EEE 526 Exam Report

Komolafe Elisha Ayobami
EEG/2015/061

November 16, 2021

OBAFEMI AWOLOWO UNIVERSITY	EEE 526 Machine Learning and Intelligent Control	Dr. Tokunbo Ogunfunmi
Final Exam (In-Class, Open-Book, Online Exam)		

Name:

KOMOLAFE ELISHA AYOBAMI

I will not give or get unpermitted aid in this exam. and I will report any honor code violations observed by me.

Signature:

← Your signature means you really

Grades:

	Possible Score	**Your Score**
Problem 1	100	
Problem 2	100	
Problem 3	100	
TOTAL	**300**	

- The Exam has a duration of up to 3 hours. Please upload your solutions (Jupiter notebook is preferred and PDF) on Google Classroom.
- You are allowed to ***use class notes and the texts*** prescribed for the course. All Python program should be written in Jupiter notebook.
- You cannot use Google search on the internet during the test and all work turned in must be your own. We expect strict adherence to the Obafemi Awolowo University Honor Code

Question Q.1

Investigating the space shuttle "Challenger" Accident using Logistic Regression

Q.1.1 Fitting a Logistic model

Q.1.1.1 data processing

Q.1.1.2 what are the coefficients?

the coefficients of the Logistic regression model for the classes 'No' and 'Yes', which tell us whether there is a probability of failure of not is, [-0.22950051]

Q.1.1.3 interpreting the coefficients

the coefficients can be interpreted to the be weight that is applied to the temperature given to the model to calculate whether there will be a probability of failure with the o-rings.

Q.1.2 removing data for flight 18

on removing the data from flight 18 the new coefficients for the model is [-0.32482337] which now shows there is a much higher reduction in the influence of temperature in the model.

Q.1.3 prediction for 31 degrees

the probability of the o-ring failing at 31 degrees is Yes, which means the plane would crash if launched.

Question Q.2

Multiple Linear Regression

Q.2.1 numerical and graphical summaries

By using the matplotlib and seaborn libraries, histograms,boxplots and the pair plot of the data was generated.
The Data shows a high linear correlation among all the variables and each other

Q.2.2 significant predictors

4 models which used each of the predictors as the dependent variable were created and based on the R^2 value the predictors can be listed based on how significant they are to the models.

- EXAM 1,this is because the drop in the model without it shows a large decrease.
- EXAM 2
- EXAM 3
- FINAL

Q.2.3 hypothesis testing

assuming a hypothesis with the final exam model which says *the final exam score of a student with average scores in the 3 exams is 50* the t test values from the models shows the exam 1 is very significant thus dismissing the null hypothesis to accept the alternative which is with average scores a student will score higher in the final exam

Q.2.4 forcast prediction

3 data for students were tested to generate the results of 83, 98 and 160 for the final scores

Question Q.3

Short Answer Questions

Q.3.1 regression model choice

In What cases would you use a Logistic Regression model instead of Linear Regression model?

The Main difference between a logistic and linear regression model is that is best used for classification problems while the other is used for regression problems.
The Linear Regression is a method that produces a straight line as the predicted relationship between the independent and dependent variables, this makes it very suitable in cases where continuous values are used, and are the predicted variables.
The Logistic Regression is a method that is best suited for classification problem where in there is distinct class for which the algorithm will be needed to choose between, e.g. classifying a set of input to either a pass or fail grade for a student.

Q.3.2 Explain why we need resampling methods?

Q.3.2.1 3 Examples of re-sampling methods

the Three types of re-sampling methods are:

- Cross-Validation
- Bootstrapping
- Pre-validation

Q.3.2.2 which re-sampling method would you use for an application and why?

for the re-sampling methods listed above the applications I used them based on their operation are:

Cross-validation: this method involves splitting the training data into blocks for training, testing, and for some variants hold out. the best application for the cross-validation resampling method is for a problem where the training data is not very large and the samples are large enough so as not to become computational intensive, and one where the training and test error might not be as low as they can be.

Bootstrapping: this method is similar to the cross-validation, but unlike the previous method where all the data available is used, not all are used here. this method involves splitting the data is chunks but instead of using all the data we replace some of it with repeated data. this method is used in problems that use time-series data such as EEG-data for stroke prediction where the data can be bootstrapped to generate more data.

Pre-Validation: unlike the previous methods which are used on data that corresponds to some output this method was designed to able to generate a fairer version of a independent used for comparison. this method is best suitable for Genetic studies and climate prediction where current data is biased with the result already and to be able to generate fair models for use.

Q.3.3 What is support Vector Machine(SVM)?

A support Vector Machine (SVM) is a method used for solving classification problems in machine learning, it classifies a input based on which side of the hyperplane it is on. it is best used when there are only 2 classes to be classified with a linear decision boundary.

Q.3.4 Differentiate between K-Means Clustering and K-Nearest Neighbor (KNN)

Both methods are supervised learning methods where the output for the data is known. the difference between both methods ares:

K-Means Clustering: This method classifies data according to a cluster, which means each unique data can be identified by a cluster, the clusters are made by generating a 'field' around all the data that have the same characteristics based on the variables used to classify them.

K-Nearest Neighbor: Unlike the method described above the KNN is not a method for clustering but a method for classification.

Q.3.5 Differentiate between Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA)

both of the listed methods are discriminant methods that are used when the classes to be divided are more than 2, but the difference between them is that while the LDA uses a linear spline to split the data, the QDA uses a higher order spline to split the data.