# Practice Final Exam. AMS 580

# Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_SBU ID:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Dear all, the Final Exam (Wednesday, May 17, 8:00-10:45am) is an in-person exam to be held in the same classroom, and with the same zoom link. This is an open-book, open-internet exam – however you must do so entirely on your own. Please email the completed Final exam to Professor Zhu (and not your TA) with one email entitled “Final Exam, AMS 580, Firstname\_Lastname\_ID” with two files attached (1. the RMD file, and 2. the Output file of your RMD in either word or pdf – the file names should be: Firstname\_Lastname\_ID.\*\*\*) no later than 10:45am, at: [wei.zhu@stonybrook.edu](mailto:song.jiecheng@stonybrook.edu)

#### Part I. Supervised Learning with the Unknown Data

The **Unknown.csv** data contain 12 predictors and one binary response variable y (= 0 or 1), which is the true class label**.**

For this dataset, **sensitivity** is defined as a case labeled 1 being classified to label 1, while **specificity** is defined as a case labeled 0 being classified to label 0.

**1.** For the entire dataset, please perform the data cleaning as instructed before; namely, delete observations with missing value(s). Please report how many cases (namely, data points) are left after this step. Then please use the random seed 123 to divide the cleaned data into 75% training and 25% testing.

**2.** Next, Please use the *neuralnet* package in R to build the various neural network classifiers. For this question, we shall NOT perform data standardization (normalization).

1. Now we shall build the best classifier to predict the class label using the training data and the Perceptron model with (i) one hidden layer with 3 neurons, (ii) the default loss function of “sse”, and (iii) the default activation function of “logistic”. Please compute the Confusion matrix and report the sensitivity, specificity, and the overall accuracy using the testing data.
2. Next we shall build the best classifier to predict the class label using the training data and the Perceptron model with (i) one hidden layer with 3 neurons, (ii) the loss function of “ce” (namely, cross-entropy, or the negative log likelihood), and (iii) the default activation function of “logistic”. Please plot the perceptron model obtained using the training data. Please compute the Confusion matrix and report the sensitivity, specificity, and the overall accuracy using the testing data.
3. Which neural network model provides the best overall accuracy among (a) and (b)? For this best model only, please add the predicted label for every test data point. (Please do not print this data set out! It will be used in Question 5 below.)
4. Now we shall use the *randomforest* function in R to build the random forest classifiers. For this question, we shall NOT perform data standardization (normalization).

1. Please first build the best random forest to predict the class label using the training data. Please compute the Confusion matrix and report the sensitivity, specificity, and the overall accuracy using the out of bag (OOB) samples.
2. Next please use this random forest to predict the class label in the testing data. Please add the predicted label for every test data point. (Please do not print this data set out! It will be used in Question 5 below.) Please compute the Confusion matrix and report the sensitivity, specificity and the overall accuracy for the testing data.
3. Please plot the variables importance measures using
   1. *MeanDecreaseAccuracy*, which is the average decrease of model accuracy in predicting the outcome of the out-of-bag samples when a specific variable is excluded from the model.
   2. *MeanDecreaseGini*, which is the average decrease in node impurity that results from splits over that variable. The Gini impurity index is only used for classification problem.
4. Please show the importance of each variable in percentage based on *MeanDecreaseAccuracy*.
5. Now we shall use the *caret* package in R to build the various SVM classifiers. For this question, we shall NOT perform data standardization (normalization).

Now we shall build the best classifier to predict class label using the training data and the SVM with radial basis kernel. We shall find the optimal tuning parameters C and sigma (**σ**) by using the command line:

tuneLength = 10

Please (i) report the optimal parameter values, and (ii) compute the Confusion matrix and report the sensitivity, specificity, and the overall accuracy using the testing data. Please add the predicted label for every test data point. (Please do not print this data set out! It will be used in Question 5 below.

1. Now, we shall build an ensemble classifier using the majority vote of: (1) the random forest, (2) the best neural network model, and (3) the SVM model obtained above. The way the majority vote works is that each case in the testing data is classified into the label with the majority vote from the three classifiers. For example, if a case is predicted to be 1, 1, 0 – then the case is classified as 1. Now, please compare the majority vote to the true class label (y) --- compute the Confusion matrix and report the sensitivity, specificity, and the overall accuracy of our new ensemble classifier using the testing data.

#### Part II. RNN & LSTM with the AMZN Data

This dataset contains the historical stock prices of **Amazon.com (AMZN)**.

We will use a Long Short-Term Memory (LSTM) neural network to predict the daily close price of **AMZN** in this assignment. The data spans from 2021-05-17 to 2022-05-13.

When applying LSTM, it is very important to **normalize** the data. One widely-used method is **min-max scaler**, which means we transform x to be **(x - min(x)) / (max(x) - min(x)).** Instead of using the min and max of that day, we shall use the min and max of the **previous** day (lagged min & max).

