

Numéro d'anonymat :

REPONDEZ SUR LE SUJET.

Exercice 1. (Questions de cours)

1. Quels sont les trois objectifs principaux d'un modèle de régression ? Décrivez brièvement chacun d'eux.

Réponse :

2. Donnez deux mesures d'évaluation de la qualité d'un modèle de régression. Discutez brièvement de leur interprétation.

Réponse :

3. Un des grands avantages des arbres de décision est la possibilité d'interpréter leurs décisions. Expliquez comment ?

Réponse :

4. Listez aux moins deux hyperparamètres importants des forêts aléatoires et expliquez brièvement leur rôle.

Réponse :

5. Quels sont les principaux avantages et inconvénients des classifieurs linéaires comme le Perceptron ou les SVM linéaires ? Expliquez comment un SVM avec noyaux permet de traiter des données non linéairement séparables.

Réponse :

6. On dispose d'une observation de 120 éléments, répartis à raison de deux tiers pour la classe A et le reste pour la classe B. On a 5 chances sur 8 de prédire correctement un A et 4 chances sur 5 de prédire correctement un B. Écrire la matrice de confusion.

Réponse :

7. Quelles sont les deux méthodes de réduction de dimension vues en cours ? Expliquez brièvement les principales différences/avantages/inconvénients de chacune elles.

Réponse :

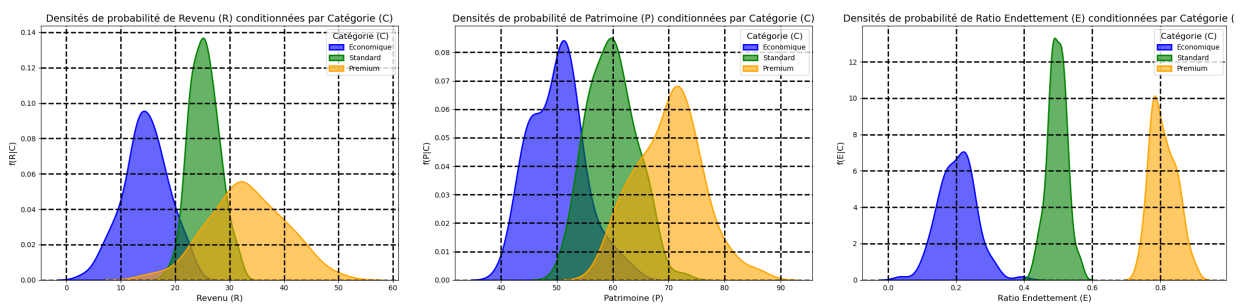
Exercice 2. (Classifieurs de Bayes)

On dispose d'un corpus de données $(R, P, E; C)$ relatives aux clients d'une entreprise de vente en ligne. Les données sont les suivantes :

- Revenu (R) : le revenu annuel du client en milliers d'euros
- Patrimoine (P) : le patrimoine du client en dizaine de milliers d'euros
- Ratio d'endettement (E) : le ratio de l'endettement du client par rapport à son revenu.
- Catégorie (C) : la catégorie du client, qui peut être Économique (1), Standard (2) ou Premium (3).

On souhaite entraîner un classifieur bayésien à prédire la catégorie c d'un client décrit par un tuple (r, p, e) .

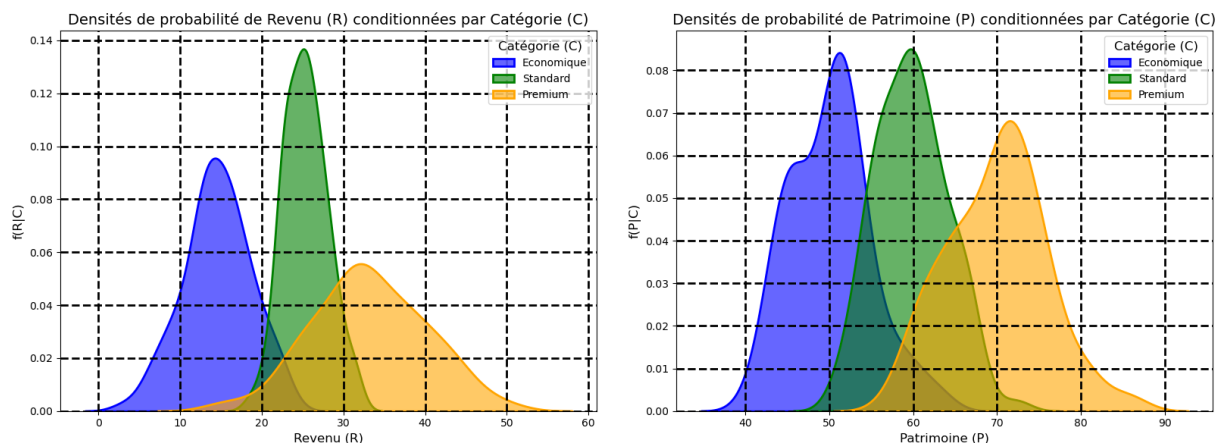
On dessine alors les trois densités de probabilité $f(R | C = c)$, $f(P | C = c)$ et $f(E | C = c)$ pour chaque catégorie c .



