Exercise Sheet 3 – Data Mining Wirtschaftsinformatik, HTW Berlin

Martin Spott

last revision 28 April 2020

This exercise is about logistic regression. Please install the library `ISLR` that contains data sets for the book *An Introduction to Statistical Learning* by James, Witten, Hastie and Tibshirani. We are interested in predicting if a person defaults (goes bankrupt) based on their bank account balance, their income and their status (student or not).

For plotting, please again refer to the Cookbook for R.

Exercise 3.1

a) Load the library `ISLR` and explore the data set `Default` used in the lecture.

```
library(ISLR)
```

b) Plot *default* against the other variables (*income*, *balance* and *student*) to see if you can spot dependencies between the variables.

c) Explore how to use colour to distinguish data points with $default = yes$ from $default = no$. Also look at *facets* in *ggplot* and other ways to visualise three or more dimensions in one graphing canvas.

Exercise 3.2

a) Build a logistic regression model to predict *default* from *balance*. The basic function is

```
model <- glm(default ~ balance, data = Default, family = binomial)
```

Explore the model using `summary()`.

b) Predict the probability of *default* for the entire data set. The basic function is

```
prediction <- predict(model, type = "response")
```

Explore the values in `prediction` to get an understanding of what the prediction does. Add `prediction` as a column to the data frame `Default`.

c) Write a function `classify(x, threshold)` that transforms a continuous value x in [0,1] into either a 1 or 0 (boolean) based on a set threshold, i.e. if $x > threshold$ then 1 else 0. Hint: Use the function `ifelse()` which allows vectors as an input x.

d) Apply the function with $threshold = 0.5$ to the column `prediction` and add the result as a column `prediction_balance_0.5` to the data frame default.

e) Explore the differences between `prediction_balance_0.5` and the original value `default` using the following commands to produce a *confusion matrix*:

```
# compute the confusion matrix (contingency table)
with(Default, table(prediction_balance_0.5, default))
```

What do the numbers tell us? Count the number of data points where the prediction is accurate.

f) Change the threshold and rerun the analysis. Compare the results of using different thresholds. Observe how the values in the confusion matrix change.

Exercise 3.3

Re-do Exercise 3.2 with other models and compare the results (summary of the models and the confusion matrices). Which is the best model?

a) Build a logistic regression model that predicts *default* from *student*.

b) Build a logistic regression model that predicts *default* from *student*, *balance* and *income*.