CY Tech : Cycle Ingénieurs : Deuxiéme Année - MI

Examen de Datamining 2

9 décembre 2020

Durée 3H - Documents de cours et de TDs autorisés

Modalités : Vous devez rendre un document électronique contenant les réponses aux questions ainsi que le fichier contenant le script R.

NOM Prénom :

Exercice 1.1	Exercice 1.2	Exercice 1.3	Exercice 1.4	Exercice 1.5

Exercice 1.6	Exercice 1.7

Exercice 2.1	Exercice 2.2	Exercice 2.3	Exercice 2.4	Exercice 2.5	Exercice 2.6

Exercice 2.7	Exercice 2.8	Exercice 2.9	Exercice 2.10	Exercice 2.11	Exercice 2.12

Notations

Notes globales :

1 Objets, Tailles & Couleurs ..

On souhaite prédire la classe d'un objet (Y=1 ou Y=0) suivant sa taille T (S=Small, M=Medium, L=Large) et sa couleur C (R=rouge, B=bleu).

Y	**T**	**C**
1	S	R
1	S	R
1	L	B
1	M	B
1	M	B
0	S	B
0	S	B
0	S	R
0	L	R
0	L	R
0	M	B

1.1 Modèle bayésien naïf

Exercice 1.1 Utiliser la formule de Bayes pour transformer les probabilités conditionnelles du modèle bayésien naïf : $P(Y = 1|T, C)$ et $P(Y = 0|T, C)$.

Exercice 1.2 Expliquer ce que signifie l'hypothèse d'indépendance dans le modèle bayésien naïf.

Exercice 1.3 Exprimer les probabilités de la question 1 en appliquant l'hypothèse d'indépendance.

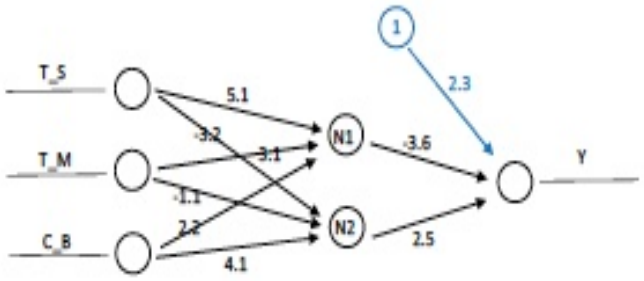
Exercice 1.4 Calculer les 10 probabilités conditionnelles du modèle.

Exercice 1.5 Quelle est la prévision obtenue avec ce modèle pour un objet de taille Small et de couleur rouge ?

1.2 Réseau de neurones

On ajuste le réseau à 2 neurones ci-dessous. La fonction d'activation pour toutes le couches est la fonction Heaviside, $f(x) = 1$ si $x \geq 0$ et 0 sinon. Les neurones d'entrée acceptent des données binaires {0,1} donnant des indications sur la taille et la couleur d'un objet donné.

— $T_S = 1$ ssi la taille de l'objet est S (Small).
— $T_M = 1$ ssi la taille de l'objet est M (Medium).
— $C_B = 1$ ssi la couleur de l'objet est B (Bleu).



Considérons un objet de taille *Small* et de couleur *Rouge*.

Exercice 1.6 Quelles sont les valeurs de sortie des 2 neurones N_1 et N_2 de la couche cachée ?

Exercice 1.7 Quelle est la classe prédite par le modèle .

