**The Analytics Edge** Fall 2018

# Test 1: The Analytics Edge

NAME: ____________________

STUDENT ID: ____________________

HONOR CODE: As a member of the SUTD community, I pledge to always uphold honourable conduct. I will be accountable for my words and actions, and be respectful to those around me. All work turned in for this test is solely my own and I take pride in this.

SIGNATURE: ____________________

The exam is 1:45 hour in duration.
The total number of points is 30.
There are a total of three questions - question 1 contains 10 sub-questions of 1 point each, question 2 contains 10 sub-questions of 1 point each and question 3 contains 5 sub-questions of 2 points each.
Answer all the questions in the boxes provided.
Immediately at the end of the exam, save the history of your R session. To do so, go to File > Save History > "Yourname.txt". Email this file to karthik_natarajan@sutd.edu.sg. This will be used to validate that your work is original.
You are allowed to use only the notes from the class for this exam. The use of the Internet is not permitted.
Good luck!

1. (10 points) A software catalog firm sells games and educational software. The company has put together a collection of items in a catalog, which it recently mailed out to its customers. This mailing resulted in 1000 purchases by customers. The file **software.csv** contains information on the purchases of the customers which include:

   - **freq**: Number of transactions in the last year for the customer
   - **last_update**: Number of days since last update to customer record
   - **first_update**: Number of days since first update to customer record
   - **web_order**: Whether customer purchased by web order at least once? (1 = Yes, 0 = No)
   - **gender**: Gender of customer (1 = Male, 0 = Female)
   - **address_res**: Whether it is a residential address? (1 = Yes, 0 = No)
   - **address_us**: Whether it is a US address? (1 = Yes, 0 = No)
   - **spending**: Amount spent by customer in purchase from the mailed catalog ($)

   Based on this data, the company wants to build a model to predict the spending amount that a purchasing customer will yield. In addition, we include a variable **partition** in the dataset to indicate to which set a customer observation will be assigned (t = training, test = test, v = validation).

   (1) Read the data into a dataframe **software**. Use a plot (or plots) to comment on the validity of the statement:
   *"The amount spent by customers in purchasing from the mailed catalog is normally distributed".*
   Which plot (or plots) did you use to arrive at your conclusion? Use all the observations in the dataset to answer this question.

(2) Compute the average amount of dollars and the variance of the dollars spent by males and respectively females in the dataset. Use all the observations in the dataset to answer this question.

(3) Run a two sample t-test to verify if the average amount of dollars spent by males and females are equal. Write down the null hypothesis, the p-value and your conclusion. Use all the observations in the dataset to answer this question.

(4) Develop a linear regression model to predict the **spending** variable using the variables - **freq**, **last_update**, **first_update**, **web_order**, **gender**, **address_res** and **address_us** as predictor variables (include the intercept). Identify all the variables that are significant at the 0.05 level. Use only the observations in the training set to build the model.

(5) Use the model you fit to answer the following question. Suppose the company is considering two possible scenarios:
Scenario 1: A customer who moves from a non-residential address to a residential address.
Scenario 2: A customer whose number of transactions in a year decreases by 1 unit.
On average, in which of the two scenarios, will the company be worse off in terms of revenue?

(6) Suppose we build a smaller model using only the significant variables with p-values below 0.05. You should leave the intercept in the model. Based on adjusted R-squared, which model would you prefer between this model and the model developed in (4)? Use only the observations in the training set to build the model.

(7) We now use the **regsubsets** function in the leaps package to do subset selection. Suppose we use the backward selection method, which variable is dropped first from the full model? Comment on how this result relates to the p-values that you estimated for the full model in (4). Use only the observations in the training set to build the model.

(8) We now use the adjusted R-squared to pick the best model from the backward selection method. Which predictor variables are included in the model you choose? Remember the intercept should always be included. Use only the observations in the training set to build the model.

(9) We will now choose among the three models identified in (4), (6) and (8) using the validation set. For each of the three models, what is the sum of squared errors in the validation set? Which model would you choose based on this?

(10) We now evaluate the three models on the test set. What are the sum of squared errors in the test set for three models? Are the results consistent with your finding in question (9)?

2. (10 points) Mammography is a commonly used method for breast cancer screening. The file **breastcancer.csv** contains information on a set of digital mammograms collected at an institute of radiology. The dataset consists of the following variables:

   - **age** = Patient's age in years
   - **shape** = Shape of mass (1 to 4)
   - **margin** = Mass margin (1 to 5)
   - **density** = Mass density (1 = high to 4 = low)
   - **severity** = Severity of breast cancer (0 = benign (low) or 1 = malignant (high))

   In addition, the dataset consists a variable **physician** which is the physician's assessment of breast cancer from the mammogram (1 = definitely benign, to 5 = highly suggestive of malignancy). In this question, we will compare how using computer aided diagnosis systems contrasts with physician recommendations from mammogram interpretations.

