Machine Learning : Using Scikit - Learn

Question 1 – Building a learning curve

The digits dataset consists of 1797 8x8 images. Each image, like the ones shown below, is of a hand-written digit. The objective is to build a model that will predict with a reasonable level of accuracy the numerical value depicted in a specific image. The objective of this exercise is to explore the use of learning curves in the context of this problem.

A selection from the 64-dimensional digits dataset

(i) Access the feature and label data from the digits dataset. The following code will allow you to import the digits dataset. It will also allow you to access the feature data (stored in X below) and the label data (stored in y below).

```
from sklearn import datasets

digits = datasets.load_digits()

digits = load_digits()

X = digits.data

y = digits.target
```

Part A.

(i) Create a DecisionTree classifier object. When creating the DecisionTree object you should specify the following parameter (max_depth=2).

(ii) Create a learning curve using this model on the digits dataset (specify the number of partitions on the X axis as 10).

(iii) Does the learning curve demonstrate a model that exhibits high bias or variance? How might you address this problem?

Part B.

(i) In the lab folder you will find a sample csv dataset. The final column in the dataset is the class that we want to predict. Read in the dataset and create a learning curve that takes in the feature and class data and a GaussianNB object.

(ii) What can you interpret from the learning curve? Is this model currently exhibiting high variance or high bias?