Samuel Mach

Nearest Neighbor Classification Using the Euclidean Distance – Using Python to Classify Handwritten Digits from the MNIST Data Set

Time estimations for full data set (10,000 test objects and 60,000 training objects)

K	Time
1	27 Minutes
3	32 Minutes
5	32 Minutes
7	31 Minutes

Full run results (K=1, N=60,000, M=10,000):

96.24% accuracy. 34:51 runtime, about 8 minutes longer than anticipated.

Accuracy for various parameters:

K	N (Training Set size)	Test Set size	Accuracy	time
1	10000	100	96%	2.7s
3	10000	100	93%	3.2s
5	10000	100	95%	3.2s
7	10000	100	94%	3.1s
1	10000	10000	93.5%	340s

Time estimates were obtained by using various values of K, keeping the training set size and test set size constant.

I used $1/6^{th}$ of the training set (N=10000) and $1/100^{th}$ of the test set (M=100). The run time in seconds multiplied by 600 was used as the estimate.

It should be noted that an additional 46ms time was spent pre-processing, to binarize the image data. This was across all 10,000 training images and 100 test images. If all 60,000 test images were preprocessed, this may take as much as 322ms.