K-Nearest Neighbors Analysis

Paul Laliberte' | CSCI-5622

1. What is the relationship between the number of training examples and accuracy?

We analyze the relationship intuitively and then in a more applied manner. The intuitive relationship is that the more samples of data available, the better we can train the KNN algorithm, the higher the accuracy. Banko and Brill give results supporting this claim (Figure 1). [1]

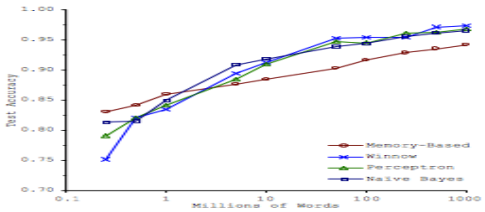

Figure 1: Banko and Brills Learning Curves

Limit	Accuracy
100	.670
300	.789
500	.831
1000	.876
2000	.909
5000	.940

Figure 2: Test Cases

Banko and Brill's result do not always hold true though. We run into issues of how "good" is the data that is being used, the variance of the models, and many other variables that are not constant over data sets.

We have arrived at an ambiguous conclusion thus far; more data is better, most of the time. To decide whether more data (training examples) is better for our algorithm we run several test cases. To do this we hold k constant, $k = 3$, for all runs, but we vary the limit (the amount of data to feed our algorithm). The results are in Figure 2. For our algorithm more data does seem to correlate with a higher accuracy.

2. What is the relationship between k and accuracy?

We will take an approach that is similar to problem 1. Intuitively, as k increases, the boundary decisions between the data becomes less distinct, and with a k too low, white noise has more of an influence (Figure 3). To mathematically prove an optimal k for the data set is beyond this analysis. However, we can simply observe how the accuracy varies with different k's in our algorithm. We will use a constant limit of 1000 and vary k. See Figure 4 for results.

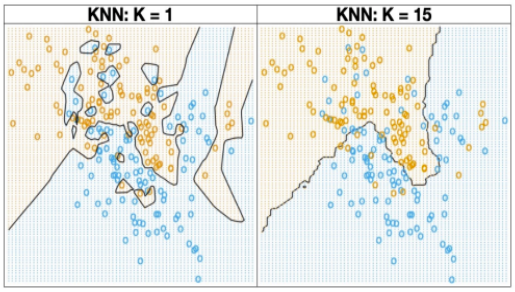

Figure 3: Decision Boundaries

k	Accuracy
1	.882
2	.786
3	.875
4	.843
5	.869
10	.846
20	.829

Figure 4: Test Cases

One note to make is that since we built our classifier from the same data we test on, a $k = 1$ should be higher. The only reason that the $k = 1$ does not result in a 100% accuracy is because we take the median if there is a tie. From our results a good choice for k would be either $k = 3$ or $k = 5$ (peaks), though these results may vary with more, and or less, data.

3. What numbers get confused with each other most easily?

To find which numbers get confused the easiest, we can look at the dictionary of dictionaries (or confusion matrix output). To get the incorrectly predicted numbers to stand out, we ran the algorithm with a limit of 10000 and a $k = 5$. We chose the the top one or two outliers per number (Figure 5). Some of the outliers were not "large" outliers, in that the KNN algorithm only incorrectly predicted a few cases (0 and 2 fall into this category).

Number	0	1	2	3	4	5	6	7	8	9
Missed	9	7,8	3	5,8	9	8	5	2,9	3	4

Figure 5: Test Cases

[1] Michele Banko, Eric Brill. 2001. "Scaling to Very Very Large Corpora for Natural Language Disambiguation." *ACM.*