1. Si on ne devait garder qu'un seul descripteur, lequel choisiriez-vous ?

Réponse :

2. On choisit de garder le revenu et le patrimoine comme descripteurs.



- (a) Quelle est la moyenne des patrimoines des clients de chaque catégorie ?

Réponse :

- (b) Quelle serait la catégorie prédite pour un client ayant un revenu de 20k€ et un patrimoine de 600k€ ? (dans vos calculs, vous pouvez arrondir à 2 chiffres après la virgule).

Réponse :

3. Que deviendrait cette prédiction si on sait que dans l'échantillon, on a 60% de clients de catégorie Économique, 30% de clients de catégorie Standard et 10% de clients de catégorie Premium ?

Réponse :

Exercice 3. (Arbres de décision)

On dispose d'un corpus de données concernant des clients d'une entreprise, avec des informations sur leur âge, leur revenu et si ces clients ont acheté un produit. La table suivante présente ces informations :

Âge	Revenu	Achat (O/N)
25 – 35	Faible	Non
25 – 35	Faible	Oui
36 – 45	Élevé	Oui
36 – 45	Faible	Non
46 – 55	Élevé	Oui
46 – 55	Faible	Non
25 – 35	Élevé	Oui
56 – 65	Élevé	Oui

Objectif : Construire un arbre de décision pour prédire si un client va acheter le produit en fonction de son âge et de son revenu.

Indication : Le tableau suivant donne quelques valeurs pour le calcul de l'information $I(p, n)$.

p	1	1	1	2	2	3	3
n	2	3	4	3	4	4	5
$I(p, n)$	2.75	3.24	3.60	4.85	5.50	6.9	7.63

1. Calculer l'information totale du corpus de données, où les classes sont "Oui" et "Non".

Réponse :

2. Partitionnez les données selon l'attribut "Âge" et calculez l'entropie pour cet attribut.

Réponse :

3. Partitionnez les données selon l'attribut "Revenu" et calculez l'entropie pour cet attribut.

Réponse :

4. Calculez le gain d'information pour chaque attribut.

Réponse :

5. Construisez l'arbre de décision pour ce problème de classification. Qu'observez-vous sur l'une des feuilles obtenues ?

Réponse :

Exercice 4. (Réduction de dimension)

Une étude environnementale est menée pour analyser les caractéristiques écologiques de différentes régions. Les données collectées incluent 5 indicateurs environnementaux mesurés dans différentes régions.

Les indicateurs environnementaux sont les suivants :

- Concentration de CO₂ : Mesure des émissions de dioxyde de carbone dans l'atmosphère.
- Température moyenne annuelle : Moyenne annuelle des températures dans la région.
- Précipitations annuelles : Quantité totale des précipitations sur une année.
- Indice de pollution atmosphérique : Indice IPA, qui mesure la qualité de l'air.
- Taux de couverture forestière : Proportion de la surface régionale occupée par des forêts.

Afin de permettre une visualisation des données, une analyse en composantes principales (ACP) a été réalisée.

1. Donnez les différentes étapes de l'ACP.

Réponse :

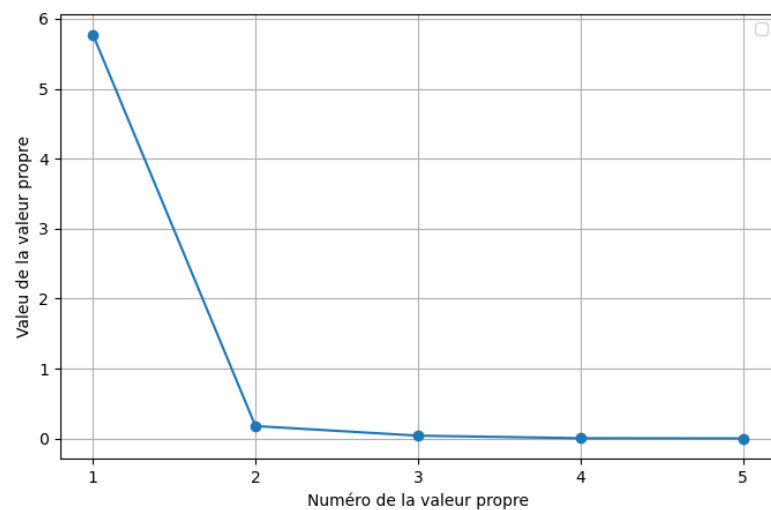
Les première étapes de l'ACP ont permis d'établir la liste suivante des valeurs propres :

Valeurs propres : [5.776, 0.178, 0.040, 0.004, 0.002]

2. Quelle est la proportion de variance expliquée par les trois premières composantes principales ?

Réponse :

On dessine l'éboulis des valeurs propres pour les 5 composantes principales. On obtient le graphique suivant :



3. Quelle est la dimension optimale pour la réduction de dimension ? Quel est le critère utilisé pour déterminer cette dimension ?

Réponse :

FIN.

