Big Data Analytics in Physics; Fall 2019; Oct 1, 2019

Name: SOLUTION
Reading Quiz 3

Give short answers to each of the questions.

1) In one or two sentences, describe one advantage of using a Decision Tree for classification.
    - Does not require normalization or scaling of input data
    - Can handle categorical data easily
    - Performance is very easy to explain

2) In one or two sentences, describe one disadvantage of using a Decision Tree for classification.
    - A small change in the input data can lead to a very different tree
    - (Related) There is a very high probability of overfitting your data
    - Trees can become very complex

3) In one or two sentences, describe why using a Random Forest for classification is an improvement over using a Decision Tree.
    - A RF is an ensemble of randomized decision trees. By randomly selecting subsets of the data as well as subsets of the input features for each of the trees in the forest, it is less sensitive to the individual features and less prone to overfitting.
    - A natural output of a RF is a table of feature importances, which tell you which of the input features is most important for classification.

4) How does logistic regression differ from linear regression?

Both are used to establish a relationship between Dependent and Independent variables.
In linear regression the dependent variable is continuous, while in logistic regression the dependent variable is binary (usually 0 or 1).

Linear Regression uses a linear function to map input variables to continuous response/dependent variables. The output of Linear Regression is a continuous value.

Logistic regression uses the logistic function to map the input variables to categorical response/dependent variables. The output is the probability that the given input is one of the two possible binary outcomes.