Rappel sur les statistiques - Introduction à R

Haute école de gestion de Genève
Informatique de Gestion
Alexandros Kalousis

h e g

Description du problème

Le directeur 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.

TP1 : Analyse exploratoire

- Dans ce premier TP, vous devez effectuer une analyse exploratoire et qualitative des données afin de mieux comprendre les données et les différentes variables.
- Vous allez utiliser R pour effectuer l'analyse et développer les scripts nécessaires.
- Sur la base des résultats de l'analyse, nous devrons rédiger un petit rapport, dans lequel vous fournirez ce qui est demandé ci-dessous et décrirez quels sont, selon votre analyse, les facteurs les plus importants qui déterminent le départ ou le maintien d'un employé.

Produits à livrer 4

D1 : Script R : générique, c'est-à-dire *exécutable sur n'importe quel ensemble de données*

- fonction de contrôle principal unique

```
runAnalysis(datasetName,
            DiscreteAttributesIndeces,
            AttrsToRemoveIndeces,
            ClassAttributeIndex)

#datasetName                : nom du fichier de l'ensemble
d'apprentissage #DiscreteAttributesIndeces : vecteur
d'indices d'attributs discrets #AttrsToRemoveIndeces    :
vecteur d'attributs à supprimer #ClassAttributeIndex    :
indice de l'attribut de classe
```

  - nous tapons runAnalysis('train.csv',...) et nous obtenons tous les résultats

- entièrement fonctionnel - nous le testons et il s'effondre en cas d'erreur = 0 point

Produits à livrer 5

meilleures 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.

D2 : Rapport

- expliquer le problème et discuter des résultats
- résumé des principaux résultats - factuel, clair et concis
- basé sur les résultats de votre code - si le rapport ✓ = code = 0 points

les meilleures pratiques standard de *rédaction de rapports* - déclaration de l'objectif et de l'auteur, structure claire, orientée vers le lecteur, distinction entre les faits et les opinions, etc.

Analyse préliminaire

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

- Combien y a-t-il d'instances et de variables (attributs) ?
- Quelle est la variable cible et est-elle quantitative ou qualitative
- ? Les autres attributs sont-ils quantitatifs ou qualitatifs ?
- Certaines variables doivent-elles être exclues de l'analyse
- ? Pourquoi ? Y a-t-il des données manquantes ?

Analyse exploratoire

Pour chaque attribut **qualitatif** $\mathbf{f}$:

- calculer la distribution de probabilité $P(\mathbf{f})$
- calculer la probabilité conditionnelle de la variable cible, *y*, compte tenu des valeurs d'attribut $P(\mathbf{y}|\mathbf{f})$

TP1 : Partie 1 9

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

- $P(f, y) = P(f|y)PyY) = P(y|f)P(f)$, 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 contiguïté.

A l'aide de vos exemples, 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.

Pour chaque attribut **quantitatif f** :

- calculer la moyenne $\mu(\mathbf{f})$ et la variance $\sigma^2(\mathbf{f})$
- calculer la moyenne $\mu(\mathbf{f}\,|target)$ et la variance $\sigma^2(\mathbf{f}\,|target)$ conditionné par la variable cible
- Classez les variables en fonction de leur importance en calculant la différence de moyenne conditionnelle de la classe mise à l'échelle par l'écart-type.

TP1 : Partie 1 11

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 respectifs que nous allons créer ci-dessous, discutez des similitudes et des différences.

TP1 : Partie 1 12

Pour chaque attribut **qualitatif** $\mathbf{f}$:

- visualiser la distribution de probabilité $P(\mathbf{f})$ et les probabilités conditionnelles $P(\mathbf{y}|\mathbf{f})$ à l'aide de diagrammes à barres

Pour chaque attribut **quantitatif** $\mathbf{f}$:

- visualiser la distribution $p(\mathbf{f})$ et les distributions conditionnelles $p(\mathbf{f}|cible)$ par des histogrammes.

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

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

Conseil : utilisez des couleurs (ou des symboles) différentes pour distinguer les classes de variables cibles pour les diverses combinaisons des deux variables continues.