Lab 2: Intro to Scikit-Learn

This lab is due at 11:59 PM ET on Sunday, September 16, 2018.

You may work individually or in groups of two. If working in a group, only one student needs to submit the group's shared work via Carmen. Every student must submit his or her own individual/exceptional work. This lab is worth 75 points.

In this lab, you'll get acquainted with the Scikit-Learn machine learning library by performing a classification task using a synthetic dataset and the k-Nearest Neighbors (KNN) classifer (**`sklearn.neighbors.KNeighborsClassifier`**).

1. Load and visualize the data (15 points)

Download the data file (**`lab2b.csv`**) from Carmen.

You do not need to conduct a full exploratory data analysis, but you should visualize the data using a scatter plot. Be sure to show the classes using different colors.

2. Prepare the data (10 points)

Create separate training and test datasets. Use the **`sklearn.model_selection.train_test_split()`** method for this step.

3. Train a classifier and classify the test records (35 points)

Train a k-nearest neighbors classifier using the training data only. Determine the accuracy score using the training data. Use the classifier's built-in **`score()`** method for this.

Classify the test records using the trained classifier.

Plot the predictions and true values side by side. Again, show the two classes in different colors.

Determine the accuracy of these predictions.

4. Conclusion (15 points)

Compare the accuracy figure you got for the training set with that of the predictions on the test set. Were the predictions better or worse? Why? Explain your conclusions, along with any other observations, in a paragraph or two.

5. Submit your work

Save your notebook(s) using a descriptive filename that includes your last name and OSU dot number (e.g. **`lab2_burkhardt_5.ipynb`**) Submit your notebook via Carmen, under Lab 2.

Do not submit the input data file.