+1/1/60+

QCM **Apprentissage supervisé**

Test
CC du 27/04/2022

Instructions :

- Le sujet comprend 4 exercices pour 22 questions au total. Les questions faisant apparaître le symbole ♣ peuvent présenter plusieurs bonnes réponses. Les autres ont une unique bonne réponse. Des points négatifs pourront être affectés à de *mauvaises* réponses.
- Seul le questionnaire à la 7ème page est à rendre. Vous commencerez par renseigner votre nom et prénom dans la case prévue ainsi que votre **numéro étudiant Rennes 2** (1 case à colorier par colonne). Un numéro d'étudiant mal renseigné entrainera la note de 0.
- Il faut **colorier** les cases correspondants aux bonnes réponses (sur la page 7), mettre une croix dans la case n'est **pas suffisant**. Les cases devront être **coloriées avec un stylo noir** (pas de crayon papier, de stabilo...).
- Le barème sera effectué de la façon suivante :
  - Aucune case coloriée entrainera une note de 0 sur la question.
  - Pour les questions à une seule bonne réponse (sans le symbole ♣), un nombre de points sera affecté (par exemple +2) si la bonne case est cochée. Un nombre de points sera retranché (par exemple -1) si une mauvaise case est coloriée ou si plusieurs cases sont coloriées.
  - Pour les questions avec plusieurs bonnes réponses (avec le symbole ♣), un nombre de points (par exemple +0.5) sera affecté pour chaque bonne réponse coloriée et pour chaque mauvaise réponse non coloriée. Un nombre de points (par exemple -0.5) sera retranché pour chaque mauvaise réponse coloriée et pour chaque bonne réponse non coloriée.
- La correction étant automatique, un non respect des consignes aura forcément un impact sur la note finale.

Durée : 1 heure 20 minutes.

Exercice 1.

On considère $(X_1, Y_1), \ldots, (X_n, Y_n)$ un n échantillon i.i.d où X_i est à valeurs dans $\mathbb{R}$ et Y_i dans $\{0, 1\}$. On désigne par g une règle de classification et par $L(g) = \mathbf{P}(g(X) \neq Y)$ son erreur de classification. On note $g^\star$ la règle de Bayes (règle optimale pour l'erreur de classification).

Question 1 ♣ Soit $x \in \mathbb{R}$. Cocher la (ou les) assertion(s) vraie(s) :

A $g : \{0, 1\} \to \mathbb{R}$

B $g^\star(x) = 1$ si $\mathbf{P}(Y = 1|X = x) = 0.45$

C $g^\star(x) = 1$ si $\mathbf{P}(Y = 1|X = x) < 0.45$

D $g : \mathbb{R} \to \{0, 1\}$

E $L(g^\star) \leq L(g)$

F $L(g^\star) > L(g)$

G $g^\star(x) = 1$ si $\mathbf{P}(Y = 1|X = x) > 0.5$

H $g^\star(x) = 1$ si $\mathbf{P}(Y = 1|X = x) = 0.65$

I $L(g^\star) \geq L(g)$

J *Aucune de ces réponses n'est correcte.*

Question 2 ♣ On suppose de plus que X suit une loi normale $\mathcal{N}(0,1)$ et que, pour $x \in \mathbb{R}$, la loi de $Y|X = x$ est

- une loi de Bernoulli $\mathcal{B}(0.90)$ si $x \geq 0$;
- une loi de Bernoulli $\mathcal{B}(0.25)$ si $x < 0$;

Cocher la (ou les) assertion(s) vraie(s)) :

A $g^\star(-1) = 0$
B $L(g^\star) = 7/40$
C $g^\star(1) = 0$
D $L(g^\star) = 9/40$
E $L(g^\star) = 31/80$
F $L(g^\star) = 9/80$
G $g^\star(x) = 1$ si $x \geq 0$
H $L(g^\star) = 0$
I $L(g^\star) = 1/4$
J $L(g^\star) = 7/20$
K $L(g^\star) = 1$
L $g^\star(x) = 0$ si $x \geq 0$
M *Aucune de ces réponses n'est correcte.*

