SF2935 Modern Methods of Statistical Learning
Project 1

The project should be done in groups of **two**.

A computer written[1], self-containing, report of the subjects presented below should be handed in no later than **Friday 28 September 2018 24:00** by email to `dabergl@kth.se`. The **subject** of the email should be

`SF2935 Project 1: Full Name 1, Full Name 2`

and **name the document**

`SF2935Project1-FullName1-FullName2.pdf`

If you don't have the course book it can be downloaded as a pdf for free at `http://www-bcf.usc.edu/~gareth/ISL/`. The lab examples in the book can be useful for examples of usages of the functions you need. The books uses R, however you can use other programming languages if you want.

When writing down the report you do not have to insert all of the code into the report, instead describe the process of solving the exercise and also why we do things. As an example in 11 a) on page 171 we are creating a new variable, instead of just giving the code for how this variable is created, mention why we are doing this, why can't we use the original variable?

Classification

In the book four different classification algorithms are described, LDA, QDA, KNN and SVM. In this project will look at these four methods. The written report should contain the following:

- In Figure 1 draw (approximately, by hand) the LDA, QDA and KNN (with $K = 1$) classification boundaries. The plot is available on the course homepage.
- Solve and present solutions to the Conceptual exercises 4 and 5 in chapter 4 on page 168-169.
- Solve and present solutions to the Conceptual exercise 3 chapter 9 on page 369.
- In this problem, you will develop models to predict whether a given car gets high or low gas mileage based on the course books *Auto* data set.[2]

  (a) Create a binary variable, *mpg01*, that contains a 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median. You can compute the median using the *median()* function. 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.

[1] Preferably using LaTeX

[2] This problem is based on exercise 4.11 and 9.7 in the course book

(b) Remove the old mpg variable from the dataset so it is not included when training the models. What would happen if it was still included in the dataset when training the prediction model for mpg01?

(c) 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 and boxplots may be useful tools to answer this question. Describe your findings. Include the plots you used in the report.

(d) Fit the follwing models to predict mpg01 with 10 fold cross-validation and using the variables that seemed most associated with mpg01 in (c); LDA, QDA, KNN. For KNN use several different values for K. What is the cross-validation errors of the models obtained? Which value of K seems to perform the best on this data set?

(e) Using the same variables as in (d) and 10 fold cross-validation fit a support vector classifier to the data with various values of cost, in order to predict mpg01. Report the cross-validation errors associated with different values of this parameter. Repeat, this time using SVMs with radial and polynomial basis kernels, with different values of gamma, degree and cost. The function *tune()* can be useful to find a good values for the parameters. Comment on your results.

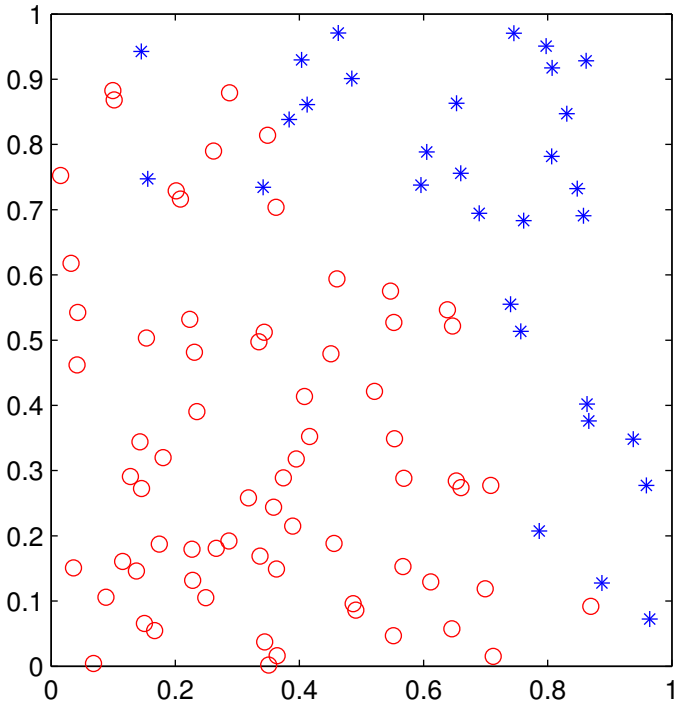

Figure 1: Plot of a two variable classification setting. The two classes are defined with stars and circles.