

Statistisches Data Mining (StDM)

Woche 6

Aufgabe 1 Lab

Read and do the excersises of chapter 4.6.1, 4.6.2, and 4.6.5 in ILSR

Aufgabe 2 Logistic Regression (based on an exerice in ISLR)

In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the Auto data set. You might want to take a look at chapter 4.6.2 in ISLR before you start the excerise.

- Create a binary variable, `mpg01`, that contains **TRUE** if `mpg` contains a value above its median, and **FALSE** if `mpg` contains a value below its median. Note you may find it helpful to use the `data.frame()` function to create a single data set containing both `mpg01` and the other Auto variables. Attention (for our German and Swiss students and the lecturer!) `mpg` is millage per gallon, the larger the less fuel the car needs.
- Explore the data graphically in order to investigate the association between `mpg01` and the other features. Which of the other features seem most likely to be useful in predicting `mpg01`? Scatterplots, tables and boxplots may be useful tools to answer this question. Describe your findings.
- Randomly split the data into a training set (80 %) and a test set (20 %). One way is to draw random number from 1 to `nrow(Auto)` using `sample`. Note that `Auto[-idx,]` takes all but the indices.
- Perform logistic regression on the training data in order to predict `mpg01` using all variables except name and `mpg`. What is the error on the training set and on the test set of the model obtained?
- Repeat d) using a knn with $k = 1, 5$ classifier. Produce two matrices with the features 'cylinders', 'displacement', 'horsepower', 'weight', 'acceleration', 'year' one from the training set and one from the test set. See
- Do a clever feature engineering. What preprocessing would you do for knn? Redo the analysis in e).
- What is problematic with the feature `origin` in logistic regression and what could you do to solve it.

Aufgabe 3 Nochmals Krabben (Decission Surface)

- Laden Sie den Krabbendatensatz `crabs` des packages `MASS` ein. Führen Sie mit den numerischen Variablen des Crabs Datensatz eine PCA durch und verwenden PC1 und PC2, als neue

Variablen für einen logistischen Klassifizierer. Bestimmen Sie die Konfusionmatrix auf dem Trainingsset (insample).

- b) Zeichnen Sie, wie im letzten Arbeitsblatt den Scatterplot von $X_1 = PC1$ und $X_2 = PC2$ und färben die Krabben nach Ihrem Geschlecht. Zeichnen Sie die ‘Decision Boundary’, d.h. die Punkte für die gilt $p(X_1, X_2) = p(Y = 1|X_1, X_2) = 0.5$ in den Plot ein. Verwenden Sie dazu die Formel für das odds-ratio:

$$\ln \left(\frac{p(X_1, X_2)}{1 - p(X_1, X_2)} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$