Problem Set 4

Preliminary Information

For the homework exercises, you will need to write code in the specified functions of "hwcode.py". The variable specified in the open and close parenthesis "(variableName)" of the function should **NOT** be reassigned. They will hold important information - such as the file path of the dataset - required to complete each task. Your job is to write code that returns the required value for each exercise. Overall, this homework will involve writing code outside of Jupyter Lab! You can use any IDE you want. I will provide an overview on how to write code in Spyder on Monday.

Similar to the prior homework assignments, please note that you do **NOT** need to create new functions. The functions have already been specified. You simply need to fill them out with code. Furthermore, do not use any packages except for "csv", "numpy", and "scikit-learn".

Submission Instructions

You will only need to submit the hwcode.py file via Blackboard.

Exercises

You are a data scientist working for a hospital. The hospital wants you to develop a classifier to predict if the sample is from a colon cancer patient given gene expression features. Unfortunately, the data only has 62 training examples, but we have 2000 features for each example. So, you need to use feature selection to avoid overfitting!

Exercise 1 (6 points)

Implement the "hospital_p1(csv_filename)" function to train an SVC (no LinearSVC) model where you grid-search over the C values (.001, .01, .1, 1. 10.), kernel type (rbf and linear), and the number of features using the chi2 feature selection method (10, 100, 500, 1000, 1500). You will need to use the scikit-learn "Pipeline" and "SelectBestK" methods for this part. Return a list containing the "best_score_" in GridSearchCV where the grid search scoring parameter is set to "macro_f1" and the "best_params_" dictionary.

Exercise 2 (2 points)

Implement "hospital_p2(csv_filename)" to return the names ("feature_1", "feature_2", ...) of the 5 most important features based on the chi2 feature selection score in scikit-learn.

Exercise 3 (2 points)

Implement "hospital_p3(csv_filename)" to return the names ("feature_1", "feature_2", …) of the 5 most important features based on the on the coefficients of a LinearSVC model. HINT: You will need to connect the coefficients learned in the LinearSVC method to the feature names that each column of the feature matrix.

Background: Not that this is different than Exercise 3. This is not performing any sort of feature selection. It is simply training a model on all of the features and we are looking at what the model identifies as the most important based on the coefficients learned for each features. Note that the clf.coef_[0] corresponds to the first column of the feature matrix, clf.coef_[1] corresponds to the second item, etc.

Exercise 4 (Extra Credit 2 points)

Write code to create a model in the function "hospital_p4(csv_filename)" that returns a better best_score_ than Exercise 1. You can explore anythign you want here.