

TD N° 3
Approches supervisées

Nous disposons des données d'apprentissage suivantes :

A	B	Classe
1	2	X
1	1	X
2	1	X
1	1	X
1	1	Y
2	2	Y
2	2	Y
2	2	Y

et des données de test suivantes :

A	B	A prédire	Classe réelle
1	1	?	Y
1	2	?	X
2	2	?	Y
2	1	?	X

1 Classification bayésienne

1) Prédisez la classe pour les données de test.

2 Arbres de décision

- 1) Construisez l'arbre de décision à partir des données d'apprentissage ci-dessus.
- 2) Quelle est la précision et le rappel lorsque vous appliquez l'arbre construit sur le jeu de données test
- 3) Soit le jeu de test ci-dessous. En supposant l'erreur tolérée de 20%, pouvez-vous élaguer l'arbre construit précédemment ?

3 K plus proches voisins

- 1) En utilisant une la distance city-block et avec $K=3$, classez les instances suivantes :
 - a. $A=0, B=0$
 - b. $A=2, B=3$
- 2) Ecrivez l'algorithme qui permet de prédire la classe d'un point $X = (x_1, \dots, x_m)$. La fonction de distance est supposée être une primitive du système.

4 Algorithme

En supposant la fonction entropie connue comme une primitive du système, donnez le pseudo-code associé au calcul du gain d'information.

5 Questions de cours

Confirmez ou infirmez et commentez les affirmations suivantes.

1. Le jeu d'apprentissage et le jeu de test peuvent contenir des éléments communs
2. La validation simple n'est utile que lorsque le jeu de données de test est volumineux.
3. Un modèle ayant une $AUC = 0.9$ est meilleur qu'un modèle dont l' AUC est 0.7
4. Il est toujours possible d'obtenir un arbre de décision parfait selon le jeu d'apprentissage.
5. La classification bayésienne fonction très bien sur des données numériques
6. Il est toujours possible de trouver une séparation parfaite avec des SVM
7. Les KPP peuvent être appliqués sur tous les types de données.