Paul Johnston

CS 478

Nearest Neighbor

1. Nearest Neighbor model

This *k* nearest neighbor model takes a data set and trains on it by simply storing the data set along with the labels. It is able to predict on both continuous and nominal labels. It uses Euclidean distance to determine the k nearest neighbors. When finding the distance between two instances, it first finds the difference between each attribute of the instances. When comparing the differences of attribute values of two instances, if either of the instances has "unknown" for the attribute value, the distance is set to 1, because 1 is the greatest distance possible between two attribute values after normalizing. All input data is normalized when given to the model to train on and all test data is normalized as well. If the attribute is nominal and the values don't match, the difference is set to 1 as well. It then takes this modified difference vector and uses it to find the Euclidean distance between the two instances. After find the distance from a novel instance to all the training instances, it find the k nearest. After finding the k nearest it is able to find the majority class (if the output class is nominal) with or without weights, or, the average output value (if the output class is continuous) with or without weights.

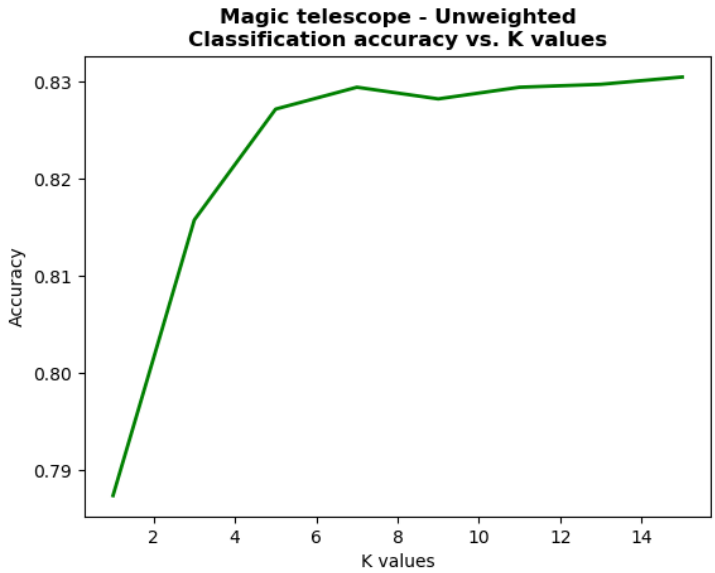
2. Magic telescope

The following data shows the model's classification accuracy on the magic telescope dataset (nominal output classes) without distance weighting, k=3, with normalization and without normalization.

	With normalization	Without normalization
Classification Accuracy	0.8157815781578158	0.8082808280828083

When the data is normalized prior to training and testing, the accuracy was slightly higher. This happens because attribute values that are inherently larger, but not necessarily more informative, affect the distance more than smaller valued attributes do, even if these smaller attribute value differences are more informative. The results above are consistent with this analysis in that normalizing the data resulted in higher accuracy when testing.

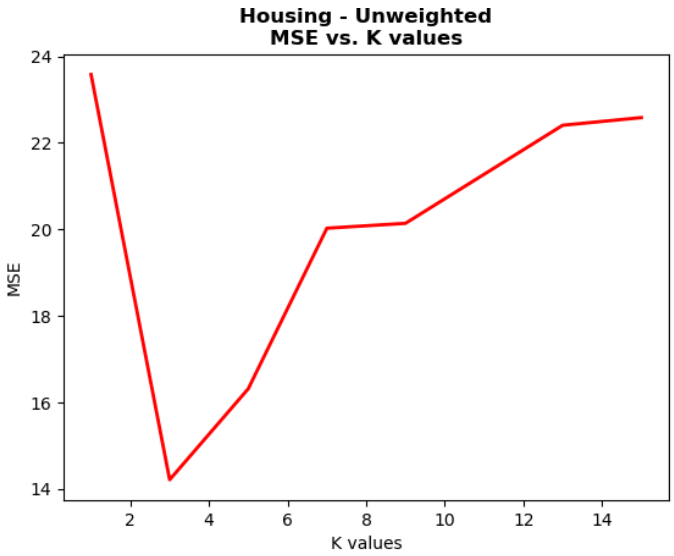
The following graph shows the classification accuracy of the model when using various values of k. The data was normalized, and no distance weighting was used.



