Interrogation 2022-2023

Exercice 1 : QCM (5 points)

Cocher la (les) bonne (s) réponse (s) :

1) Quelle est la fonction d'erreur la plus couramment utilisée en régression linéaire?

a) La fonction de perte de Hinge
b) La fonction de perte logistique
c) La fonction d'erreur quadratique moyenne
d) La fonction de perte de Perceptron

2) Qu'est-ce qu'un arbre de décision?

a) Un algorithme de classification linéaire
b) Un algorithme de régression linéaire
c) Un algorithme de classification non linéaire
d) Un algorithme de clustering

3) Comment est déterminée la meilleure variable de division dans un arbre de décision?

a) En maximisant l'information gain
b) En minimisant l'information gain
c) En maximisant le sur-apprentissage de l'arbre
d) En minimisant sur-apprentissage de l'arbre

4) Comment est déterminée la profondeur optimale d'un arbre de décision?

a) En maximisant l'information gain
b) En minimisant l'information gain
c) En maximisant sur-apprentissage de l'arbre
d) En minimisant sur-apprentissage de l'arbre

5) Qu'est-ce que la règle de Bayes?

a) Un théorème mathématique qui permet de calculer la probabilité conditionnelle
b) Une heuristique utilisée pour choisir la meilleure variable de division dans un arbre de décision
c) Une méthode d'optimisation utilisée pour minimiser la fonction de coût dans la régression linéaire
d) Une méthode d'optimisation utilisée pour minimiser l'impureté de Gini dans un arbre de décision

6) Quel est l'avantage de la régression linéaire par rapport à d'autres algorithmes de régression?

a) Elle est plus rapide à entraîner
b) Elle peut modéliser des relations non linéaires
c) Elle est plus précise que les autres algorithmes
d) Elle est plus facile à comprendre et à interpréter

7) Qu'est-ce que la descente de gradient?

a) Une méthode pour minimiser la fonction de coût dans la régression linéaire

b) Une méthode pour maximiser l'information gain dans les arbres de décision

c) Une méthode pour calculer la probabilité conditionnelle dans le classificateur de Bayes

d) Une méthode pour maximiser la précision de la classification linéaire

8) Qu'est-ce que l'entropie?

a) Une mesure de la pureté d'un nœud dans un arbre de décision

b) Une mesure de la distance entre les données dans l'espace des caractéristiques

c) Une mesure de l'erreur de prédiction pour un modèle de classification

d) Une mesure de la qualité d'un modèle de régression linéaire

9) Qu'est-ce que le classificateur de Bayes naïf?

a) Un algorithme de classification linéaire qui utilise une approche probabiliste

b) Un algorithme d'apprentissage non supervisé qui recherche des motifs dans les données

c) Un algorithme de classification qui utilise des règles conditionnelles simples et indépendantes

d) Un algorithme de régression linéaire qui utilise une approche bayésienne

10) Quelle est la différence entre l'overfitting et l'underfitting?

a) L'overfitting se produit lorsque le modèle est trop simple pour les données, tandis que l'underfitting se produit lorsque le modèle est trop complexe pour les données.

b) L'overfitting se produit lorsque le modèle s'adapte trop bien aux données d'entraînement, tandis que l'underfitting se produit lorsque le modèle ne s'adapte pas suffisamment bien aux données d'entraînement.

c) L'overfitting se produit lorsque le modèle a une variance élevée, tandis que l'underfitting se produit lorsque le modèle a une erreur de biais élevée.

d) L'overfitting et l'underfitting sont des termes interchangeables qui décrivent tous deux une mauvaise adaptation du modèle aux données.

Exercice 2 : Classificateur de Bayes (5 points)

Nous souhaitons réaliser un classifieur bayesien permettant de classifier les emails en « Spam » ou « Ham (not spam) ». Pour ce faire, chaque mot w_i d'un e-mail, quel que soit l'endroit où il se trouve dans l'e-mail, est supposé avoir une probabilité $P(W= w_i|Y)$, où W prend des mots dans un dictionnaire prédéterminé (la ponctuation est ignorée). Y prend une valeur binaire (Spam ou ham).

I. Supposons que nous avons trois emails comme ensemble d'apprentissage.

(Spam) dear sir, if you could answer my questions I would be most grateful.

(Ham) see you at 12

(Ham) well, prepare it for tomorrow.

A partir de cet ensemble d'entrainement, calculer les probabilités bayésiennes suivantes.

- $P(W=sir \mid Y = spam)$
- $P(W=see \mid Y = ham)$
- $P(W=today \mid Y = ham)$
- $P(Y = ham)$

II. Le tableau suivant montre les probabilités estimées d'un ensemble de mots spams entrainés sur un large corpus d'emails.

W	*good*	*to*	*fine*	*luck*	*pay*
P(W|Y=spam)	*1/6*	*1/8*	*1/4*	*1/8*	*1/4*
P(W|Y=ham)	*1/8*	*1/3*	*1/4*	*1/12*	*1/12*

On vous donne un nouvel email à classer, avec seulement deux mots :

Good luck

1. Calculer la décision estimée pour cet email, sachant que :
   P(Y = spam)=1/5.
2. Quelle est l'intervalle de probabilités de *P(Y = spam)* pour lequel le classifieur bayesien classe ce nouvel email comme spam ?

Exercice 3 Régression linéaire avec la bibliothèque Scikit-learn/Python (10 points)

Nous avons un ensemble de données comprenant deux variables, X et Y. Nous souhaitons créer un modèle de régression linéaire pour prédire la variable Y en fonction de la variable X.

Les données sont les suivantes :

X = [23, 45, 12, 67, 87, 43, 65, 34, 56, 78]

Y = [450, 678, 340, 980, 1200, 600, 900, 520, 800, 1100]

1. Tracer un graphique représentant les données.
2. Diviser les données en ensembles d'entraînement (80%) et de test (20%).
3. Créer un modèle de régression linéaire à l'aide de l'ensemble d'entraînement.
4. Prédire les valeurs de Y pour l'ensemble de test en utilisant le modèle de régression linéaire.
5. Évaluer les performances du modèle en utilisant la métrique de l'erreur quadratique moyenne (Mean Squared Error - MSE).
6. Tracer un graphique représentant les prédictions du modèle par rapport aux vraies valeurs pour l'ensemble de test.

Bon courage