Question 3 ♣ On considère g_n un algorithme de prévision règle qui souffre de surapprentissage. Cocher la (ou les) assertion(s) vraie(s) :

A g_n possède un biais élevé mais une variance faible
B les données d'apprentissage seront très bien ajustées par g_n
C g_n possède un biais faible mais une variance élevée
D les données d'un échantillon test seront très bien classées par g_n
E *Aucune de ces réponses n'est correcte.*

Question 4 ♣ Soit k un entier plus petit que n. On désigne par $g_{n,k}$ l'algorithme des k plus proches voisins. Cocher la (ou les) assertion(s) vraie(s) :

A $g_{n,k}$ possède généralement une variance faible pour de grandes valeurs de k
B $g_{n,k}$ possède généralement un biais élevé pour de petites valeurs de k
C $g_{n,k}$ aura tendance à surajuster si k est trop grand
D $g_{n,k}$ aura tendance à surajuster si k est trop petit
E $g_{n,k}$ possède généralement un biais faible pour de petites valeurs de k
F $g_{n,k}$ possède généralement une variance élevée pour de grandes valeurs de k
G *Aucune de ces réponses n'est correcte.*

Exercice 2.

On dispose de $n = 20$ observations $(x_i, y_i), i = 1, \ldots n$ avec $x_i \in \mathbb{R}^2$ et $y_i \in \{0, 1\}$. On souhaite construire un arbre de classification pour expliquer les y_i par les x_i. La figure 1 propose deux coupures différentes du nœud racine. On considèrera comme mesure d'impureté d'un nœud N l'impureté de Gini. On rappelle qu'elle se met sous la forme $\mathcal{I}(N) = 2p(1-p)$ où p désigne un paramètre présenté en cours.

Question 5 On désigne par N le nœud racine. L'impureté $\mathcal{I}(N)$ de N vaut :

A 0.5 B 11/100 C 9/20 D 11/20 E 9/100 F 99/200
G 101/200 H 0 I Aucune réponse n'est correcte

Question 6 L'impureté $\mathcal{I}(N1)$ de $N1$ vaut :

A 2/5 B 3/25 C 12/50 D 1/4 E 12/25 F 6/25
G 0.5 H Aucune réponse n'est correcte

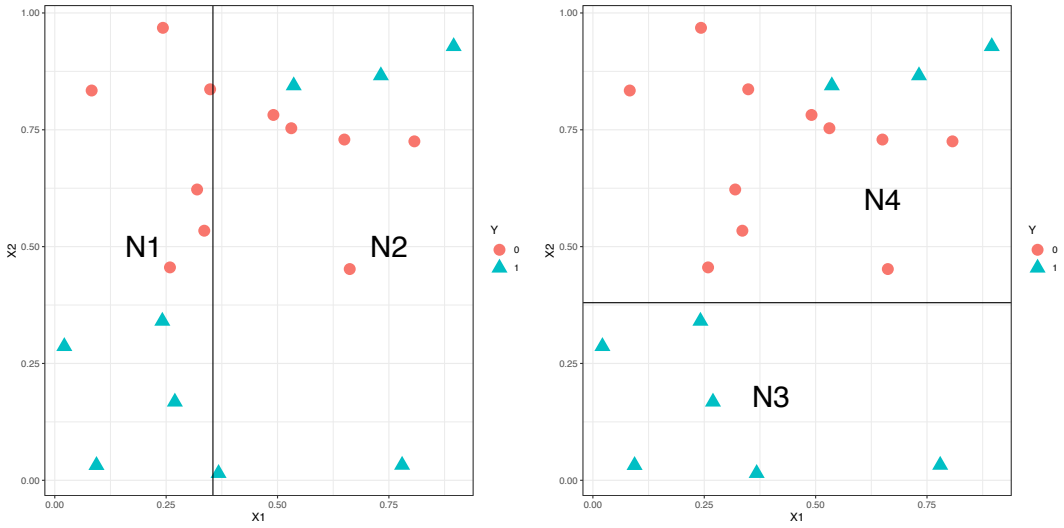

Figure 1: 2 exemples de coupure pour le même échantillon.

