Probabilités et statistiques pour l'ingénieur (M1 IAD, VMI)

Devoir Maison

Novembre 2021

Instructions

- Dans ce devoir maison vous devrez analyser un jeu de données en `R`.
- Pour cela, vous pouvez consulter les fonctions de bases que nous avons introduites dans le document `Intro R: types` disponible dans le moodle. Si nécessaire, vous pouvez aussi consulter les polys suivants, que j'ai préparé pour un autre cours: https://helios2.mi.parisdescartes.fr/~vperduca/cours/programmation/R1_intro.pdf et https://helios2.mi.parisdescartes.fr/~vperduca/cours/programmation/R2_stat_desc_graphs.pdf
- Pour la rédaction des réponses, vous avez deux options:
  1. Vous pouvez rédiger vos réponses en utilisant votre éditeur de text préferé (open office, word...) dans le quel vous collerez les graphiques éventuels. Vous **convertirez impérativement** ce fichier en format `.pdf` et vous rendrez ce dernier. Vous rendrez aussi un fichier `.R` avec vos codes.
  2. Vous pouvez rédiger un rapport `R Markdown`, un format qui permet de combiner dans un seul document du texte, du code `R` et les sorties du code. Vous trouverez au lien suivant une introduction très rapide à la syntaxe `R Markdown`: https://helios2.mi.parisdescartes.fr/~vperduca/cours/programmation/R_intro_RMarkdown.pdf
- Vous nommerez le(s) fichier(s) avec vos réponses `NOM_prénom_DM`. Si vous avez plusieurs fichiers vous nommerez vos fichiers `NOM_prénom_DM_fichier1`, `NOM_prénom_DM_fichier2`, ...
- Vous téléverserez vos fichiers dans moodle.
- La date limite est le **21 novembre à minuit**. Les réponses parvenues après ne seront pas évaluées.
- Vous avez le droit de discuter parmi vous des solutions, mais vous rédigerez les solutions de façon individuelle. Les solutions trop similaires entre elles seront pénalisées.
- Vous vous engagerez formellement sur votre honneur à respecter scrupuleusement toutes ces consignes. Le non respect des consignes comportera des points de pénalités voir l'annulation de votre devoir maison.

Engagement sur l'honneur

Vous recopierez et compléterez au début de la première feuille le texte suivant

NOM :

Prénom :

Numéro étudiant :

Master:

Je m'engage sur l'honneur à respecter les consignes pour le DM, et en particulier à rédiger seul(e) les solutions.

Date et signature :

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'age et la classe.

Description des données

Q.1

Télécharger les données (dite *d'apprentissage*) `titanic_train.Rdata` disponibles à l'adresse

https://helios2.mi.parisdescartes.fr/~vperduca/cours/programmation/data/titanic_train.Rdata ou dans moodle.

Charger en `R` ces données à l'aide de

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

Le data frame `train` contient un échantillon de passagers du Titanic que vous analyserez.

Q.2

Explorer la structure des données:

- donner le nombre d'observations (c'est à dire le nombre de lignes du data frame) et le nombre de variables (le nombre de colonnes)
- donner le nom des variables et dire si elles sont quantitatives ou qualitatives (indication: on utilisera les fonctions `str()`, `names()`, `class()`...)

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`, dénotée A

Décrire S, Sx, P et A en utilisant les *statistiques descriptives* et/ou les graphiques les plus appropriées.

Q.4

Construire une nouvelle variable `cAge` qui catégorises `Age` à l'aide de la fonction `cut()` (consulter l'aide!). On considérera les catégories d'ages 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.

Q.7

A l'aide de tests statistiques:

- vérifier si l'age moyenne des passagers est différente de 30
- vérifier si l'age moyenne des passagers ayant survécu est inférieure à 30
- vérifier si l'age moyenne des passagers n'ayant pas survécu est supérieure à 30

Conclure les tests au niveau de confiance $1 - \alpha = 0.90$.

Q.8

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_{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.9 (Bonus facultatif I)

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'age 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

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

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

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

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

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

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.10 (Bonus facultatif II)

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. Prédire la survie de chaque passager du dataset.

Indication: la syntaxe pour définir une fonction en `R` est donnée dans à la page 18 du poly https://helios2.mi.parisdescartes.fr/~vperduca/cours/programmation/R1_intro.pdf