Data Analysis and Visualization Exercise 11

Daniela Klaproth-Andrade, Felix Brechtmann, Julien Gagneur

Quiz

1. Which of the following is true for logistic regression?

a) It assigns classes to the datapoints.
b) There is an analytical solution for the estimation of the parameters.
c) It predicts probabilities for each of the two classes.

2. Suppose you are given a fair coin, p(heads) = 0.5. Which of the following are true about odds and log-odds of head?

a) The odds are 0, and the log-odds are 1.
b) The odds are 0.5, and the log-odds are approximately -0.693.
c) The odds are 1, and the log-odds are 0.
d) The odds are 1, and the log-odds are 1.

3. Let sigm() denote the sigmoid function. Which of the following statements are possible for some value of x or y?

a) sigm(x) = 10
b) sigm(10) = y
c) sigm(x) = - 1
d) sigm(x) = 0.5

4. We are tasked with fitting a logistic regression to detect possible bank frauds that will be further investigated by the bank. Bank frauds are rare but when they occur can be very costly for the bank. Moreover, dealing with false alarms by manual inspection is not too costly. Which of the following is preferable? More than one answer can be correct.

a) High accuracy
b) High recall
c) High precision
d) None of the above

Tutorial

Section 00 - Getting ready

1. Make sure you have already installed and loaded the following libraries:

```
library(ggplot2)
library(data.table)
library(magrittr)
library(tidyr)
library(ggrepel)

library(plotROC)
```

Section 01 - Logistic regression on Diabetes dataset

In this section we are considering the dataset `pima-indians-diabetes.csv` which is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. A more detailed description of the data can be obtained from Kaggle: https://www.kaggle.com/uciml/pima-indians-diabetes-database.

Load the dataset with the following lines of code:

```
diabetes_dt <- fread("extdata/pima-indians-diabetes.csv")
diabetes_dt[, Outcome := as.factor(Outcome)]

# Store feature variables that we will need for later
feature_vars <- colnames(diabetes_dt[,-c("Outcome")])

diabetes_dt
```

1. How balanced are the classes of the diabetes dataset?

2. Create an appropriate plot to visualize the relationship between the `Outcome` variable and the feature variables `Glucose`, `BloodPressure` and `Insulin`. What do you conclude from your visualization?

3. Fit a logistic regression model for predicting `Outcome` only based on the feature `Glucose`. Inspect the coefficients of the model's predictors. According to the model, how much do the odds of getting diabetes increase upon increasing the blood glucose level by 1 mg/dL?

4. Collect the predictions for the model from above for all samples in the dataset. Store the scores in a new column of the original dataset. Visualize the distributions of the scores with an appropriate plot. Which type of distribution would you ideally expect? Hint: Use the `predict()` function.

5. Now, create a function for computing the confusion matrix based on the predicted scores of a model and the actual outcome. The function takes as input a threshold, a data table, the name of a scores column and the name of column with the actual labels. Then, use the implemented function for computing the confusion matrix of the model for the thresholds -1, 0 and 1. Are there any differences? What is the amount of false positives for the last cutoff? You can use the following definition of the function:

```
confusion_matrix <- function(dt, score_column, labels_column, threshold){ }
```

6. Use the implemented function to create a second function for this time computing the TPR and FPR for a certain threshold of a classification model given the predicted scores of a model and the actual outcome. What is the TPR and the FPR of the first model for the thresholds -1, 0 and 1? Plot these values in a scatter plot. Your function should take the same parameters as before and return a data table as follows:

```
tpr_fpr <- function(dt, score_column, labels_column, threshold){
  tpr <- NULL # TODO
  fpr <- NULL # TODO
  return(data.table(tpr=tpr, fpr=fpr, t=threshold))
}
```

Homework

1. Create two further logistic regression models as in section 1.3 for predicting `Outcome`. For one model, use only the feature variable `BloodPressure` for building the model. For the other model, use only the feature variable `Insulin`. Which models have a significant feature?

2. Collect the predictions of each model for all samples in the dataset. Store the scores of each model in a separate column of the original dataset. Visualize the distributions of the scores with an appropriate plot. Which type of distribution would you ideally expect? Hint: Use the `predict()` function.

3. For a systematic comparison of the previously built three models, plot a ROC curve for each model into a

single plot using the function `geom_roc` from the library `plotROC`. Add the area under the curve (AUC) to the plot. Which is the best model according to the AUC?

4. Now, fit a logistic regression model with all feature variables (stored in `feature_vars`). Visualize the distribution of the predicted scores for positive and negative classes. What can you conclude from this visualization regarding the separation of the two classes by the model? Plot once again the previous ROC curves and include the ROC curve of the full model for comparison.