Prof. Dr. Steffen Wagner
Angewandte Statistik
Fachbereich II
Berliner Hochschule für Technik

Exercise 07: Classification

Machine Learning I – SoSe 2024

This workshop covers hierarchical clustering and soft clustering. At the end of the worksheet there are a couple of written exercises for you to do at home, which should be good practice for the exam.

1 Preparations

1.1 RStudio Project

1. Open your Machine Learning 1 RStudio Project
2. Create an R Script file to perform this exercise

1.2 Required Packages

For this exercise you require the following additional R packages. Please make sure that you have installed them on your computer before coming to the workshop session for the case that Eduroam is not working.

```
# check if packages can be loaded, i.e. they are already installed
library(ISLR2)          # for data sets
library(corrplot)
library(pROC)
```

If you get an error at this stage, you need to install the packages.

1.3 Required Data

In this Worksheet we will use the data set `Diabetes.Rda` that is available via Moodle.

2 Classification

2.1 Using logistic regression in R

Work through the the first two Sections of Lab 4.7 in James et al[1], sections 4.7.1 and 4.7.2.

Remarks:

1. The 8×8 correlation matrix that you obtain is not so easy to interpret. A standard way to plot a correlation matrix is using a heat map. The easiest way to do so in R is using the package `corrplot`:

   ```
   library(corrplot)
   corrplot(cor(Smarket[,-9]))
   ```

   There is also a way to plot this using `ggplot` graphics, but the commands to do this are more complex.

2. Don't forget to `detach` the data frame `Smarket` once you've finished this exercise.

2.2 Classifier for Diabetes

The `Diabetes` data set in Moodle is part of a larger data set collected by the *US National Center for Health Statistics*. Later in ML1 you will use a data set from the same source using more variables, and in ML2 you will learn how to cope with the missing values in these data. For the moment we keep things simple by using just 3 variables.

We will fit a logistic regression model to the variable `YN`, which takes the values "Yes" for patients with Diabetes and "No" otherwise. This dataset contains two explanatory variables `BMI` (Body mass index) and `Age`.

The R script file `Classification_Diabetes.R` is provided as a template for this exercise.

First of all, to get to know the data. Read in the `Diabetes` data set and obtain the following summary statistics.

a) How many observations are there?
b) obtain the frequency table for diabetes status.
c) What is the mean and standard deviation of `BMI` and `Age`?
d) Plot a histogram of `BMI` and `Age`, and a scatter plot of the two.
e) Create a box plot of `BMI` against `YN` and `BMI` against `YN`.

Outline of the logistic regression classification given in the template file is:

- Split the data into a training and test data set, with 2000 ($\approx 20\%$) observations in the test data set.
- Fit the logistic regression model to the training data and look at the model summary.
- Define "High Risk of Diabetes" using a cut off of $\alpha = 0.5$. Obtain the classification matrix
- Compute the sensitivity, specificity and accuracy.
- Plot the ROC curve and calculate the AUC using the package `pROC`.
  Notice that the x-axis is reversed so that the value of the specificity can be read off directly, but the shape of the ROC curve is the same as when the false positive rate is plotted.

[1] https://www.statlearning.com/

- Use `roc()` to obtain the sensitivity and specificity for $\alpha = 0.5$.

Once you have worked through this model, copy and adapt your code to carry out the following tasks.

- Fit a logistic regression model `YN` depending on `Age`.
- Fit a third model with `YN` jointly depending on `BMI` and `Age`.
- Obtain the ROC Curve and AUC for all 3 models using the test data and compare them.
- Obtain the classification matrix, sensitivity, specificity and accuracy for "High Risk of Diabetes" with $\alpha = 0.5$.
- Now change the cut off probability for "High Risk of Diabetes" by setting $\alpha = 0.19$. Obtain the new classification matrix and calculate the sensitivity, specificity and accuracy for this alpha. Interpret these results by completing the following text.