   (1) Read the data into the dataframe "breast". Which variable has the most number of missing entries in this dataset? How many entries are missing?

(2) A balanced dataset has roughly around 50% observations of each of the type of tumors (benign or malignant). Of course, such a partition will not be always available in practice and for most predictive techniques, a little imbalance is not a problem. Only when the class imbalance is high, roughly around 90% points for one class and 10% for the other, standard fitting methods and performance measures may not be as effective. Is your dataset balanced?

(3) We start by dropping all observations with missing entries using the na.omit function. Now split the dataset into a training and testing set, putting 60% of the data in the training set. Set the seed to 1000 before making the split. Remember to use the TRUE values from sample.split function in the caTools package to put into the training set. Write the estimated equation that predicts the **severity** using **age**, **shape**, **margin** and **density** as predictors (plus the intercept) using the observations in the training set.

(4) Consider the odds ratio for the dependent variable **severity** and the continuous predictor variable **age**. We define this as the ratio of the odds of having a breast cancer for a person who is one year older to the odds of having a breast cancer for a person who is of a certain age. What is the odds ratio in this example?

(5) Compute the log-likelihood of the model on the test set. What is the accuracy of this model with a threshold of 0.5?

(6) Now use only the statistically significant predictor variables (include the intercept) and rebuild the model on the training set. What is the log-likelihood of the model on the test set and the corresponding accuracy with a threshold of 0.5?

(7) In this example, it is better to falsely classify a benign region as malignant rather than to miss a breast cancer by classifying a malignant region as benign. In other words, we are not just interested in the area under the curve but rather a related measure which is the area under the partial ROC curve. Suppose, we are interested in calculating the partial area under the ROC curve for false positive rates which is 0.7 or larger (this corresponds to high true positive rates). What is the area under the partial ROC curve for the model developed in (6)? Hint: Check the "auc" argument in the **performance** function in the ROCR package.

(8) In this question, you will test the robustness of your findings by trying an alternative to logistic regression - namely probit regression where the equation to be estimated is given as

$$P(y = 1) = \Phi(\beta_0 + \beta_1 x_1 + \dots \beta_p x_p)$$

where Φ is the standard normal cumulative distribution function. Build the model with only the statistically significant predictors that you identified in (6) on the training set. Hint: You can fit the model by using **glm** by modifying the family argument with link="probit". What is the log-likelihood of the model on the test set and the corresponding accuracy with a threshold of 0.5?

(9) Based on your results for the models in (6) and (8), select the best option.

A. Logistic regression is slightly better than probit regression.

B. Probit regression is slightly better than logistic regression.

C. Logistic regression is significantly better than probit regression.

D. Probit regression is significantly better than logistic regression.

(10) One of the worries with the physician's estimates captured by the **physician** variable is that the low predictive value of breast biopsy resulting from mammogram interpretation leads to a lot more unnecessary biopsies with benign outcomes. Suppose, we use a threshold of 4 to classify a tumor as malignant. How many biopsies would be unnecessarily done?

3. (10 points) Each question is worth 2 points below.

(1) Suppose you want to calculate the sum given by:

$$\frac{4}{5} + \frac{4 \times 6}{5 \times 7} + \frac{4 \times 6 \times 8}{5 \times 7 \times 9} + \ldots + \frac{4 \times 6 \times \ldots \times 100}{5 \times 7 \times \ldots \times 101}.$$

Provide a R command that can help you do this. Hint: Check the **cumprod** command.

(2) In a logistic regression model, suppose we have $p$ predictor variables and an additional intercept term to predict the output variable $y \in \{0, 1\}$. Show that in the best fit model, the average probability of predicting $y = 1$ in the dataset is equal to the fraction of 1's in the dataset.

(3) Consider two different predictive models.

Model 1 makes a constant prediction for all observations while Model 2 uses enough parameters to fit the data exactly. Select the best option.

A. Model 1 has low bias and low variance while Model 2 has high bias and high variance.

B. Model 1 has low bias and high variance while Model 2 has high bias and low variance.

C. Model 1 has high bias and low variance while Model 2 has low bias and high variance.

D. Model 1 has high bias and high variance while Model 2 has low bias and low variance.

(4) You have built a multinomial logit model that predicts the choice of customers on their last mile transportation mode among three possibilities - walking, cycling and bus. The average predicted probabilities of choosing these modes are 0.4, 0.2 and 0.4 respectively. Suppose the bus breaks down. What would the new probabilities of choosing walking and cycling respectively be?

(5) In linear regression, we compute a correlation coefficient $\rho$ to measure the strength of the linear relationship between the independent and the dependent variable. A large absolute value of $\rho$ is a proof of cause-effect relationship between the variables. True or False?

END OF PAPER