Statistical learning

Due: 13:00, Friday 21/5/2021 and Friday 22/5/2020 respectively, electronic submission process to be confirmed during the lecture of 18/6/2021

Objectives:

This assignment aims to achieve the following general learning objectives:

- To gain experience with a moderately large real-world data set;
- To apply various statistical learning techniques for regression and classification;
- To gain experience in formal, scientific writing.

Plagiarism:

The Department of Computer Science regards plagiarism as a serious offence. Your submission will be subject to plagiarism checks and appropriate action will be taken against offending parties. You may also refer to the the Library's website at *www.library.up.ac.za/plagiarism/index.htm* for more information.

Report,hand in and mark allocation:

Marks are awarded as indicated in each question.

NOTE the submission dates

Suggested Textbook

Title: An Introduction to Statistical Learning with Applications in R
Authors: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani
Publisher: Springer
Available on-line: *http://www-bcf.usc.edu/ gareth/ISL*

Note that there are also slides and videos on the topics covered in this book - see the above-mentioned link.

or any other relevant material

1: Question 1 - due 21/5/2021

Give a detailed discussion of *linear regression* with refererence to the following concepts. Write a report that covers the aspects below without addressing each as a seperate question. Add any other relevant concepts that might be of importance.

1. Simple and multiple regression
2. The regression model
3. Correlation
4. Goodness of fit / model accuracy
5. Significance of parameters
6. Training and test datasets
7. Reducible and ireducible error
8. Interpretation of modelling results
9. Benefits and disadvantages of using regression analysis (**20**)

2: Question 2 - due 21/5/2021

Give a detailed description of *logistic regression* with refererence to the following concepts. Write a report that covers the aspects below without addressing each as a seperate question. Add any other relevant concepts that might be of importance.

1. The regression model
2. The properties of logistic regression
3. Predicted values and their properties
4. The use of the modelling results w.r.t. classification
5. Benifits and disadvantages of using logistic regression (**20**)

3: Question 3 - due 21/5/2021

Give a brief discussion of linear discriminant analysis. (**10**)

4: Question 4 - due 18/6/2021

Consider the attached data files *insurance.xlsx* and *check2021.xlsx* and the associated description on *http://archive.ics.uci.edu/ml/datasets/Bank+Marketing*. Use the data to fit both a linear and logistic regression model to the data, with the last 10% of the data as test data. Write a report summarising your findings based on the questions below.

1. Give a detailed description of the models that you fit.
2. Fully discuss the estimated results as well as the interpretation thereoff. Include the estimated regression coefficients, significance and interpretation as part of your answer.
3. Evaluate the model accuracy with specific reference to specificity and sensitivity.
4. Use your "best" model to classify the observations in the file check2021.xlsx. Submit your results electronically, using the same structure and variables with an additional column with your classification.
5. Which of the two models is the best model to use for the given data? Fully motivate your answer. (**50**)

Total [**100**]