1. Plot the close price vs. date to visualize the data we will analyze. Then, use the ‘min-max scaler’ to normalize our stock price data. Scaled\_x = (x-lagged\_min(x))/(lagged\_max(x) – lagged\_min(x)). Please report the values of the last 5 scaled close prices.
2. Divide the cleaned dataset into two parts, the last 5 prices for testing y and the rest for training (the last 5 prices in the training set will be used as testing x). Please report how many days of stock price are divided into the training set.
3. LSTM algorithm creates predictions based on the lagged values, which means we need to look back as many previous values as many points we wish to predict. Here we want to do a 5-day ahead forecast, so we need to base each prediction on 5 data points. (We lag the data 5 times, so that each prediction is based on 5 values, and arrange lagged values into columns) Additionally, keras LSTM expects specific tensor format of shape of a 3D array of the form [samples, timesteps, features] for predictors (X) and for response (y) values. Please create matrices for training and testing predictors and response in the 3D form, and report their dimensions by using dim().
4. Please first build the predictive model to predict 5-day stock close price using the training data and the LSTM method with only one LSTM layer with 100 hidden units, and the loss function of ‘mse’. Please make predictions on the 5 observations in the testing set by using the last 5 in the training dataset and compute the Test MSE using the testing data. Scale the predicted stock price back and plot the 5-day predictions and the true stock close price in the same figure. Also, try to predict the close price of 2022-05-16, in its original scale. (Note: the true closing price on 2022-05-16 was 2216.21.)
5. Please first build the predictive model to predict 5-day stock close price using the training data and the RNN method with only one RNN layer with 100 hidden units, and the loss function of ‘mse’. Please make predictions on the 5 observations in the testing set by using the last 5 days in the training dataset and compute the Test MSE using the testing data. Scale the predicted stock price back and plot the 5-day predictions and the true stock close price in the same figure. Also, try to predict the close price of 2022-05-16, in its original scale.
6. Please compare the performance of the algorithms in Question 4 and Question 5 for the 5-day ahead forecast using the testing data.

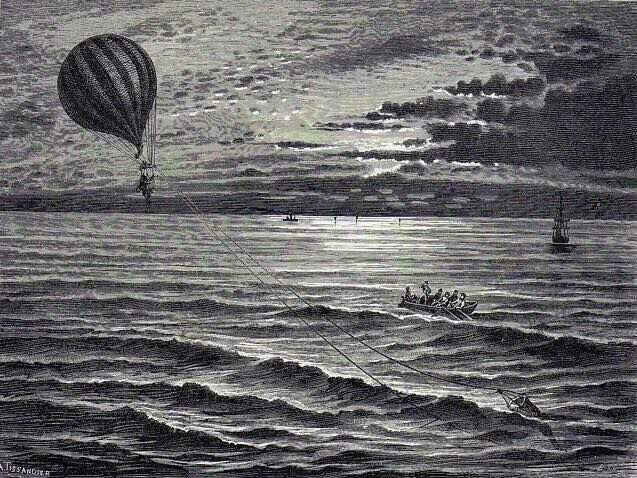
#### Part III. Unsupervised Learning with the Iris Data

The ***iris*** data set, a built-in R data set, was introduced by the British statistician and biologist Ronald A. Fisher in his 1936 paper entitled “The use of multiple measurements in taxonomic problems”. It is also referred to as Anderson's Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species. The data set consists of 50 samples from each of three species of Iris (Iris Setosa, Iris virginica, and Iris versicolor). This sums to 150 records under 5 attributes - Petal Length, Petal Width, Sepal Length, Sepal width and Class (Species).

1. Our goal is to predict the iris Species based on the other 4 attributes given. Subsequently we wish to compare our data-driven clustering to that of the true species classification. For each of the following six clustering methods (K-mean; Hierarchical: Ward, Single-linkage, Complete-linkage, Average-linkage, Centroid), we need you to:
2. Perform the cluster analysis;
3. Draw scree-plot to see whether three clusters are reasonable or not;
4. Show the 2D representation of the Cluster solution;
5. Build a confusion matrix to evaluate the clustering performance;

For Hierarchical clustering, please also draw the dendrogram showing the 3 clusters obtained;

1. Build a confusion matrix to compare the clustering results of the K-means and the Ward’s method;
2. Finally, make a comparison of all six clustering methods, and rank their performance for the given problem. Now imagine you are meeting with a biologist, would you recommend the cluster analysis as an effective way to classify future unknown new flower species based on your analysis of the iris data? If your answer were yes, which method(s) would you recommend?
3. Please write up the entire Rmd code necessary to answer the following questions use the same data set ‘**iris**’:
4. Please compute the Principal Components (PC’s) using the four quantitative attributes and print out a summary of the analysis – and in particular, please point out what percentage of the variations each PC would explain.
5. Please make a biplot, which includes both the position of each sample in terms of PC1 and PC2 and will also show you how the initial variables map onto this.
6. Now you will utilize the Species information binning the iris flowers into three groups: Iris Setosa, Iris virginica, and Iris versicolor. You will visualize the biplot by setting the ellipse argument to be TRUE, which will draw an ellipse around each group.
7. Last but not the least, we shall print out the PC1 as linear combinations of the original variables.
8. Now imagine you are meeting with a biologist; would you recommend the principal component analysis method as an effective way to classify future unknown new flower species based on your analysis of the iris data? Please let the scientist know why your answer is yes or no.



**Unknown Artist?**