

# Projet R

## Cours de Programmation

*Vittorio Perduca, Master 1 Mathématiques et Applications*

*UFR Math-Info, Université Paris Descartes, automne 2018*

## Instructions

- Vous avez le droit de travailler en binômes.
- Vous rendrez un rapport écrit sous format **R Markdown** : vous écrirez vos programmes **R** complets dans des blocs de code et vos commentaires et explications dans des paragraphes de texte.
- Vous enverrez vos fichiers **NOM1\_NOM2.Rmd** et **NOM1\_NOM2.pdf** (ou **NOM1\_NOM2.html**) dans un email avec le sujet *Projet R 2018 : NOM1 Prénom1, NOM2 Prénom2* à mon adresse vittorio.perduca@parisdescartes.fr
- La notation prendra en compte **le soin et la clarté des réponses**.
- Une **penalité** est prévue si les instructions ci-dessus ne sont pas respectées.

Dans ce projet vous serez amenés à analyser les données du naufrage du Titanic (1912). Vous utiliserez les données de la compétition de machine learning *Titanic : Machine Learning from Disaster* de Kaggle, une plateforme web organisant des compétitions en science des données. Si certaines de ces compétitions reposent sur des problèmes difficiles et offrent un prix en argent (ou le recrutement) pour les gagnants, beaucoup d'autres sont réservées aux débutants et constituent un cadre idéale pour apprendre à travailler sur des données réelles. La compétition *Titanic* consiste à prédire la survie des passagers du Titanic sur la base de variables telles que le sexe, l'âge et la classe.

## Description des données

### Q.1

Télécharger les données (dite *d'apprentissage*) **titanic\_train.Rdata** disponibles à l'adresse [w3.mi.parisdescartes.fr/~vperduca/programmation/data/titanic\\_train.Rdata](http://w3.mi.parisdescartes.fr/~vperduca/programmation/data/titanic_train.Rdata) et les charger en R à l'aide de

```
load('REPERTOIRE_DE_TRAVAIL/titanic_train.Rdata')
```

Le data frame **train** contient un échantillon de passagers du Titanic que vous utiliserez pour construire un modèle de prédiction de la survie.

### Q.2

Explorer la structure des données :

- donner le nombre d'observations et le nombre de variables
- donner le nom des variables et dire si elles sont quantitatives ou qualitatives
- donner le nombre de valeurs manquantes. Quelles sont les variables avec le plus de données manquantes ?

### Q.3

On considère les variables

- **Survived** dénotée  $S$  dans ce document : survivant ou pas (1/0)
- **Sex**, dénotée  $Sx$
- **Pclass**, dénotée  $P$  : classe de voyage (1, 2 ou 3)
- **Age**

Décrire  $S$ ,  $Sx$ ,  $P$  et  $A$  de manière appropriée.

### Q.4

Construire une nouvelle variable **cAge** qui catégorise **Age** à l'aide de la fonction `cut()` (consulter l'aide!). On considérera les catégories d'âges par tranches de 20 ans, allant de 0 à 80 ans :  $(0, 20]$ ,  $(20, 40]$ ,  $(40, 60]$  et  $(60, 80]$  ans. Décrire cette nouvelle variable, dénotée  $cA$  dans la suite de ce document .

## Liens entres les variables

### Q.5

En utilisant les statistiques descriptives et/ou les graphiques les plus appropriés, décrire le lien entre

- $Sx$  et  $S$
- $P$  et  $S$
- $A$  et  $S$
- $cA$  et  $S$ .

Par la suite nous ne considérerons pas la variable  $A$ , préférant travailler avec  $cA$ .

### Q.6

Commenter les résultats obtenus en formulant une première hypothèse quant à la survie des passagers selon les différentes valeurs de  $P$ ,  $Sx$ , et  $cA$ .

## Prédiction de la survie

### Q.7

On peut estimer la probabilité de survie conditionnellement à la valeur d'une autre variable, à l'aide de formules du type

$$\hat{\mathbb{P}}(S = 1 | Sx = \text{female}) = \frac{n_{1,\text{female}}}{n_{\text{female}}}$$

avec  $n_{1,\text{female}}$  = nombre de survivants parmi tous les passagers femmes et  $n_{\text{female}}$  = nombre total de passagers femmes. Estimer

- $\mathbb{P}(S = 1 | Sx = \text{female})$
- $\mathbb{P}(S = 1 | Sx = \text{male})$
- $\mathbb{P}(S = 1 | P = 1)$
- $\mathbb{P}(S = 1 | P = 2)$
- $\mathbb{P}(S = 1 | P = 3)$

- $\mathbb{P}(S = 1|cA = (0,20])$
- $\mathbb{P}(S = 1|cA = (20,40])$
- $\mathbb{P}(S = 1|cA = (40,60])$
- $\mathbb{P}(S = 1|cA = (60,80])$

