Midterm Exam 2. AMS 580. Spring 2023

Name:__________________________________SBU ID:___________________________________

Dear all, you have the entire lecture to work on the midterm – and it is due at 10:00am. 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 "Midterm2, 580" with two file attachments (the Rmd file and the Output file – with all answers included) no later than 10:00am, at: wei.zhu@stonybrook.edu

(Note: If you wish to turn in solution to the extra credit problem, you will have three attachments.)

Part I. Classification Tasks with the Puzzle2 Data

The **Puzzle2.csv** data contain 14 predictors and one binary response variable y (= 0 or 1). **The first 13 predictors are quantitative variables, however, the 14th predictor, *x14* 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. 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.

3. 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", (iii) the default activation function of "logistic", and (iv) in building this neural network model, please use only the **top 10 predictors** identified by the random forest procedure in Question 2 above. (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.

4. Now we shall build the best classifier to predict the class label using the training data and the SVM with polynomial basis kernel. We shall find the optimal tuning parameters C, degree and scale by using the command line:

tuneLength = 4

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.

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. Is this ensemble classifier better than those in Questions 2, 3 and 4 above in terms of the overall accuracy? If not, which one is the best?

Part II. Regression Analyses & Variable Selections with the Maze2 Data

The **Maze2.csv** data contain 15 predictors and one continuous response variable y.
The first 14 predictors are quantitative variables, however, the 15th predictor, *x15* 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.

6. 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 75% training and 25% testing. Please normalize the data using the R scale() function or other equivalent function.

7. 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.

8. Please first build the best random forest to predict **y** using the training data. (1) Please use the 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.

9. Please first find the best Ridge Regression model using the training data. Please (1) find the best **λ** value through cross-validation and display this value; (2) display the coefficients of the fitted model; and (3) make prediction on the testing data, and report the RMSE and the Coefficient of Determination R^2.

10. Please take the mean of the predicted response values obtained in Questions 2, 3 and 4 above. Please compute the **RMSE** of this mean for the testing data. Is this ensemble classifier better than those in Questions 2, 3 and 4 above in terms of the RMSE? If not, which one is the best?

11. (**Extra Credit** – a handwritten solution is required – please scan the solution as a PDF file and submit as the third attachment to this midterm.) For a feedforward neural network model with no hidden layer and the identity activation function, when would the loss function of “sse” and “ce” yield identical results? Please show the entire derivation for full credit.

That’s all, folks!