  *A **logistic** regression model was fitted to predict the probability of diabetes based on the variables **BMI and Age**. When predicting that a person is more likely to have diabetes than not have diabetes, the true positive rate was a disappointing **0.019** (using the **test** data set). To identify more people that might have diabetes, the definition of "high risk" was changed to the probability of diabetes being **0.19** or more. This increased the **sensitivity** to **49**%.*

3 Written Exercises

3.1 Logit function

logit(p) is defined to be $\log\left(\frac{p}{1-p}\right)$ for $p \in [0,1]$. Show that if logit(p) $= a$ then

$$p = \frac{e^a}{1+e^a}.$$

Solution:

$$\begin{aligned}
\text{logit}(p) = \log\left(\frac{p}{1-p}\right) &= a \\
\frac{p}{1-p} &= e^a \\
p &= e^a(1-p) \\
p &= e^a - e^a p \\
p + e^a p &= e^a \\
p(1+e^a) &= e^a \\
p &= \frac{e^a}{(1+e^a)}
\end{aligned}$$

3.2 Logistic regression coefficients

A logistic regression with one predictor variable x gives the *linear predictor* logit(p) $= 4 - 2x$

a) What value does p take when $x = 0$.
b) Which value of x gives $p = 0.5$.
c) Let $f(x)$ be the predicted probability p as a function of x. Is $f(x)$ monotonic increasing or decreasing with x?
d) For the linear predictor $logit(p) = \beta_0 + \beta_1 x$, summarise in words the effect of the coefficients β_0 and β_1 on the probability curve?

Solution:

a) $\frac{\exp(4)}{1+\exp(4)} = 0.982$
b) $p = 0.5 \Leftrightarrow 4 - 2x = 0 \Leftrightarrow x = 2$
c) Monotonic decreasing because of the negative coefficient -2.
d) When $x = 0$, logit(p) $= \beta_0$. The larger β_0 is, the larger p is at $x = 0$. The sign of the β_1 coefficient determines whether $f(x)$ is increasing or decreasing. The larger the absolute value of β_1, the steeper the function is.

3.3 Classification matrix

A company has designed a pattern recognition program to identify iPhones from a set of photos of either iPhones or of Android phones. In an experiment, the program correctly identifies 98 from 124 of the iPhones and correctly identifies 90 of the 117 Android phones.

For the purposes of this exercise consider the iPhone as the “positive case”.

a) Construct the classification matrix for this experiment.
b) Calculate the sensitivity and specificity of the program.

Solution:

		Identified as Android	iPhone
Actual	Android	90	27
	iPhone	26	98

- Sensitivity = 98/(98+ 26)=98/124= 0.79.
- Specificty=90/(90+ 27)=90/117=0.769.

3.4 Prosecutor's fallacy

The police in a large city of (1 million inhabitants) arrest a man for theft. A DNA test shows a positive match with a DNA sample taken at a murder crime scene which so far remains unsolved. There is no other evidence linking the thief with the murder case.

This type of DNA matching claims that:

- if he was at the crime scene the probability of a positive test result is is 1 (certain)
- if he was not at the crime scene the probability of a positive test result is is 10^{--5}, i.e. one in a hundred-thousand.

The thief appears in court being tried for the murder case. The prosecutor claims "the DNA test shows that probability this man is innocent is one in a hundred-thousand".

a) What is wrong with the prosecutor's probability statement?
b) Draw a tree-diagram illustrating the given scenario.
c) Use Bayes' theorem to calculate the probability that this thief is the murderer.

Solution:

a) The prosecutor is confusing $P(DNA - Match|Innocent)$ with $P(Innocent|DNA - match)$. This is the "prosecutor's fallacy".
b) "the DNA test shows that probability this man is innocent with probability p" means we need to find:

$$p = P(\text{Innocent}|\text{DNA-match}) = 1 - P(\text{Guilty}|\text{DNA-match}).$$

Let G be the event "the thief is guilty" and M the event that a DNA-match was found. Knowns:

- $P(G) = 10^{-6}$ (with no evidence at all one million inhabitants of this city are equally likely to be the murderer!).
- $P(\overline{G}) = 1 - 10^{-6} = 0.999999$
- Test probabilities are $P(M|G) = 1$ and $P(M|\overline{G}) = 10^{-5}$ We need to find $p = 1 - P(G|M)$.

Using Bayes' Theorem:

$$\begin{aligned} P(G|M) &= \frac{P(M|G)P(G)}{P(M|G)P(G) + P(M|\overline{G})P(\overline{G})} \\ &= \frac{1 \cdot 10^{-6}}{(1 \cdot 10^{-6} + 10^{-5} \cdot 0.999999)} = 0.09091 \\ p &= 1 - P(G|M) = 0.9091 \end{aligned}$$

The probability that he is guilty given the test is positive and no other evidence is about 9%.
“the DNA test shows that probability this man is innocent is 90%” is much less convincing.