## Q.8

Dans le but de construire un modèle de prédiction de la survie en fonction de plusieurs variables, on pourrait imaginer d'estimer les probabilités  $\mathbb{P}(S = 1|Sx, P, cA)$  en adaptant la formule ci-dessus. Par exemple on pourrait prendre

$$\hat{\mathbb{P}}(S = 1|Sx = \text{female}, P = 3, cA = (20,40]) = \frac{n_{1,\text{female},3,(20,40]}}{n_{\text{female},3,(20,40]}}$$

où  $n_{\text{female},3,(20,40]}$  = nombre total de passagers femmes, voyageant en troisième classe et d'âge comprise entre 20 et 40 ans et  $n_{1,\text{female},3,(20,40]}$  = nombre de survivants dans cette même catégorie de passagers. Cette approche pose un problème majeur : en prenant l'intersection de nombreuses strates, il se peut que la catégorie résultante soit vide, ce qui donnerait un dénominateur nul dans la formule précédente. On préfère donc appliquer le théorème de Bayes

$$\mathbb{P}(S = 1|Sx, P, cA) = \frac{\mathbb{P}(Sx, P, cA|S = 1)\mathbb{P}(S = 1)}{\sum_{i=0}^1 \mathbb{P}(Sx, P, cA|S = i)\mathbb{P}(S = i)} \quad (1)$$

et faire l'hypothèse que les variables explicatives  $Sx$ ,  $P$  et  $cA$  sont indépendantes conditionnellement à l'évènement de survie :

$$\mathbb{P}(Sx, P, cA|S = i) = \mathbb{P}(Sx|S = i)\mathbb{P}(P|S = i)\mathbb{P}(cA|S = i). \quad (2)$$

En injectant (2) dans la formule (1) on obtient le modèle dit de *classification naïve bayésienne* :

$$\mathbb{P}(S = 1|Sx, P, cA) = \frac{\mathbb{P}(Sx|S = 1)\mathbb{P}(P|S = 1)\mathbb{P}(cA|S = 1)\mathbb{P}(S = 1)}{\sum_{i=0}^1 \mathbb{P}(Sx|S = i)\mathbb{P}(P|S = i)\mathbb{P}(cA|S = i)\mathbb{P}(S = i)}. \quad (3)$$

Pour coder une fonction qui implémente le classificateur naïf de Bayes, on peut commencer par construire les tables de probabilité conditionnelle correspondantes à  $\mathbb{P}(Sx|S)$  (2 ligne, 2 colonnes),  $\mathbb{P}(P|S)$  (3 lignes, 2 colonnes) et  $\mathbb{P}(cA|S)$  (4 lignes, 2 colonnes). Par exemple, la table nous donnant  $\mathbb{P}(P|S)$  pour toute valeur de  $P$  et  $Sx$  est

```
(S_P <- prop.table(table(train$Pclass, train$Survived), margin=2))
```

On peut donner des noms aux lignes et aux colonnes pour faciliter l'accès aux différents éléments de la table :

```
rownames(S_P) <- c('1', '2', '3')
colnames(S_P) <- c('0', '1')
# Pour extraire P(Pclass = 3 | Survived = 1):
S_P['3', '1']
```

Construire les tables suivantes :

- $S_{Sx}$  pour  $\mathbb{P}(Sx|S)$
- $S_{Ca}$  pour  $\mathbb{P}(cA|S)$ .

On construira aussi la table

```
S <- prop.table(table(train$Survived))
names(S) <- c('0', '1')
```

nous donnant  $\mathbb{P}(S = 0)$  et  $\mathbb{P}(S = 1)$ .

## Q.9

Coder une fonction `prob_prediction(Sex, Pclass, cAge)` qui implémente le classificateur naïf de Bayes de l'équation (3) et rend en sortie la probabilité  $\mathbb{P}(S = 1|Sx, P, cA)$  correspondante aux valeurs données en entrée. On utilisera les tables de probabilité construites au point précédant.

# Evaluation de la performance du classificateur

## Q.10

Télécharger les données (dite *de test*) `titanic_test.Rdata` disponibles à l'adresse

`w3.mi.parisdescartes.fr/~vperduca/programmation/data/titanic_test.Rdata`

et les charger en R. Le data frame `test` contient un deuxième échantillon de passagers qui permet de évaluer la qualité du modèle de prédiction. `test` contient uniquement les variables `Survived`, `Sex`, `Pclass`, `cAge` et n'a pas de valeurs manquantes.

## Q.11

A l'aide de la fonction `prob.prediction()` prédire la probabilité de survie de chaque passager dans `test`. Indications : normalement, en suivant les instructions ci-dessus, `prob.prediction()` doit

- prendre en compte des entrées vectorielles (comme les colonnes de `test`) en rendant en sortie un vecteur de probabilités
- prendre en argument des vecteurs de caractères. Il faudra donc convertir `test$Sex` et `test$Pclass` et `test$cAge` en vecteur de type `character` avant d'appliquer la fonction.

## Q.12

Prédire la survie de chaque passagers dans `test` en utilisant la règle du *Maximum a Posteriori Probability (MAP)* : un passager sera classifié comme survivant si et seulement si sa probabilité de survie est  $> 0.5$ .

## Q.13

Comparer le vecteur des prédictions construit à la question précédente avec le vecteur contenant le vrai status de survie des données `test` à l'aide d'une table de contingence. Calculer la proportion des passagers bien classés, c'est à dire la proportion de passagers se trouvant sur la diagonale de la table de contingence crée. Cette mesure de performance est dite *accuracy* de l'algorithme de prédiction.