INSTRUCTION AND GROUND RULES

1) **This is an INDIVIDUAL assignment, discussion or collaboration with fellow students is not allowed.**
2) **You are welcome to use your notes, text books, and other references. Please acknowledge any sources you use beyond the course notes and texts.**
3) **Due Date: Wednesday, May 7th, at 11:59pm**
4) **Submit electronically via Collab in the form of a SINGLE pdf file. Please print an appendix R-code, but preserve your .R script files in the event we need to evaluate them.**
5) **Good luck!**

QUESTION #1 – BIAS AND VARIANCE

Sketch typical bias-squared, variance, training error, test error and Bayes (irreducible) error curves on the same plot as a function of flexibility of a given model. The x-axis should represent flexibility and the y-axis represents mean-squared error.

a) What does it mean that a model is more or less flexible? Explain the advantages and disadvantages of a highly flexible model.
b) Explain the shape of, and relationships between, the curves you have sketched.
c) What do resampling techniques, such as cross-validation, LOOCV and boot-strapping have to do with these curves?
d) What is the advantage of using a resampling technique as opposed to a standard validation set approach? How do the resampling techniques affect bias and variance? (If you addressed this question in part d, you do not need to repeat the answer here. Just indicate that it has been addressed.)

QUESTION #2 – LINEAR REGRESSION

In this question we will analyze the training and test data found in the files SYS6018-FinalQ2_train.csv and SYS6018-FinalQ2_test.csv posted under the "Exam" folder on Collab. The column "y" represents the target variable and the columns x1,…,xi represent available explanatory variables.

f) Explore the structure of the data set using visual inspection and basic summary statistics. Describe any characteristics or properties that will be important to consider when constructing a model.
g) Fit a simple linear least squares regression of y ~ x1 + x2 + x3 +x4. Use model selection techniques to choose the which variables to include
    i. Show is the structure of your final model. Why did you choose this form?
    ii. What is the statistical significance of the parameters and the model as a whole?
    iii. Explain in words the meaning of the t-statistic and p-value for the coefficient on x2
    iv. Calculate the mean RSS and R^2 for both the training and test sets
h) Repeat step b) with the full set of explanatory variables.
    i. Show is the structure of your final model.
    ii. How did the included explanatory variables and their respective statistical significance change?
    iii. Explain in words the meaning of the coefficient on x6
    iv. Calculate the mean RSS and R^2 for both the training and test sets
i) Add an interaction term between x2 and x6 to the variables included in your model from part c)
    i. Describe and changes to model structure or significance
    ii. Explain in words the meaning of the coefficients on x6 and the new interaction term and sketch a graph showing the full incremental contribution on the forecast due to x6 depending on the value of x2
    iii. Calculate the mean RSS and R^2 for both the training and test sets

QUESTION #3 – SHRINKAGE METHODS IN REGRESSION

Using the data sets from Question #2, fit a linear *ridge regression* model to the full set of explanatory variables for a range of values for $\lambda = 0$ to $\sim\infty$

a) Describe your procedure
b) Plot the coefficients as a function of λ
c) Plot the mean RSS as a function of λ for both the training and test sets. Describe the meaning of the graph.
d) What value of λ appears best? Why?
e) Compare the structure and performance of the model in iv. to your model from 2c)
f) What are the benefits of the ridge procedure in this case?
   If you fit both a standard least squares model and a ridge regression model on the same set of explanatory variables why would you always expect training error to be lower in the least squares case? Does this also apply to the test error?

QUESTION #4 – NON-LINEAR REGRESSION

In this question we will analyze the training and test data found in the files SYS6018-FinalQ4_train.csv and SYS6018-FinalQ4_test.csv posted under the "Exam" folder on Collab. The column "y" represents the target variable and the column x represents the sole explanatory variable.

a) Explore the using visual inspection and basic summary statistics. Describe any characteristics or properties that will be important to consider when constructing a model.
b) Fit a series of cubic spline models to the training data varying the number of knots
   i. Plot the mean square error as a function of knots
   ii. Using performance on the test set, select the best model
   iii. Write out the expressions for and plot the basis functions for the best model
   iv. Explain how a cubic spline model treats extreme values for the explanatory variables
c) Using the anova() function in R, compare the cubic spline above with the following:
   i. Degree-5 polynomial
   ii. Smoothing Spline (using smooth .spline() with 6 degrees of freedom)
   iii. Local regression (using loess() with a span of 0.2)
d) How does the performance of the above models differ across training and test sets? Are these results expected? Why?

