

Statistics Reminder-Introduction to R

First TP

La directrice des ressources humaines tente de comprendre les raisons pour lesquelles les travailleurs quittent leur emploi. Pour ce faire, le département des ressources humaines a créé une base de données contenant la description d'un certain nombre d'employés au cours de l'année écoulée. Votre objectif est de comprendre quels sont les principaux facteurs qui déterminent si un employé va rester ou partir.

- Vous devez effectuer une analyse exploratoire afin de mieux comprendre les données et les différentes variables.
- L'un des principaux objectifs de cette analyse préliminaire est de découvrir quels sont les facteurs les plus importants (attributs) qui déterminent si une personne va quitter l'entreprise ou pas. Votre analyse sera principalement basée sur des inspections visuelles de graphiques décrivant des distributions de probabilités.
- Vous allez utiliser R pour effectuer l'analyse et développer les scripts nécessaires.
- Vous rédigerez un rapport dans lequel vous documenterez les résultats de votre analyse et aborderez également les différentes questions présentées dans les diapositives suivantes.

Analyse exploratoire II

L'analyse exploratoire se déroulera en deux phases :

- Dans la première phase, TP1-A, vous ferez l'analyse exploratoire des attributs *qualitatifs*, vous livrerez un notebook jupyter et un pdf dans lequel vous présenterez vos résultats et répondrez aux différentes questions que nous posons. Cette phase *n'est pas notée*.
- Dans la deuxième phase, TP1-B, vous ferez l'analyse exploratoire pour les attributs *quantitatifs*, vous livrerez un carnet jupyter et un pdf dans lequel vous présenterez vos résultats et répondrez aux différentes questions que nous posons. Cette phase *est notée*.
 - TP-1B étendra les livrables du TP-1A, c'est-à-dire le jupyter notebook et le rapport.
- Dans les deux phases, vous devez répondre à des questions pratiques et théoriques. Prêtez attention aux questions théoriques car elles seront très utiles dans les TP suivants (naive Bayes)

- Vous feriez l'analyse exploratoire qui est demandée/décrit dans les diapositives suivantes en utilisant de Jupyter notebooks. L'objectif est de développer une compréhension des différents concepts de statistiques et de probabilités, des commandes R qui mettent en œuvre ces concepts, les résultats produits par ces commandes et l'interprétation correcte de ces résultats.

Les livrables finaux, décrits en détail dans les diapositives suivantes, sont les notebooks et le rapport contenant les résultats de votre analyse.

Livrables: I, R script/jupyter notebooks

- Vos notebooks doivent être entièrement fonctionnel, sans erreur.
- Ils doivent être générique (c'est-à-dire exécutable sur différents ensembles de données).
- par exemple, vous pouvez avoir une seule fonction de contrôle

```
#####  
# runAnalysis is the main function that controls all your analysis  
#####  
runAnalysis <- function(datasetName="iris.csv", #filename that contains the data  
  hasHeader=TRUE, #does your dataset has a header  
  sep=",", #what is the separator character  
  IndexTarget=5, #the index of the target variable  
  IndexesQualitative=c(0), #a vector with the indeces of the qualitative variables  
  IndexesRM=c(0), #a vector with the indeces of the attributes that should be removed (not analysed)  
  colClasses=c("numeric", #a vector with the types of the different attributes  
    "numeric",  
    "numeric",  
    "numeric",  
    "factor")) {
```

que vous pouvez appeler avec les paramètres appropriés pour votre jeux de données.

bonnes pratiques de codage - déclaration de l'objectif et de l'auteur, ne se répète pas (fonctions, boucles), documentation utilisateur/technique (commentaires), indépendant des données (pas de valeurs codées en dur, réutilisable pour d'autres ensembles de données), structure claire, etc.

Votre rapport doit être rédigé comme s'il était adressé à un manager.

- Introduction : Expliquer brièvement le problème (TP1-A)
- Analyse préliminaire (TP1-A)
- Analyse exploratoire. Dans cette section, vous rendrez compte de votre analyse exploratoire. Elle doit être structurée en deux parties :
 - Analyse exploratoire des attributs qualitatifs (TP1-A)
 - Analyse exploratoire des attributs quantitatifs (TP1-B).
- Résumé des principaux résultats et conclusions - factuel, clair et concis (TP1-A, TP1-B)

pratiques standard *best report writing* - déclaration de l'objectif et de l'auteur, structure claire, orientation vers le lecteur, distinction entre les faits et les opinions, etc. Soumettre le rapport au format .pdf.

Section: Analyse préliminaire I

Avant de commencer l'analyse, vous devez comprendre votre ensemble de données. Pour cela, vous devez répondre clairement aux questions suivantes dans votre rapport :

- Combien d'instances et combien de variables (attributs) y a-t-il ?
- Quelle est la variable cible et est-elle quantitative ou qualitative ?
- Les autres attributs sont-ils quantitatifs ou qualitatifs ?
Fournissez un tableau dans lequel vous indiquez le type de chaque attribut (quantitatif ou qualitatif) et la raison de ce choix. et la raison de ce choix)
- Certaines variables doivent-elles être supprimées de l'analyse ?
Pourquoi ?
- Y a-t-il des données manquantes ?
- Fournir une description de la variable cible et de sa distribution.

