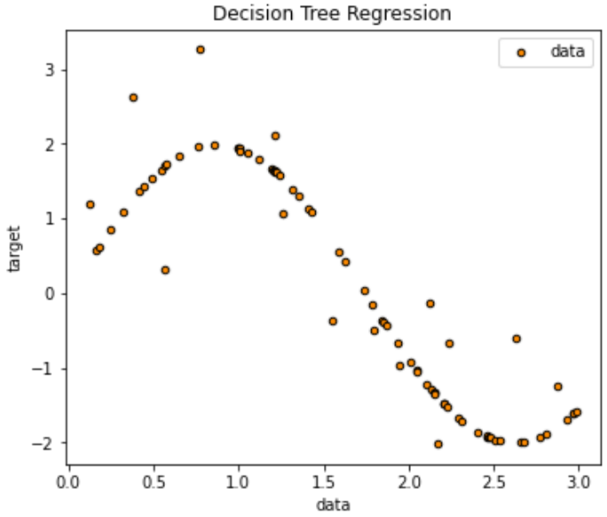


to try more complex polynomial models for this dataset. Which ones would you check, and which one would be the most complex that is meaningful to try ?

d) Explain the principle of Ridge Regression (RR).

e) For a linear regression problem, you are given N=50 datapoints, each one described by d=200 features. Which regression method would you prefer ? If you were asked to use specifically OLS, describe one way or more in which you would proceed.

Exercice 3. Decision Trees



a) Draw directly on the above figure the way (roughly) a regression tree (RT) approximates the data of the time-series, for depth = 1 and depth = 2.

b) Draw on the right to the figure, a (rough) RT with depth = 2.

c) How would an ensemble of RTs (each tree using random subsets of the data to train with) approximate the region $x \in [1, 2.5]$ of this dataset ? Make a drawing and explain.

d) You have trained your $RT_{initial}$ to approximate the unknown function $f(\cdot)$ using the dataset of the figure. Then, your intention is to use it also for unseen data of the same nature, which however are a bit transformed. Specifically, let $f_1(\cdot)$, $f_2(\cdot)$ and $f_3(\cdot)$ be the new functions that can be expressed with repect to $f(\cdot)$:

case d1) $f_1(5x) = f(x)$

case d2) $f_2(x) = 5f(x)$

case d3) $f_3(x) = (f(x))^5$

Do you have to retrain from scratch your RT to apply it in these cases ? If yes, why ; if no, what is the alernative approach you would follow for (d1)-(d3) ?

e) What would be the difference in (d) if the $RT_{initial}$ is a single tree or an ensemble model of several trees ?

Exercice 4. Clustering

a) Consider a special case of the 2-means algorithm : a k-means with $k = 2$ where, after the centers' initialization, only one of them (defined by the user) gets updated while the second one remains always fixed. Give a pseudocode of this particular algorithm.

b) You have the dataset $X = \{10, 15, 20, 25, 55, 60, 90, 100\}$. Starting with centers at positions 35 and 80 (fixed), compute numerically all the steps of the algorihm of (a) until convergence to solution $\hat{c}$. For each step, fill the following table with the positions of the centers, the cluster to which datapoints is assigned, and the Mean Squared Error (MSE).

		assignment to cluster								
step	centers	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	MSE
1										
2										
3										
4										
5										
6										

c) Given the centers of $\hat{c}$, summarize the dataset by providing two 'typical' data examples. Compare how well each cluster is represented by its center, and why.

d) Given the centers of $\hat{c}$, now apply the standard 2-means (no restriction to center updates) and fill in the table below.

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

e) During k-means, explain how is it possible that one cluster becomes empty ? In such case, describe a couple of reasonable heuristics that would allow the process to continue without reducing the cluser number k.