MA5832-Assessment 2

Weighting: 35% *Total marks: 90*

Due date: Week 4 - Sunday, 31st May 2020, 11:59pm AEST

Overview

In this assessment, you will implement and compare two machine learning algorithms learnt from Week 3 and Week 4 on a real data. In addition, you will solve some analytical questions to develop a conceptual and mathematical understanding of conventional support vector classifiers.

The assessment addresses the following learning outcome(s):

- developing a conceptual and mathematical understanding of conventional support vector classifiers;
- identifying and translating a data science problem into a supervised learning problem;
- identifying appropriate tree-based methods, and support vector classification for descriptive problems;
- application of support vector classifier and tree-based methods covered in Week 3 and 4 to a dataset using the computer language R and the software environment RStudio.

Submission

You will need to submit the following:

- A PDF file clearly shows the assignment question, the associated answers, any relevant R outputs, analyses and discussions.
- **Rmarkdown/R** script file to reproduce your work.
- The task cover sheet

You have up to three attempts to submit your assessment, and only the last submission will be graded.

A word on plagiarism:

Plagiarism is the act of using another's words, works or ideas from any source as one's own. Plagiarism has no place in a University. Student work containing plagiarised material will be subject to formal university processes.

1 Part I: An analytical problem (25 marks)

Note: `svm()` in `e1071` is not allowed to be used in this question.

We consider a training data with 12 observations with two dimensions, (X_1, X_2). For each observation, there is an associated class label $Y = \{-1, 1\}$ as follows

X_1	X_2	Y
1	1	-1
2	4	-1
4	2	-1
6	3	-1
4	5	-1
8	3	-1
1	5	1
2	7	1
4	9	1
3	8	1
3	10	1
0	6	1

1. Draw a scatter plot to represent the points with Red colour for the class $Y = 1$ and Blue colour for $Y = -1$. X_1 is on the vertical axis while X_2 is on the horizontal axis.

2. - Provide the equation for the optimal separating hyperplane (in the form which is similar to equation (4.1) in Topic 2, Week 4).
   - Provide the values of β by using the function `solve.QP()` from `quadprog` package to solve the optimisation problem.
   - Sketch the optimal separating hyperplane in the scatter plot obtained in Question 1.

   (**Hint: `R` might send you an message about the matrix is not positive definite. You can address this problem by adding a small value on the diagonal of the matrix, e.g.** `1e-8`).

3. Describe the classification rule for the maximal margin classifier.

4. Compute the margin for the maximal margin hyperplane.

5. Indicate the support vectors for the maximal margin classifier.

2 Part II: An application

2.1 Background on Credit Card Dataset

The data, **"CreditCard_Data.xls"**, is based on Yeh and hui Lien (2009). The data contains 30,000 observations and 23 explanatory variables. The response variable, *default payment*, is a binary variable. "Yes", equal to 1, indicates clients are able to pay their credit card debt. "No", equal to 0, indicates clients are not able to pay off their credit card debt. The description of 23 explanatory variables is as follows:

- X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
- X2: Gender (1 = male; 2 = female).
- X3: Education (0 = unknown; 1 = graduate school; 2 = university; 3 = high school; 4 = others; 5 = unknown; 6 = unknown).
- X4: Marital status (0 = unknown; 1 = married; 2 = single; 3 = others).
- X5: Age (year).
- X6 - X11: History of past payment. The data was tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -2= no consumption, -1=pay duly, 0 = the use of revolving credit; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
- X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.
- X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

2.2 Assessment Tasks

2.2.1 Data

(a) Select a random sample of 70% of the full dataset as the training data, retain the rest as test data. Provide the code and print out the dimensions of the training data. **(5 marks)**

2.2.2 Tree Based Algorithms

(a) Use an appropriate tree based algorithms to classify credible and non-credible clients. Specify any underlying assumptions. Justify your model choice as well as hyper-parameters which are required to be specified in `R` to estimate the selected model. **(10 marks)**

(b) Display model summary and discuss the relationship between the response variable versus selected features. **(10 marks)**

(c) Summarise the error rates obtained from the training data. Comment on the error rates. **(5 marks)**

2.2.3 Support vector classifier

(a) Use an appropriate support vector classifiers to classify the credible and non-credible clients. Justify your model choice as well as hyper-parameters which are required to be specified in `R` to estimate the selected model. **(10 marks)**

(b) Display model summary and discuss the relationship between the response variable versus selected features. **(10 marks)**

(c) Summarise the error rates obtained from the training data. Comment on the error rates. **(5 marks)**

2.2.4 Prediction

Now apply your fitted models in 2.2.2 and 2.2.3 to make prediction on the test data. Use appropriate methods (at least 2) to summarize the error(s) and/or precision of your models for prediction. Which models do you prefer? Justify your answers. **(10 marks)**

References

Yeh, I.-C. and hui Lien, C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2, Part 1):2473 – 2480.