DATA-LA 322 Final Exam

Fall 2020

Due: Thursday, December 17th by 11:59pm

1. (10 points) You are dealt a hand of 2 cards. What is the probability of getting a pair (two cards of the same number or type (King, Queen, Jack, Ace))? Please compute this in two ways: 1) as a Monte Carlo simulation, 2) as a calculated probability.

2. (5 points) Women's shoe sizes can be represented as a normal distribution with a mean of 9.6 inches and standard deviation of 0.5 inch. What is the **probability** that a randomly selected woman has feet smaller than 8 inches?

3. (5 points) We will use the dataset: **College**, which can be found in the ISLR package (see the end of this doc for more info). It provides information on US colleges in 1995. Please plot a histogram of college graduate rates (variable: Grad.Rate). Discuss any problems that you see, and exclude any values (if needed).

4. (5 points) Assuming normality, please construct a 95% Confidence Interval (CI) for the college graduate rates plotted above (variable: Grad.Rate). What does this CI tell us?

5. (5 points) What is the t-distribution? When do we use it? Calculate the 95% CI for the Grad.Rate variable from the previous question using a t-distribution. Is it more appropriate to use a standard normal distribution or t-distribution to calculate a CI for this variable?

6. (10 points) Run a linear regression model to see if you can predict the number of applications a university/college receives (variable: Apps).
   a. First, you should examine this variable: Apps to see if we should transform it. Given the skew in the data, it may make sense to take a log transformation of App.
   b. Create a model to predict Apps (log transformed) that includes (as predictors): the number of students from top 10% of their high school class (Top10perc), out of state tuition (Outstate), the percent of faculty with PhDs (PhD), and the graduation rate (Grad.Rate). It makes sense to standardize our independent variables so that we can easily compare them in our model (and also so that our machine learning algorithms are more appropriate). Mutate each independent variable so that it is standardized. This can be done using the scale() function. Example: std_Top10perc = scale(Top10perc).
   c. Examine the regression model summary and determine 1) whether or not each predictor is statistically significant and 2) the direction of the effect (is there a positive association or negative association).
   d. Next, check the assumptions of this model, including: 1) the residuals are normally distributed with equal variance across the predictor (and no outliers), 2) no influential points, 3) no multicollinearity (check for correlations), 4) independent errors (residuals aren't correlated with each other).

7. (5 points) What is a p-value? What does it tell us? What are the problems with using p-values?

8. (5 points) Now let's use the College dataset to predict Apps (log transformed) using our machine learning algorithms (still using our standardized predictors listed above). **Note: We will run the machine learning algorithms in the next question.** Create a training and test dataset.

9. (15 points) Run a linear model and kNN model on the train dataset. Predict the number of (log transformed) applications and determine which model is better at predicting the results and collect metrics on them. For the kNN model, tune on hyperparameter k ranging from 1 to 71 by 2. Plot the total within sum of squares for each k value to see which is best.

10. (25 points) Next, we will use the **Smarket** data (part of the ISLR library). This provides data on the stock market. It consists of percentage returns for the S&P 500 stock index over 1250 days from the beginning of 2001 to 2005. It includes data on the percentage returns for each of the five previous trading days (Lag1 through Lag5). We have also recorded Volume (the number of shares traded on the previous day, in billions), Today (the percentage return on the date in question) and Direction (whether the market was Up or Down on this day). We want to predict whether or not the returns should be classified as Up or Down on a particular day. Let's use the two previous days (Lag1, Lag2) as features/predictors.
    a. You will need to create a test and training datasets. (5 points)
    b. Then examine this using **three different classification algorithms** (5 points each).
        i. Several of these algorithms have variables that can be tuned. List what variables can be tuned for each function. Use tune_grid to identify good tuning parameters for the methods that have it.
    c. Summarize your findings with metrics. Which method was best at predicting the outcome? (5 points)

11. (10 points) This question will use the iris data (built into R).

a. Use an unsupervised clustering algorithm to calculate iris clusters based on two or more continuous variables. Tune your hyperparameters to identify the 'best' arguments for your algorithm (using an elbow plot). (5 points)

b. Compare a plot of two continuous variables from the iris dataset colored by cluster to a plot of the same two variables colored by species. How well did your clustering algorithm identify Iris species from the data? (5 points)

Appendix:

The **College** dataset (located in ISLR package, which you may need to install) includes data on 777 colleges/universities in the United States in 1995. It has the following variables:

Private: Indicates if the university / college is public or private
Apps: Number of applications received
Accept: Number of applicants accepted
Enroll: Number of new students enrolled

Top10perc: New students from top 10% of high school class
Top25perc: New students from top 25% of high school class
F.Undergrad: Number of full-time undergraduates
P.Undergrad: Number of part-time undergraduates
Outstate: Out-of-state tuition
Room.Board: Room and board costs
Books: Estimated book costs
Personal: Estimated personal spending
PhD: Percent of faculty with PhDs
Terminal: Percent of faculty with terminal degree
S.F. Ratio: Student/faculty ratio
perc.alumni: Percent of alumni who donate
Expend: Instructional expenditure per student
Grad.Rate: Graduation Rate

The **Smarket** dataset (located in ISLR package) includes data on 1250 daily percentage returns for the S&P 500 stock index between 2001 and 2005. It includes these variables:

Year: The year that the observation was recorded
Lag1: Percentage return for previous day
Lag2: Percentage return for 2 days previous
Lag3: Percentage return for 3 days previous
Lag4: Percentage return for 4 days previous
Lag5: Percentage return for 5 days previous
Volume: Volume of shares traded (in billions)
Today: Percentage return for today
Direction: A factor with levels Down and Up indicating whether the market had a positive or negative return on a given day.