Claire Sato Feb. 13, 2014
COMP135: Intro to Machine Learning

Assignment 4: Normalization, KNN, and Perceptron

Accuracy

KNN

	Sonar			
	Normalized		Unnormalized	
	Original	Noisy	Original	Noisy
k=1	55.00%	55.00%	55.00%	55.00%
k=3	84.00%	71.00%	77.00%	67.00%
k=5	81.00%	81.00%	77.00%	64.00%

	Vertebrate			
	Normalized		Unnormalized	
	Original	Noisy	Original	Noisy
k=1	68.00%	68.00%	68.00%	68.00%
k=3	75.00%	71.00%	79.00%	68.00%
k=5	77.00%	66.00%	83.00%	65.00%

Perceptron

	Sonar		Vertebrate	
	Original	Noisy	Original	Noisy
η = .01	63.00%	55.00%	76.00%	71.00%
η = .001	71.00%	63.00%	77.00%	74.00%

1. For KNN normalization helps more with the sonar datasets. When noisy features are added the accuracy for both the original and normalized is lower. This may be because KNN measures the euclidean distance between a test point and the training points, adding noise will create more dimensions that that may lead to an incorrect classification.

2. If we had to measure the performance of even values for k, we would have to write code to resolve ties. For example, if k = 2 and the two nearest points for the test point were classified as 1 and -1, we would have to compare the distance between the test point and neighbor 1 against that of the test point and neighbor 2. This comparison would require additional code.

3. The perceptron tests had more varied performance with sonar data. Both the original and noisy versions increased accuracy by ~7% from = .01 to = .001. The .001 learning rate performed better because it allowed the weights to increment by a smaller amount, enabling them to come within closer range of the actual desired weights.

4a. For the original vertebrate data, the KNN tests had slightly better performances than those of the perceptron tests. The highest accuracy was KNN for unnormalized data with 83% while the highest accuracy within the perceptron tests 76%. There was a lack of difference between the two because the original vertebrate data set does not, comparatively, have many features, it is only 37 dimensions. KNN has a worse performance with many features but under a certain threshold, the two algorithms will have roughly the same accuracy.

4b. With added noise features, the perceptron tests had more consistent and accurate results than the KNN tests. Under the perceptron tests for both = .01 and = .001 the accuracy was about 73%-75%. With added noise features, the KNN tests consistently had greater differences between original and noisy data. The KNN tests are weaker with higher dimensional data.

4c. For the sonar dataset without noise features, the KNN tests had a higher accuracy. The highest accuracy within regular data was 81% for KNN was K=5, normalized, while the highest within the perceptron tests was 71% with = .001. The KNN classifier was stronger because the test points in the sonar datasets have roughly little variance. Data points for the sonar are greater than 0 and less than 1, and this ensures that calculated distances will not be skewed by one feature having a much larger range. However, the lack of variance may work against perceptron classification because there would be even less difference amongst weights, creating a larger margin of error.

4d. For the sonar dataset, the KNN classifier was still more effective when noisy features were added. The greatest accuracy for the noisy sonar data within KNN was 81% with the normalized data, while the greatest for perceptron was 71%. However, KNN is not more effective when handling unnormalized data, which only had a max of 64%. This may indicate that normalization has a greater or equal effect on the accuracy of the sonar dataset than the noise features.

4e. The performance drop from noisy features is greater for the vertebrate dataset. Noisy features have less of an effect on sonar data because all the features, original and additional noise, had a range 0-1. However, the vertebrate data points had much greater variance, some values being <10 and some >100. Adding more dimensions of an already large variance has a stronger effect on the accuracy than adding dimensions of smaller variance. The proportion of noise added was also different between the two. The original sonar data had 60 features and the noisy file had 90, while the vertebrate data set originally had 7 features and it's noisy counter part had 30. While the noisy sonar dataset was 1.5x the size of the original, the noisy vertebrate dataset was more than 4x the size of the original.