Data Mining

Elian Cruz Pastor

Cruz Elian

Table des matières

Partie 1

Combien y a-t-il d'instance et d'attributs ?

Il y a 10'000 instances et 11 variables.

Code :

```
# read csv file
myData <- read.table("C:/Users/elian/Documents/HEG/3eme annee/Semestre
6/DataMining/Data Mining/untitled/TPS/HR_prediction-train.csv", sep=",",
header=1)
#get the number of instances and variables
dim(myData)
```

Result :

[1] 10000 11

10000 = Le nombre d'instances

11 = variables

Quelle est la variable cible et est-elle quantitative ou qualitative ?

La variable cible est « Left ».

- left - quitter l'entreprise - variable cible

C'est une variable qualitative car on ne peut pas des valeurs numériques. C'est une réponse avec 2 choix possibles qui sont : Oui / Non.

Les autres attributs, sont-ils quantitatifs ou qualitatifs ?

Id : Quantitatif

Satisfaction_level : Quantitatif

Last_evaluation : Quantitatif

Number_project : Quantitatif

Average_monthly_hours : Quantitatif

Time_spend_company : Quantatif

Work_accident : Qualitatif

Promotion_last_5years : Qualitatif

Departement: Qualitatif

Salary: Qualitatif

Certaines variables devraient-elles être enlevées de l'analyse ? Pourquoi ?

La variable qui pourrait être enlevée est « Id » car elle ne rapporte aucune information expiatoire. C'est un identifiant ***unique*** pour chaque personne.

Y a-t-il des données manquantes ?

Il n'y a aucune donnée manquante.

Code :

```
# Check if there is any missing data
sum(is.na(myData)) # returns the number of missing values
any(is.na(myData)) # returns TRUE if there is at least one missing value
```

Result:

[1] 0

[1] FALSE

0 = Indique le nombre de données manquantes.

FALSE = Indique qu'il ne manque aucune valeur.

Partie 2

Analyse exploratoire

Pour chaque attribut qualitatif f :

- calculer la distribution de probabilité P(f)
- calculer la probabilité conditionnelle de la variable cible, y, données les valeurs de l'attribut P(y|f)

Partie 3

Choisissez une variable qualitative, f, et une variable cible y, et montrez des exemples de la façon dont les règles de probabilité suivantes s'appliquent à elles:

P(f, y) = P(f|y)P(y) = P(y|f)P(f), connue sous le nom de règle de multiplication.

P(y|f) = P(f|y)P(y)/P(f), connue sous le nom de règle de Bayes.

À l'aide des matrices de probabilité que vous avez établies ci-dessus et de la matrice de contingence, expliquez la signification de chacun des termes que vous voyez ci-dessus, c'est-à-dire P(f, y), P(f|y), P(y), ... etc.

Partie 4

Pour chaque attribut quantitatif f :

calculer la moyenne μ(f) et la variance σ2(f)

calculer la moyenne μ(f|cible) et la variance σ2(f|cible)

conditionnée par la variable cible

Classer les variables en fonction de leur importance en calculant la différence moyenne conditionnelle de classe, mise à l'échelle par l'écart type.

Partie 5

Choisissez une variable quantitative, f, et écrivez la distribution normale pour les cas suivants : P(f), P(f|y), expliquez ce qui change entre les différentes distributions. Tracez ces distributions normales, comparez-les aux histogrammes correspondants que nous créerons ci-dessous, discutez des similitudes et des différences.

Partie 6

Pour chaque attribut qualitatif f :

visualisez la distribution de probabilité P(f) et les probabilités conditionnelles P(y|f) sous forme de diagrammes à barres.

Pour chaque attribut quantitatif f :

visualisez la distribution p(f) et les distributions conditionnelles p(f|target) sous forme d'histogrammes.

Quels attributs semblent les plus et les moins utiles pour prédire la variable cible ?

Choisissez deux variables continues que vous estimez être les plus utiles et visualisez leur effet sur la variable cible dans un nuage de points.