Question 7 L'impureté $\mathcal{I}(N3)$ de $N3$ vaut :

A 0 B 0.25 C 0.75 D 3/10 E 0.5 F 0.125
G 1 H Aucune réponse n'est correcte

Pour les questions suivantes, nous considérons les commandes R suivantes.

```
> arbre <- rpart(Y~.,data=df,cp=0.0001,minsplit=2)
> printcp(arbre)

Classification tree:
rpart(formula = Y   ., data = df, cp = 1e-04, minsplit = 2)

Variables actually used in tree construction:
[1] X1 X2

Root node error: ????

n= 20

      CP nsplit rel error   xerror    xstd
1 0.66667    0   1.00000   1.00000   0.24721
2 0.22222    1   aaaaaaa   0.33333   0.17743
3 0.11111    2   bbbbbbb   0.66667   0.22771
4 0.00010    3   ccccccc   0.22222   0.085434
```

Question 8 La colonne `xstd` permet d'estimer la variance (ou l'écart-type) des termes calculés dans la colonne

A `xerror` B `CP` C `nsplit` D `rel error`
E Aucune réponse n'est correcte

Question 9 Retrouver la valeur du terme manquant `????` après `Root node error`.

A 0.45 B 0 C 0.75 D 0.6 E 0.125 F 0.5 G 1
H 0.55 I 0.25 J Aucune réponse n'est correcte

On entre ensuite les commandes suivantes.

```
> arbre1 <- prune(arbre,cp=0.2223)
> arbre2 <- prune(arbre,cp=0.0001)
> rpart.plot(arbre1)
> rpart.plot(arbre2)
```

On obtient les arbres représentés sur la figure 2.

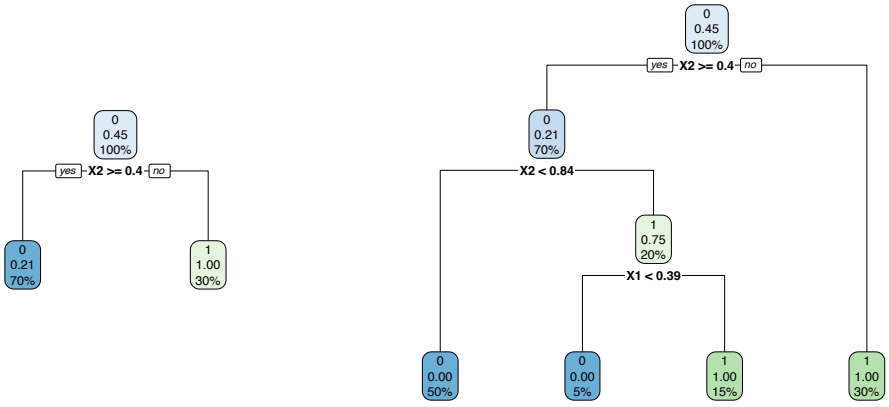

Figure 2: 2 arbres obtenus avec la fonction `rpart.plot` (`arbre1` à gauche et `arbre2` à droite).

Question 10 Retrouver la valeur du terme manquant `aaaaaaa` dans le tableau obtenu avec la commande `printcp`.

A 3/4 B 1 C 1/4 D 3/10 E 1/3 F 7/20
G 1/2 H 3/20 I 3/14 J Aucune réponse n'est correcte

Question 11 Retrouver la valeur du terme manquant `bbbbbbb` dans le tableau obtenu avec la commande `printcp`. **Pas de point négatif à cette question, on peut donc répondre sans craintes**.

A 0 B 1/4 C 1/8 D 1 E 1/9 F 1/11
G 3/20 H 1/2 I 1/3 J 1/20
K Aucune réponse n'est correcte

Question 12 Retrouver la valeur du terme manquant `ccccccc` dans le tableau obtenu avec la commande `printcp`.

A 1/3 B 1/2 C 1/9 D 0 E 1/8 F 1/11 G 1
H 1/20 I 3/20 J 1/4 K Aucune réponse n'est correcte

Question 13 Parmi les arbres suivants, lequel est le plus pertinent selon vous (selon la méthode d'élagage présentée en cours) ?