Attributs qualitatifs (TP1-A)

Section: Analyse exploratoire des attributs qualitatifs I

Pour chaque attribut **qualitatif** f :

- Calculer la distribution de probabilité $P(f)$
- Calculer la probabilité conditionnelle de la variable cible, y , compte tenu des valeurs des attributs, $P(y|f)$.
- Les visualiser à l'aide des utils appropriées (barplots).

Analyser les probabilités conditionnelles. Existe-t-il des attributs qualitatifs utiles pour distinguer les différentes classes? Si oui, sélectionnez deux ou trois attributs qui, selon vous, sont les plus prédictifs de l'attribut cible y . Expliquez clairement votre raisonnement, comment chacun de ces attributs affecte la valeur de y . Dans le rapport, *incluez uniquement les résultats et la discussion pour les deux ou trois attributs les plus importants* que vous avez sélectionnés.

Section: Analyse exploratoire des attributs qualitatifs II

En plus des deux ou trois attributs les plus importants, sélectionnez et décrivez également un attribut qui n'affecte pas la variable cible. Expliquez clairement pourquoi cet attribut n'affecte pas la cible.

Section: Analyse exploratoire des attributs qualitatifs, Questions théoriques I

- Sélectionnez un attribut qualitatif, f , de votre choix, idéalement avec un petit nombre de valeurs distinctes.
- Ecrire une petite fonction nommée `QualitativeAttrsTheory` qui prend en entrée le jeu de données, l'index de l'attribut y et l'index de la variable cible et :
 - Établit la distribution conjointe $P(f, y)$ où y est votre variable cible.
 - Utilise $p(f, y)$ pour obtenir les distributions marginales $p(f)$ et $p(y)$.
 - Utilise $p(f, y)$ et les distributions marginales $p(f)$ et $p(y)$ pour obtenir les distributions conditionnelles $p(f|y)$ et $p(y|f)$.

Pour les trois derniers points, expliquez clairement dans votre rapport comment vous passez de la distribution jointe à la distribution demandée en utilisant comme exemple les attributs que vous avez sélectionnés.

- Utilisez les distributions que vous venez de calculer pour donner un exemple simple du théorème de Bayes.

Attributs quantitatifs (TP1-B)

Section: Analyse exploratoire des attributs quantitatifs I

Pour chaque attribut **quantitative** f :

- calculer la moyenne μ_f et la variance σ_f^2 .
- calculer les moyennes conditionnelles par classe $\mu_{f|y_1}, \mu_{f|y_2}$ et les variances conditionnelles par classe $\sigma_{f|y_1}^2, \sigma_{f|y_2}^2$
- Ordonner les variables en fonction de leur score d'importance en calculant la différence de moyenne conditionnelle de la classe mise à l'échelle par l'écart-type, $\frac{|\mu_{f|y_1} - \mu_{f|y_2}|}{\sigma_f}$.
- Résumez vos résultats dans un tableau de la forme suivante :

Attribut	Moyenne	Variance	M. cond. classe 1	M. cond. classe 2	V. cond. class 1	V. cond class 2	Score d'importance
----------	---------	----------	-------------------	-------------------	------------------	-----------------	--------------------

Section: Analyse exploratoire des attributs quantitatifs II

Pour chaque attribut **quantitatif** f :

- visualiser la distribution $p(f)$ et les distributions conditionnelles $p(f|y)$ avec l' aide de histogrammes.
- En utilisant le score d'importance des attributs quantitatifs que vous avez calculé ci-dessus, sélectionnez les deux attributs quantitatifs les plus importants et incluez dans le rapport *uniquement les visualisations de leurs distributions*. Décrivez ce que révèlent les distributions conditionnelles. S'agit-il d'attributs qui permettent une bonne discrimination entre les différentes classes ?
- Créez un diagramme de dispersion avec les deux attributs quantitatifs les plus importants et utilisez différentes couleurs (ou symboles) pour distinguer les différentes valeurs de variable cible. La visualisation conjointe des deux attributs permet-elle de séparer les différentes classes ?

Section: Analyse exploratoire des attributs quantitatives, Questions théoriques I

Choisissez une variable quantitative, f , et

- écrivez la distribution normale pour les cas suivants :
 $P(f)$, $P(f|y)$ en utilisant les moyennes et les écarts-types que vous avez déjà calculés, Expliquer ce qui change entre les différentes distributions.
- Ecrire une petite fonction nommée QuantitativeAttrsTheory qui prend en entrée le jeu de données, l'index de l'attribut f et l'index de la variable cible et:
 - Visualise ces distributions normales en utilisant des données artificielles et comparez-les aux histogrammes respectifs que nous obtenons à partir des données.

Discutez des similitudes et les différences. Votre attribut f suit-il approximativement une distribution normale ? Expliquez votre réponse