CST 383: Intro to Data Science
Dr. Glenn Bruns

Test/Train split and cross validation

Purpose. The test/train split and cross validation are used with every kind of supervised learning algorithm. These problems will reinforce your understanding of these key topics.

Instructions. The following reading is recommended but not required. Please at least skim the readings and read what you find most interesting or important. The books for our course can be found at this playlist.

- *Python Data Science Handbook*
  - Section 'Hyperparameters and Model Validation' in Chapter 5, up to, but not including subsection 'The bias-variance trade-off'. (about 5 pages)

Answer the following questions by downloading and editing cross-val.txt.

1. Which of the following is the correct way to use Scikit-Learn's train_test_split:
   a. `X_train, y_train, X_test, y_test = train_test_split(...)`
   b. `X_train, X_test, y_train, y_test = train_test_split(...)`

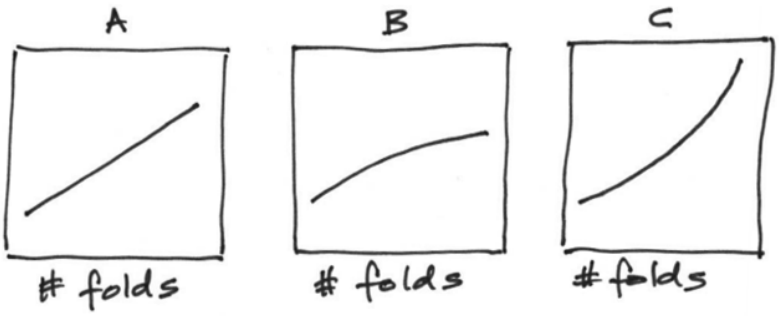
2. If we have 5 predictor variables, then which of the following is a possible value of X_train.shape()?
   a. (1000, 6)
   b. (,5)
   c. (1000, 5)
   d. (,6)

3. If clf is a KNN classifier object, then which of the following can be used to compute test accuracy, assuming y_pred = clf.predict(X_test) ?
   a. `(y_pred = y_test).mean()`
   b. `(y_pred = y_train).mean()`
   c. `(y_pred = y_test).max()`
   d. `(y_pred = y_train).max()`

4. If 10-fold cross validation is used, how many times is training performed?
   a. 1
   b. 9
   c. 10
   d. 10*9
   e. 10*10

5. If 10-fold cross validation is used, how many folds are used each time training takes place?
   a. 1
   b. 9
   c. 10

6. Which of the pictures below best represents how the total time needed to perform cross validation changes as a function of the number of folds?



Submission. Submit cross-val.txt on Canvas.