A `arbre` B `arbre1` C `arbre2` D Aucun

Question 14 ♣ Cette question porte sur les arbres en général (elle n'est pas en lien avec les questions précédentes). Cocher la (ou les) assertion(s) vraie(s)) :

- A Un arbre très profond possède généralement peu de biais
- B Le risque de surapprentissage augmente avec la profondeur de l'arbre
- C Plus un arbre est profond, mieux il ajuste les données
- D Un arbre peu profond possède généralement une forte variance
- E La qualité de prédiction d'un arbre augmente avec sa profondeur
- F *Aucune de ces réponses n'est correcte.*

Exercice 3. Les questions suivantes portent sur les forêts aléatoires.

Question 15 ♣ Cocher la (ou les) assertion(s) vraie(s) :

- A Il n'est pas possible de faire des forêts aléatoires en présence de variables explicatives qualitatives
- B Une forêt aléatoire surapprend lorsqu'on agrège trop d'arbres
- C On doit choisir un nombre d'arbres le plus grand possible pour une forêt aléatoire
- D Une forêt aléatoire permet de réduire le biais des arbres qu'elle agrège
- E On peut faire de la classification supervisée et de la régression avec des forêts aléatoires
- F *Aucune de ces réponses n'est correcte.*

Question 16 ♣ Cocher la (ou les) assertion(s) vraie(s) :

- A On doit choisir des arbres avec un biais élevé pour une forêt aléatoire.
- B On doit choisir des arbres avec un biais faible pour une forêt aléatoire.
- C On doit choisir des arbres avec une variance élevée pour une forêt aléatoire.
- D Il faut utiliser des arbres peu profonds pour les forêts aléatoires.
- E *Aucune de ces réponses n'est correcte.*

Question 17 Au niveau temps de calcul, l'erreur OOB d'une forêt aléatoire se calcule plus rapidement que l'erreur par validation croisée 10 blocs.

A Faux B Vrai

Exercice 4. Les questions suivantes portent sur le logiciel R.

Question 18 ♣ La fonction **tune_grid** du package tidymodels permet

- A d'estimer des risques pour des algorithmes d'apprentissage supervisé.
- B de sélectionner les paramètres de certains algorithmes d'apprentissage supervisé.
- C *Aucune de ces réponses n'est correcte.*

Question 19 ♣

- A L'option `minsplit` dans **rpart** permet de modifier la profondeur des arbres construits
- B L'option `cp` dans **rpart** permet de calculer des erreurs de prévision
- C L'option `cp` dans **rpart** permet de calculer des erreurs d'ajustement
- D L'option `cp` dans **rpart** permet de modifier la profondeur des arbres construits
- E La fonction **prune** du package rpart permet de sélectionner un arbre dans une sous-suite d'abres.
- F *Aucune de ces réponses n'est correcte.*

Question 20 Le paramètre `num.trees` de la fonction **ranger** doit être

A ça dépend B petit C grand

Question 21 Le paramètre `mtry` de la fonction **ranger** doit être

A grand B ça dépend C petit

Question 22 Le paramètre `min.node.size` de la fonction **ranger** doit être

A petit B ça dépend C grand

Feuille de réponses :

0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9

⟵ codez votre numéro d'étudiant ci-contre, et inscrivez votre nom et prénom ci-dessous.

Nom et prénom :

.....

.....

Les réponses aux questions sont à donner exclusivement sur cette feuille : les réponses données sur les feuilles précédentes ne seront pas prises en compte.

Question 1 : A B C D E F G H I J

Question 2 : A B C D E F G H I J K L M

Question 3 : A B C D E

Question 4 : A B C D E F G

Question 5 : A B C D E F G H I

Question 6 : A B C D E F G H

Question 7 : A B C D E F G H

Question 8 : A B C D E

Question 9 : A B C D E F G H I J

Question 10 : A B C D E F G H I J

Question 11 : A B C D E F G H I J K

Question 12 : A B C D E F G H I J K

Question 13 : A B C D

Question 14 : A B C D E F

Question 15 : A B C D E F

Question 16 : A B C D E

Question 17 : A B

Question 18 : A B C

Question 19 : A B C D E F

Question 20 : A B C

Question 21 : A B C

Question 22 : A B C