PROJECT WORK IN SF2935: MODERN METHODS OF STATISTICAL LEARNING (2019)
PART 1

General information

The course has a project part that counts for 3 ECTS (out of the full 7.5 for the course). This part is an essential learning activity in the course and is mandatory in the sense that you cannot obtain a passing grade for the full course without completing the project part.

Regarding programming, although the book focuses on R, you can use other languages if you prefer. When writing the report, you should not include all your code. Instead, focus on describing the methods/techniques used and any active choices you make (the use of a regularisation technique, choice of learning method etc.). Moreover, be sure to describe what any existing libraries/functions that you are using do - check the documentation.

Group organisation and guidelines

The project work is to be done in groups of 2–3 students - exceptions from this must be cleared by the instructor *before* the project work begins. Each group must send an email to pierren@kth.se no later than **September 20**, containing

- a group name
- a list of group members
- the name and email of a "group representative"

The "group representative" will be responsible for all correspondence between the instructors and the group. Information to the group will only be sent to the group representative, who in turn is responsible for relaying relevant information to the other group members.

Grade and bonus points. All members of a group will receive the same grade (pass/fail) for the project. All members of a group are equally responsible for the group work and are obliged to equally contribute to it. At the end of the course, if a group member is of the opinion that this principle did not apply to his/her group, please contact the instructor.

Project 1

The project is divided into three parts, A, B and C, dealing with the classification methods discussed in lectures thus far. Some parts are vague by design, to allow you to explore the different methods in ways you find interesting. You are then required to use your own judgement in deciding on what to include in your report and how to present it.

To complete the assignment you are free to use any software and corresponding packages of your own choice. Please state explicitly what you are using and what the specific function does.

You should *not* describe the theory underlying the tools you chose for your analysis in full detail, unless you make use of techniques not covered in the course. For the latter, a brief review of the technique and the most relevant theoretical results should be included in the report. Make sure to explicitly state what different variables, results etc. refer to.

Part A

The first part of this problem deals with the data set *Auto* in the R library ISLR[1]. Specifically, you will use the data set to train different classifiers to classify a car as having high or low gas mileage.

a) Create your training set by creating a new variable with value '1' if the mpg for a car is above the median of the data set and value '0' if the mpg is below the median. Now remove the 'mpg' variable from the data set in order to exclude it from your training - what would happen if this variable was still available when training the prediction model?
b) Explore the 'new' data set graphically: What features seem most likely to be useful for predicting how a car is classified (high vs. low mpg)? Describe your findings and include relevant plots (scatterplots, box plots, correlation plots etc.).
c) Split the data into a training set and a test set - note the sizes of these sets in your report. Using the variables selected in part (b), train the following models: LDA, QDA, logistic regression and k-nearest neighbours. Record training and test errors for each method. For kNN use different values of k - how does this impact the results?
d) What effect does making the training or test sets smaller/larger have on your results? You can also use some type of cross-validation to evaluate the different models.

Part B

In this part you will explore the differences between logistic regression and LDA using simulated data. Throughout this part you will work on $\mathbb{R}^2$ to enable visualisation of the results in a straightforward manner. Pick a size N for the number of examples you will create - you will then split these into training and test sets as in Part A. Report and visualise your results.

a) For a fraction M/N, $M < N$, for $i = 1, \ldots, M$, construct $x_i = (x_{i,1}, x_{i,2})$ by sampling a Gaussian distribution on $\mathbb{R}^2$; you decide on the parameters of the distribution. These points are classified as '1'. Next, choose different parameters in the Gaussian distribution, sample a new set $\{x_j\}_{j=M+1}^{N}$, a total of $N - M$ points, and classify them as '-1'. After splitting your data set into training and test sets, train logistic regression, LDA and QDA models to predict whether a point belongs to class $y = -1$ or $y = 1$.

   What model performs the best/worst? Are there drastic differences in performance? How does the choice of parameters in the two Gaussian distribution affect your results (for example, how can you make the task more or less difficult)?

[1] If you prefer to work in another programming language you can simply export the relevant parts of the data set

b) Consider the same procedure as in (a) but now rather than having both components of the x_is having a Gaussian marginal, assign the first component of each data point to have some other distribution (examples: pareto, exponential, Poisson, etc.). The second component should still have a Gaussian distribution and the two components sampled independently. Now proceed as before: pick two different sets of parameters, classify the corresponding data points as ± 1 and train logistic regression, LDA and QDA models. Report your results as in (a).

c) Consider the same procedure as in (a) and (b) but now *both* components of each x_i should follow a distribution different from Gaussian (independent between components). Train logistic regression, LDA and QDA models for the classification task (defined as in (a) and (b)) and report your results.

Are there any differences in the performance of any of the methods when comparing the three settings? If so, what potential explanations are there for these differences? Try as much as possible to relate your answer to the (mathematical) concepts discussed during lectures and in the book.

Part C

This part is a conceptual part that does not require any computational work,

a) Throughout this problem, suppose we have a collection of points $\{x_i\}_i$ in $\mathbb{R}^p$, that is there are p features for each input, and corresponding responses $\{y_i\}_i$ in a discrete set of size $K \geq 2$. For each choice of p suppose that X is uniformly distributed over the unit hypercube in $\mathbb{R}^p$. For prediction we want to use nearest-neighbours with only 10% of the range of X closest to that test observation. For example, for $X = 0.6$, consider only observations in $[0.55, 0.65]$. Now consider the following cases:

I) $p = 1$,
II) $p = 2$,
III) $p = 100$.

In each setting, what fraction of the available observations will be used **on average**, in each case, to predict the corresponding response? What is the implication, in terms of large p, for k-nearest neighbours methods? What about LDA or QDA?

b) Consider applying LDA and QDA to a training set. Which of the two will perform better - answer for both training and test set - in the following settings:

i) The Bayes decision boundary is linear.
(2) The Bayes decision boundary is non-linear .
3) When sample size increases, will the prediction accuracy of LDA and/or QDA improve, decline, or remain same as before?

Justify your answers.

Schedule and deadlines - Part 1

- September : Email group information (see above) to pierren@kth.se
- October 11: Deadline for submission of final report (23:59).

The final report should be submitted as a PDF, with the subject of the email reading

SF2935 Project 1: Names

where names refer to the full names of all group members. The document should. be namned in a similar way:

SF2935Project1-Name1-Name2-Name3.pdf.

Good luck!