2 Etude de cas : Heart Disease Recognition

Heart Disease data set consists of 14 attributes data. All the attributes consist of numeric values. The first 13 variables will be used for predicting 14th variables. The target variable is at index 14. Le fichier contenant les données est *heart_ tidy.csv*

Number	Feature Title	Variable Data Type	Feature Categorization
V1	age	Continuous Variable	[29,77]
V2	sex	Categorical Variable	1 = male ; 0 = female
V3	cp : chest pain type	Categorical Variable	1 : typical angina 2 : atypical angina 3 : non-anginal pain 4 : asymptomatic
V4	trestbps : resting blood pressure	Continuous Variable	[94,200]
V5	chol : serum cholestoral	Continuous Variable	[126,564]
V6	fbs : fasting blood sugar > 120 mg/dl	Categorical Variable	1 = true ; 0 = false
V7	restecg :resting ECG result	Categorical Variable	0 : normal 1 : having ST-T wave abnormality
V8	thalach : maximum heart rate achieved	Continuous Variable	[71,202]
V9	exang : exercise-induced angina	Categorical Variable	1 = yes ; 0 = no
V10	oldpeak : ST depression induced by exercise relative to rest	Continuous Variable	[0,6.2]
V11	slope : slope of the peak exercise ST segment	Continuous Variable	[1,3]
V12	ca : number of major vessels	Continuous Variable	[0,3]
V13	thal	Categorical Variable	3 = normal ; 6 = fixed defect ; 7 = reversible defect
V14	Target Variable	Categorical Variable	0 : Absence of Heart Disease 1 : Presence of Heart Disease

Exercice 2.1 Lire le fichier avec la commande

```
heart_df <- read.csv("heart_tidy.csv", sep = ',', header = FALSE)
```

Changer le types des données catégorielles en utilisant la commande *as.factor*.

Exercice 2.2 Séparer les données en deux ensemble : un ensemble d'apprentissage contenant 70% des données et un ensemble de test contenant le reste.

Exercice 2.3 Construire une arbre de décision à partir de l'ensemble d'apprentissage avec comme paramètre de contrôle $minsplit = 10$. Donner les règles qui permettent de prédire la classe *Presence of Heart Disease*

Exercice 2.4 Quelle est l'erreur sur l'ensemble de test ? Calculer la sensibilité et la spécificité de la classe : *Presence of Heart Disease*

Exercice 2.5 Ajuster une forêt aléatoire avec le paramétrage par défaut. Quels sont les 9 variables les plus discriminants ?

Exercice 2.6 Quelle est l'erreur sur l'ensemble de test. Calculer la sensibilité et la spécificité de la classe : *Presence of Heart Disease*. Comparer avec l'arbre de décision.

Exercice 2.7 Construire un classifieur bayésien naïf. Comparer ses performances avec les les deux méthodes précédentes.

Exercice 2.8 Observer les tables de sortie. Choisir deux variables, une catégorielle et une continue. Expliquer les deux tables correspondantes à chacune de ces variables. A quoi correspondent les valeurs dans les tables ?

Exercice 2.9 Donner la valeur de superficie sur la courbe Roc (Area Under Roc) pour le classifier bayésien naif.

Exercice 2.10 Nous souhaitons construire un réseaux de neurone en prenant seulement les 6 variables continues qui sont les plus importantes selon les résultats du RandomForest. Pour construire un dataFrame en prenant en compte certaines colonnes d'un autre dataFrame, vous pouvez utiliser :

```
newDF<-data.frame(oldDF$colName1, .., oldDF$colNamek, ..)
```

Préparer les données avant d'ajuster un réseau de neurones : vous avez besoin de 6 neurones d'entrée et 2 neurones de sortie. Centrer et réduire les données. Si vous êtes ramenés à diviser de nouveau les données en deux ensembles d'apprentissage et de test, utiliser le même *seed* utilisé lors de la première division. Ceci vous permettra de comparer les résultats avec les méthodes précédentes.

Exercice 2.11 Ajuster un réseau de neurones avec une couche cachée ayant 8 neurones avec un maximum d'itérations de 100000. Quelle est l'erreur sur l'ensemble d'apprentissage ? Quel est le nombre de poids à ajuster dans ce réseau. Justifier votre réponse.

Exercice 2.12 Quelle est l'erreur sur l'ensemble de test ? Calculer la sensibilité et la spécificité par rapport à la classe : *Presence of Heart Disease*. Comparer avec les méthodes précédentes.

.