Midterm Exam. AMS 580. Spring 2023

Name:_______________________________SBU ID:________________________________

Dear all, you have the entire lecture to work on the midterm – and it is due at 9:50am. This is an open-book, open-internet exam – however you must do so entirely on your own. Please email the completed midterm exam to Professor Zhu (and not your TA Ian) in one email entitled "Midterm, 580" with two file attachments (the Rmd file and the Output file – with all answers included) no later than 9:50am, at: wei.zhu@stonybrook.edu

Part I. Classification Tasks with the Puzzle Data

The **Puzzle.csv** data contain 15 predictors and one binary response variable y (= 0 or 1). **The first 14 predictors are quantitative variables, however, the 15th predictor, *x15* is a categorical (qualitative) variable with 3 categories (1, 2, 3).**
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.

For this part, we shall NOT perform data standardization (normalization).

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 80% training and 20% testing.

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 "sse", and (iii) the default activation function of "logistic". (1) Please plot the perceptron model obtained using the training data. (2) Please compute the Confusion matrix and report the sensitivity, specificity, and the overall accuracy using the testing data. (3) Please add the predicted class label to the testing data.

3. Now we shall use the *randomforest* function in R to build the random forest classifiers. (1) Please first build the best random forest to predict the class label using the training data. (2) Please add the predicted class label using this random forest to the testing data. (3) Please compute the Confusion matrix and report the sensitivity, specificity and the overall accuracy for the testing data. (4) Please plot the variables importance measures using *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.

4. For this question, we shall use *rpart* to build the classification tree and *rattle* to plot the tree. Please first build a fully grown tree using the training data, and then prune the fully grown tree using the training data with 10-fold cross-validation. Please (1) show the complexity plot, (2) report the best CP value, and (3) draw the pruned tree using *rattle*. (4) Please add the predicted class label using this optimal pruned tree to the testing data. (5) Please compute the Confusion matrix and report the sensitivity, specificity, and the overall accuracy for the testing data.

5. Please take the majority vote of the predicted class label obtained in Question 2, 3 and 4 above. Please compute the Confusion matrix and report the sensitivity, specificity, and the overall accuracy of this majority vote for the testing data.

Part II. Regression Analyses & Variable Selections with the Maze Data

The **Maze.csv** data contain 16 predictors and one continuous response variable y.
The first 15 predictors are quantitative variables, however, the 16th predictor, *x16* is a categorical (qualitative) variable with 2 categories (y, n).

For this part, we shall perform data standardization (normalization) for each variable (including both the response variable and the predictors) by subtracting its sample mean and then divide by its sample standard deviation.

1. Please perform data cleaning by checking whether there are any missing values and if so, please delete observations with missing values. Please report how many observations with missing values we have in our dataset. Please use the random seed 123 to divide the data into 80% training and 20% testing. Please normalize the data using the R scale() function or other equivalent function.

2. Please first build the predictive model to predict the house price using the training data and the NN model with (i) one hidden layer with 3 neurons, (ii) the default loss function of "sse", and (iii) the default activation function of "identity". (1) Please plot the model obtained using the training data. (2) Please compute the **RMSE** using the testing data. (3) Please add the predicted **y** to the testing data.

3. Please first build the best random forest to predict **y** using the training data. (1) Please use 10-fold cross-validation to obtain the best tuning parameter **mtry**, which is the best number of random variables to select (without replacement) for each tree node. (2) Next please add the predicted **y** using this random forest to the testing data. (3) Please compute the **RMSE** for the testing data.

4. Please first find the best Elastic Net model using the training data. Please (1) find the best **tuning parameter** values through cross-validation and display these values; (2) display the coefficients of the fitted model; and (3) add the predicted label to the testing data, and report the RMSE and the Coefficient of Determination R^2.

5. Please take the mean of the predicted class label obtained in Question 2, 3 and 4 above. Please compute the **RMSE** of this mean for the testing data.