Data mining 2
Examen de rattrapage

janvier-2023

Duréee : 2 heures. **Documents autorisés :** 2 feuilles recto-verso.

---

1 Questions de cours et de réflexion

Il vous est demandé de répondre de manière brève et précise aux questions suivantes :

1. Quel(s) avantage(s) présentent les forêts aléatoires (RF) par rapport aux arbres de décision.
2. Rappelez le principe du bootstrapping.
3. Quels sont les deux hyperparamètres principaux des RFs.
4. Que contient chaque noeud de chaque arbre d'une RF ?
5. Comment l'algorithme (appelé CART) de construction des arbres de décision formant une RF choisit-il ce contenu ?
6. Rappelez les deux erreurs calculées pour une RF.
7. Expliquez comment l'algorithme calcule l'importance des variables.

2 Généralités

Dans cette partie, nous considérons un problème de classification binaire dans lequel les deux classes sont C et $\neg C$ (exemple : *Malade* et *Sain*). Soit $CONF$ la matrice de confusion d'un classifieur binaire. $CONF$ peut être notée comme suit :

		Valeurs réelles	
		C	$\neg C$
Valeurs prédites	C	TP	FP
	$\neg C$	FN	TN
	$Total$	P	N

1. Soit un classifieur qui prédit systématiquement la classe C. Considérons un ensemble D d'exemples contenant 90 exemples de la classe C et 10 exemples de la classe $\neg C$. En utilisant D comme ensemble de test, donnez la matrice de confusion du classifieur et calculez son taux d'erreur ta correspondant à la précision (accuracy). Que constatez-vous ?
2. Proposez et calculez un autre taux d'erreur t_b permettant d'éviter les faiblesses du taux t_a.

Une autre façon d'évaluer un classifieur binaire est d'utiliser les deux mesures suivantes :

— La sensibilité : $S_e = \frac{TP}{P}$.
— La spécificité : $S_p = \frac{TN}{N}$.

3. Calculez la spécificité et la sensibilité pour les exemples suivants :
$M_1 = \begin{pmatrix} 9 & 8 \\ 1 & 1 \end{pmatrix}$ $M_2 = \begin{pmatrix} 1 & 1 \\ 9 & 8 \end{pmatrix}$ $M_3 = \begin{pmatrix} 9 & 1 \\ 1 & 8 \end{pmatrix}$.
4. Quelle signification donnez-vous à chacune de ces mesures ?
5. Expliquez pourquoi il est facile d'avoir "une bonne valeur" pour **l'une ou l'autre** de ces deux mesures, mais que cela ne suffit pas pour avoir un bon classifieur.

3 Classifieurs bayésiens

Nous supposons dans un premier temps que nos observations (exemples) sont décrites par des variables exclusivement catégorielles.

1. Rappelez en expliquant chacun de ses éléments la formule mathématique utilisée par les classifieurs bayésiens.
2. Quels sont les entrées et les sorties de l'algorithme permettant de construire un classifieur bayésien ?
3. Une fois qu'il est défini, quelle réponse un classifieur bayésien donne-t-il lorsqu'on lui soumet une nouvelle instance ?

4. Utilisez l'ensemble de données ci-dessous pour construire un classifieur bayésien permettant de prédire la valeur de la variable '*Emprunt accordé*'.
5. Appliquez ce classifieur à un exemple de votre choix.
6. Expliquez brièvement comment on teste la qualité d'un tel classifieur.
7. Dans certains cas, le résultat de cette validation n'est pas satisfaisant car certaines probabilités calculées par l'algorithme sont nulles. Comment expliquez-vous cela ? Comment remédier à ce problème ?

Num.	Âge	A un emploi	A un logement	Qualité dossier	Emprunt accordé
1	Jeune	Faux	Faux	Moyen	Non
2	Jeune	Faux	Faux	Bon	Non
3	Jeune	Vrai	Faux	Bon	Oui
4	Jeune	Vrai	Vrai	Moyen	Oui
5	Jeune	Faux	Faux	Moyen	Non
6	Intermédiaire	Faux	Faux	Moyen	Non
7	Intermédiaire	Faux	Faux	Bon	Non
8	Intermédiaire	Vrai	Vrai	Bon	Oui
9	Intermédiaire	Faux	Vrai	Excellent	Oui
10	Intermédiaire	Faux	Vrai	Excellent	Oui
11	Senior	Faux	Vrai	Excellent	Oui
12	Senior	Faux	Vrai	Bon	Oui
13	Senior	Vrai	Faux	Bon	Oui
14	Senior	Vrai	Faux	Excellent	Oui
15	Senior	Faux	Faux	Moyen	Non

Nous supposons à présent que les variables indépendantes de notre problème sont numériques.

8. Quelle autre formule mathématique utilisent les classifieurs bayésiens ?
9. Que suppose cette formule au sujet des variables indépendantes ?

4 Réseaux de neurones

On considère un réseau de neurones composé de 3 couches telles que la couche d'entrée contient n neurones, la couche cachée contient n neurones et la couche de sortie contient 1 neurone. La fonction d'activation des couches cachée et de sortie est la fonction de Heaviside. Les entrées ne prennent que les valeurs 0 et 1. Nous avons les valeurs suivantes des poids et des biais :

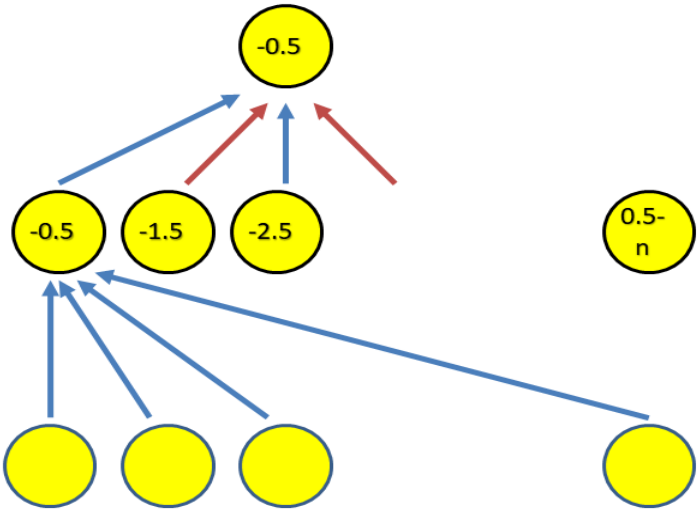

FIGURE 1 – Réseau de neurones

— Neurones de la couche cachée : tous les poids sont égaux à 1, et en parcourant les neurones de gauche à droite, les biais sont égaux à $-0.5, -1.5, -2.5, \ldots, 0.5 - n$ respectivement.
— Neurone de la couche de sortie : en parcourant les liens de gauche à droite les poids sont égaux à $1, -1, 1, -1, \ldots$ respectivement, et le biais est égal à -0.5.

La figure 1 récapitule ces caractéristiques (les flèches bleues correspondent à un poids 1 et les flèches rouge à un poids -1).

1. **Cas n=2 :**
   (a) Représentez graphiquement le réseau.
   (b) Calculez la sortie de chaque neurone des couches cachée et de sortie.
   (c) Quelle fonction calcule ce réseau ?
2. **Cas n=3 :**
   (a) Représentez graphiquement le réseau.
   (b) Calculez la sortie de chaque neurone des couches cachée et de sortie.

(c) Quelle fonction calcule ce réseau ?

3. **Cas genéral :**

(a) Quelle fonction calcule ce réseau ? (vous justifierez de manière détaillée votre réponse)