CS 324 Data Mining **Problem Set 1**
Il Shan Ng **April 2, 2017**

Collaborators: I collaborated with Yuanshen Li on this assignment.

1. The k-nn classifier works on the assumption that if two images show the same digit, then large pixel values (indicating gray colors) are likely to occur in the same positions in both vectors. If two same digits are off-center in their respective images, then this assumption breaks down and so the computed distance metric for the two vectors may be large, indicating that the two images are dissimilar when they in fact show the same digit.

Another pre-processing step could be to convert the pixel values (0 to 255) into binary integers (0 or 1). This removes any noise introduced by the intensity of the pixels, and allows distance calculations to reflect only the difference in location of pixels within each image, which is what the algorithm uses to determine similarity. Implementation-wise, this re-mapping of pixel values can be done using a threshold; convert all values above 200 (very faint gray) to 0, representing a purely white pixel, and values 200 and below to 1, representing a purely black pixel.

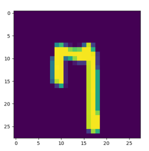
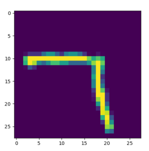
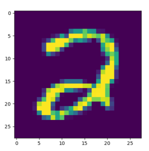
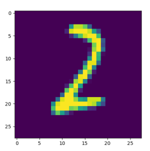
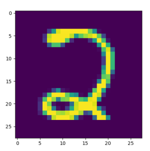
3. Using $k = 3$ with Euclidean distance gives an accuracy of 94.33% and confusion matrix

		Predicted classes		
		1	2	7
Actual classes	1	100	0	0
	2	4	90	6
	7	6	1	93

Using $k = 1$ with Manhattan distance gives an accuracy of 93.33% and confusion matrix

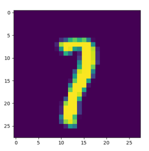
		Predicted classes		
		1	2	7
Actual classes	1	100	0	0
	2	10	85	5
	7	5	0	95

4. Below are five images that were classified correctly as 7, 7, 2, 2, and 2 respectively.

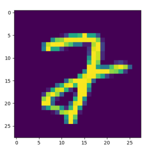






Below are five images that were classified incorrectly. The k-nn algorithm tends to do poorly for digits that were not written clearly and are ambiguous even to humans. For example, the second image below looks like both a 2 and a 7.

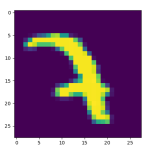
classified as 1; correct label is 7



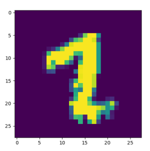
classified as 7; correct label is 2



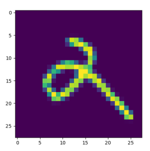
classified as 1; correct label is 7



classified as 7; correct label is 2

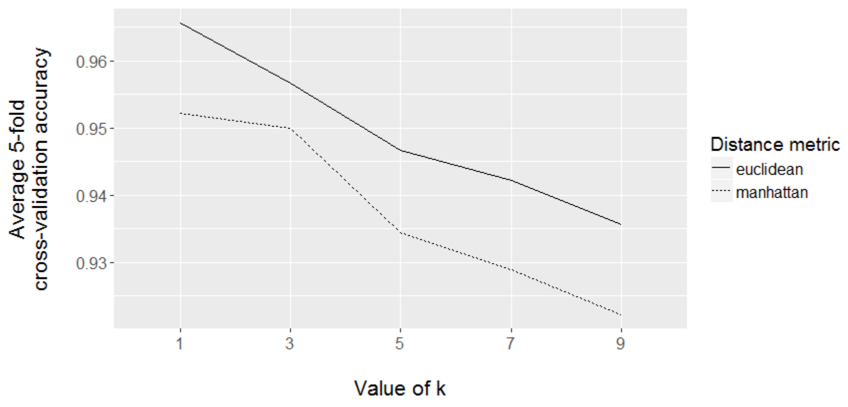


classified as 1; correct label is 2



5. The requested plot is shown below. For a given value of k, using Euclidean distance produces better accuracy than using Manhattan distance. For both distance metrics, using a larger value of k reduces accuracy. The most accurate classifier uses $k = 1$ and Euclidean distance. Its average 5-fold cross-validation accuracy is 96.55%, which is higher than the test accuracy, using the same classifier, of 95.67%.

Plot of cross-validation accuracy against k



6. With 300 test images and and an original training set of 900 images, the k-nn algorithm would have to compute 300×900 distances. With 300 test images and a new training set of 18,000 images, the algorithm now needs to compute 300×18000 distances. Computational cost increases by a factor of 20.

In the cross-validation case, with the original training set of 900 images, the knn-cv algorithm would have to compute $180 \times 720 \times 5$ distances. With 18,000 training images, the algorithm now needs to compute $14400 \times 3600 \times 5$ distances. Computational cost increases by a factor of 400.