Data & Apprentissage

—

Novembre 2024

—

Examen final

—

Le sujet est imprimé sur 3 pages – Durée 2 heures

—

Les documents et notes de cours ne sont pas autorisés. 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. General course understanding *[4 points]*

Please provide concise replies to each question.

1. *Quote one example of algorithm in each of the three main categories of machine learning methods (local, global, ensemble).*
2. *Explain what is the trade-off between estimation error and approximation error in machine learning and how can this be solved ?*
3. *What is the optimization strategy in deep learning ?*
4. *Give an example of hyperparameter in a specific machine learning method and explain how it can be calibrated.*

Exercice 2. Course understanding - Decision trees *[5 points]*

1. *What is a decision tree in the context of binary classification ?*
2. *Consider a new point x, what is the label y assigned by the decision tree ?*
3. *Consider a set of supervised classification data. Provide the pseudocode in order to train a decision tree classifier.*
4. *Give at least two stopping rules of the training procedure of decision trees.*
5. *Give at least two hyperparameters of decision trees.*
6. *Is it possible to monitor the number of cells of the tree using the training data ?*
7. *What are the pros and cons of decision trees ?*
8. *Is it possible to solve the drawbacks of decision trees ?*

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 controler 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* 2×2 *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 ?*

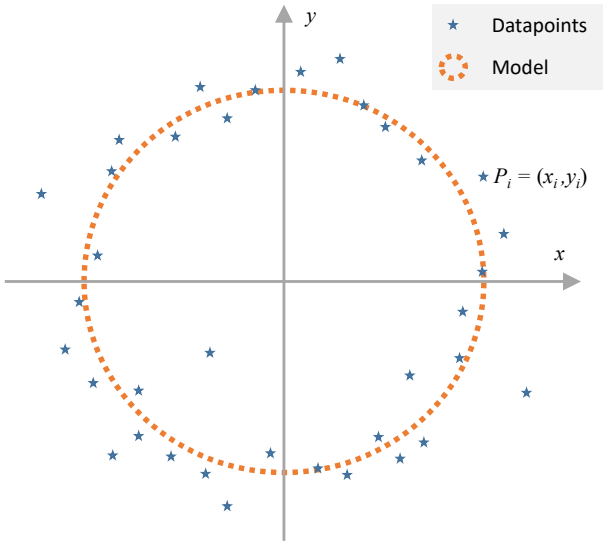
Exercice 4. Regression and robust estimation *[3 points]*

a) *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 $P = \{(x_i, y_i), i = 1...4\}$.

b) *In a given dataset* P*, the datapoints are assumed to be distributed in a circular shape (e.g. see the figure below). We are interested to learn a circular model* $C(c, r)$ *that best fits the data. Recall that the standard equation of a 2D circle centered at* $c = (x_c, y_c)$ *and of radius* r *is :* $(x - x_c)^2 + (y - y_c)^2 = r^2$*. Develop in detail how this problem could be solved with the generalized RANSAC approach.*



Exercice 5. 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 how it can be used to approximate the globally optimal solution containing the best possible centroids.*

b) *We are interested about a* soft *k-*means *that would use a probabilistic datapoint-to-cluster assignment (similar to Gaussian Mixture Modeling). Propose a way to define this probabilistic assignment, give an updated pseudocode for this case, and argue whether or not this would converge.*

c) *If we initialize k-means with the solution of soft k-means, would the clusters change or not, and why ? Give 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 ?*

d) *Describe a procedure that would use a bootstrap approach, i.e. multiple randomized k-means solutions, to compute an outlier score for each datapoint. Discuss its properties in perspective with Isolation Forest.*