DEPARTMENT *of* COMPUTING
College of Business & Technology
EAST TENNESSEE STATE UNIVERSITY

CSCI 5260 – Artificial Intelligence

Lab 10 – Scikit-Learn

Overview

The most common machine learning library in Python is scikit-learn. This lab walks you through the sklearn library to perform machine learning tasks.

Step 1 – Tutorial

Review the tutorial at https://scikit-learn.org/stable/tutorial/basic/tutorial.html.

Step 2 – Explore

Download the **lab10.py** file, which has the following features:

- Imports:
  - **numpy** – for Linear Algebra
  - **matplotlib.pyplot** – for Plotting results
  - Scikit-Learn Data Manipulation:
    - **datasets**
    - **sklearn.model_selection.train_test_split**
    - **sklearn.model_selection.learning_curve**
  - Scikit-Learn ML Classifiers
    - **sklearn.linear_model.LinearRegression**
    - **sklearn.svm.SVC**
    - **sklearn.tree.DecisionTreeClassifier**
    - **sklearn.neighbors.KNeighborsClassifier**
- Functions:
  - **linear_regression**(X_train, Y_train, X_test, Y_test)
  - **support_vector_machine**(X_train, Y_train, X_test, Y_test)
  - **decision_tree**(X_train, Y_train, X_test, Y_test)
  - **k_nearest_neighbors**(X_train, Y_train, X_test, Y_test)
  - split_test_train(test_percent, X, Y)
  - plot_learning_curve(estimator, title, X, y, axes=None, ylim=None, cv=None, n_jobs=None, train_sizes=np.linspace(.1, 1.0, 5))

Step 3 – Complete the Code

A. Setup Classifiers

Each of the four ML functions should have the following basic layout. Use this layout to complete the code for each.

1. Set an `estimator` variable equal to the appropriate classifier.
2. Set a `model` variable equal to the estimator's `fit()` method, passing in the `X_train` and `Y_train` parameters. This returns the model created by the classifier based on the training set.

3. Set a `score` variable by calling the `score()` method, passing in the `X_test` and `Y_test` parameters. This returns the **average accuracy** of the model.
4. Return `estimator`, `model`, `score`

B. Setup Data Sets

1. Load the iris data set from the sklearn.datasets.load_iris() method.
2. Store observed data from iris.data in a variable named X.
3. Store labels for the observed data from iris.target in a variable named y.

   Note that X[0] is a vector whose label is y[0].

4. Explore the data. Determine how many classes exist, and how many observations exist within each class. Is the data balanced?
5. Create a test and training set. Note that split_test_train returns in this order: X_train, X_test, Y_train, Y_test. Note also that you should decide how large the test set should be. A typical train/test split is 70/30 or 80/20.

C. Perform Machine Learning

1. Call each function to perform the machine learning and create models based on each classifier. Collect the output in variables, keeping in mind that each function returns estimator, model, and score.
2. Plot the learning curve for each classifier. Call the plot_learning_curve method, passing the appropriate estimator, a title, and the X and y train variables.
3. Output the value of the testing score results for each model to the console.

D. Analyze Results

Use the information above to analyze the results. Include screenshots of your learning curves and include the average accuracy scores.

1. Which classifier performed best based on your train/test split? Why do you think it outperformed the others? Use the screenshots to justify your answer.
2. Try a different train/test split. Did this affect the results? If so, how? Record the screenshots for the new train/test split.

As a note, the train/test split will be different each time you run the program, which can affect results. Most often, people run several iterations to determine an overall average accuracy. You only need to run once here.

Submission

Submit your completed **lab10.py** file and your **Lab10.docx** file.

Submit to the Lab 10 dropbox at or before Monday, April 12, 2021 by 11:59 PM.

Grading

A letter grade will be assigned for each response. The letter grades are based on both correctness and the adequacy of answers. Points are assigned as follows:

		A	B	C	D	F	Zero
		Excellent	Above Average	Average	Below Average	Poor	No Attempt
		10	8	6	4	2	0
Complete the Code	Step A						
	Step B						
	Step C						
	Step D						
	Analysis						