Homework 3

DS440 Data Mining: Fall 2022

Due: November 7 (Monday), 11:59pm

100 points

Classification

For this assignment, we will be looking at following classification algorithms:

1. Logistic Regression
2. Decision Trees
3. Random Forest

Here we will be working with the breast cancer dataset (it is available from sklearn *load_breast_cancer*()). Each row/sample in this dataset contains 30 features and is either put in class 0 or class 1. Hence it is a binary classification problem. The total number of samples are 569.

Question 1: Preprocessing (5 points)

1. Divide the dataset into 2 parts: D_{tr} (training set), and D_{te}(testing set) by randomly placing 75% of the data into D_{tr} and 25% in D_{te}. For this, use train_test_split() from sklearn with a random state of 0.

Question 2: Logistic Regression (40 Points)

1. (**15 points**) Fit a logistic Regression model on D_{tr} with random state 0 and explore differnet solvers. Addtionally use l2 penalty. For details refer to the sklearn documentation: Link
2. (**15 points**)Plot a barplot of the coefficients of logistic Regression model learnt in part 1. On x-axis display the feature names. What information does this barplot give you ?
3. (**5 points**) Predict the Class labels for samples in testing dataset: D_{te}.
4. (**5 points**) Print the accuracy score

Question 3: Decision Tree (25 points)

1. (**2.5 points**) Fit a Decision Tree model on D_{tr} with random state 0. More details on the parameter options available can be found here Link.
2. (**2.5 points**) Print the accuracy on D_{te} using this fitted decision tree model.
3. (**12.5 points**) Vary the parameter $min\_samples\_leaf$ from 1 to 50. For each case fit a decision tree model on D_{tr} and compute the accuracy on D_{te}.
4. (**7.5 points**) Plot accuracy on D_{te} (y-axis) vs $min\_samples\_leaf$ (x-axis) computed in part 3. Print the top accuracy achieved on D_{te}.

Question 4: Random Forest (30 points)

1. (**2.5 points**) Fit a random forest model on D_{tr} with random state 0. More details on the parameter options available can be found here Link.
2. (**2.5 points**) Print the accuracy on D_{te} using this fitted random forest model.
3. (**12.5 points**) Vary the parameter max_depth from 1 to 50. For each case fit a random forest model on D_{tr} and compute the accuracy on D_{te}.
4. (**7.5 points**) Plot accuracy on D_{te} (y-axis) vs max_depth (x-axis) computed in part 3. Print the top accuracy achieved on D_{te}.
5. (**5 points**) Compare the top accuracy achieved by Random forest algorithm (question 4.4) to the decision tree algorithm (question 3.4) in question 3. Comment on what you observe. Does it follow your intuition ?