QUESTION #5 – CLASSIFICATION

a) Describe the advantages of using a random forest approach as opposed to a bagging approach on a given dataset for the purposes of classification.
b) If the bagging results and random forest results are similar for a given dataset, what does this imply about the features of the dataset? Explain.
c) Explain how one can determine the importance of a given feature from a random forest model.

QUESTION #6 – PRINCIPLE COMPONENT ANALYSIS

Continuing with the data we used in class to demonstrate PCA, let's incorporate the PCA approach into a predictive model.

Participant	Price	Software	Aesthetics	Brand	OS
P1	6	5	3	4	0
P2	7	3	2	2	0
P3	6	4	4	5	0
P4	5	7	1	3	0
P5	7	7	5	5	1
P6	6	4	2	3	0
P7	5	7	2	1	0
P8	6	5	4	4	0
P9	3	5	6	7	1
P10	1	3	7	5	1
P11	2	6	6	7	0
P12	5	7	7	6	1
P13	2	4	5	6	1
P14	3	5	6	5	1
P15	1	6	5	5	1
P16	2	3	7	7	1

In this dataset, OS is the classification that an individual purchases a particular OS (1) or that a given individual does not purchase a particular OS (0).

a) Using the 'princomp' package in R (you will likely need to install this package first), and the commands shown in our last two classes, compute the principal components of the dataset for yourself.
b) Express Component 3 as a function of the original feature set.
c) Now, use the components 1 and 2 as the features for a quadratic discriminant analysis (QDA) model. How well does the model perform?
d) Now use components 1, 2 and 3 in a QDA model. How does this model perform? (Simple misclassification calculation can be used as the performance metric)
e) Finally, perform QDA on the original features and report on its performance.
d) Explain the difference in the results of each of these three approaches.

PART #2 – OPEN ENDED ANALYSIS

After reading about the growing popularity of data science techniques and their successful application across a diverse range of fields a local wine merchant hires you to consult for him. He hopes that using data mining techniques will allow him to identify "undiscovered" wines in the latest vintage before they are available for tasting. At this pre-release stage he is able to buy any wine for $10/bottle. If the wine turns out to be rated below average (i.e. rating <=5) he has to "fire sale" it for $5/bottle to customers who's palates can't distinguish good from bad. If it turns out to be better than average but not exceptional (i.e. rating >5; 8) he can sell it for a modest profit at $12/bottle. If the wine turns out to be exceptional (rating >=8) he can sell it for $25/bottle.

He has hired you to help him develop a predictive model and a purchasing strategy. You have agreed to provide him with the following deliverables:

a) A regression model that forecasts the expected rating based on a set of observed characteristics
b) A classification model that predicts whether a wine will be below average, above average, or exceptional.
c) I purchasing strategy based on the predictions of one or both of the two models above
d) Full description of your analysis and modeling decisions, including
    i. Exploratory analysis and assessment of the explanatory variables and any exclusions, transformations, interactions, etc applied to the data set
    ii. Structure of the model and how any parameters, settings, functional forms, techniques were decided. Include a description of any other model techniques seriously investigated and why they were not chosen
    iii. Training and validation performance including expected error rates
    iv. Description of your purchasing plan
    v. Forecast of profit for your proposed strategy. For this purpose, assume he has up to $100,000 to purchase inventory for the coming year and there are no limits on the number of bottles he can purchase from a given producer
    vi. Stating any necessary assumptions, can you put a 90% confidence interval on his profit?

To design and build your models, you have data available on last year's vintage including its ultimate rating (in SYS6018-FinalOEA_train.csv) Your model will also be evaluated using performance on a selection of bottles that received early ratings for this year. We have provided the explanatory data in SYS6018-FinalOEA_test.csv. In addition to your write up you need to populate three columns in SYS6018-FinalOEA_predict_*YOURNAME*.csv . The three columns correspond to the forecasts from your regression and classification models as well as how many bottles of each wine he should purchase (assuming $25,000 budget for these early release bottles). When you submit please replace "*YOURNAME*" in the file name with your actual name.