Big Data Analytics in Physics; Fall 2019; Sept 19, 2019

Name: _____

Reading Quiz 2

Give the best answer(s) among those provided. Some questions have more than one answer.

1) What is multi-class classification (as opposed to binary classification?
   a. It refers to the circumstance where there are 2 and only 2 labeled classes.
   b. It refers to the circumstance where there are more than 2 labeled samples.
   c. It refers to the circumstance where there are 2 or more labeled classes, and at least some of the classes have multiple (more than one) labels

2) Recall that macro-recall is the average of the per-class recall for multi-class classification, while micro-recall is simply the total number of correct predictions divided by the total number of samples. Why might you prefer macro-recall to micro-recall?
   a. Averaging is always better!
   b. If you have many more of a single class, you might prefer macro-recall, since this will not bias your estimate of recall to that of the single largest class.
   c. As usual, it depends on the problem you are studying. There are cases where even if there is significant imbalance in your dataset you might prefer to use micro-recall, while in others you might prefer to use macro-recall.

3) In the context of machine learning, what does k-fold refer to?
   a. This is another name for multi-class classification.
   b. It is a method of cross-validation, where the training data is divided up into sections, each containing k-samples.
   c. It is a method of cross-validation, where the full data sample divided up into "k" sections. Training is run on one section at a time, and the testing is done on the data from the remaining k-1 sections. This process is iterated until all "k" sections have been used in training.
   d. It is a method of cross-validation, where the full data sample divided up into "k" sections. Training is run on data from k-1 sections, and the testing is done on the data from the remaining section. This process is iterated until all "k" sections have been used in testing.

4) In the context of machine learning, what does data augmentation refer to?
   a. Adding new data samples which have been generated in some fashion from data you already have.
   b. Cleaning your data samples so that nulls and categorical data can be properly processed by your estimator.
   c. Getting new original data from sources similar to where you got you original data from.