The k value which produced the best result was 15 surprisingly, which resulted in an accuracy of 0.83. This surprises me because I thought that by consulting lots of neighbors, the overall accuracy would decrease because the prediction would be pulled away by noisy neighbors. My hypothesis is that the training data is accurate and since there is so much data, the more neighbors (to a certain point) is better.

3. Housing price

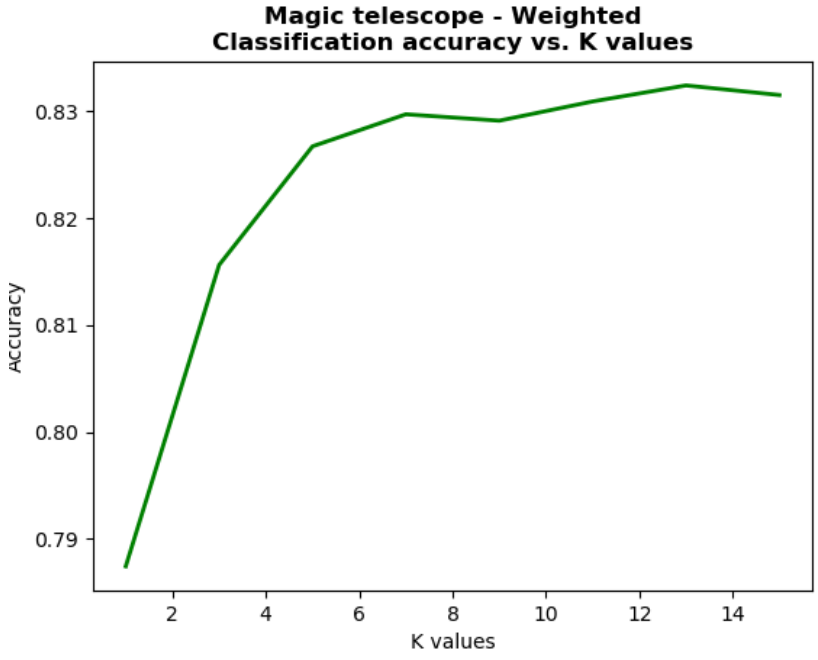
The following data was found by using the regression version of the model (continuous output class) with normalization and no distance weighting. Different values of k were tested. The output classes were not normalized.



The k value that resulted in the lowest MSE was 3, which resulted in an MSE of 14.215. This dataset is smaller in comparison to the magic telescope dataset which leads me to believe that this sudden increase in error as k increases can be attributed to the noisy/sparse training data and the fact that it's outputting continuous labels instead of just yes/no like the magic telescope data.

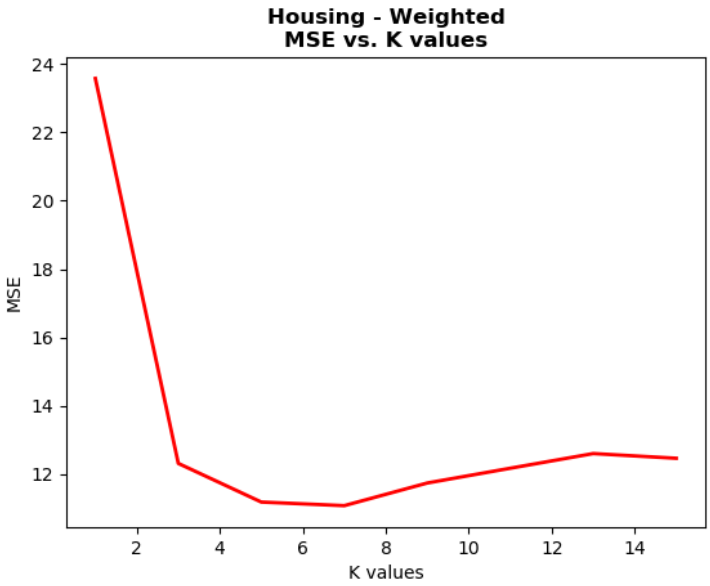
4. Distance weighting

The following data was produced using distance weighting (inverse of distance squared) with normalization on the magic telescope dataset. Different k values were tested, and the classification accuracies were as follows:



The k value which resulted in the highest classification accuracy was 13 resulting in an accuracy of 0.8324, slightly higher than before with no distance weighting. Interestingly, with k set to 15, the accuracy decreased slightly, which leads me to believe that with distance weighting, it is not getting as "lucky" with asking more neighbors what their outputs were, and maybe those outside neighbors would have increased the accuracy slightly, but their votes were less overall because they were far away.

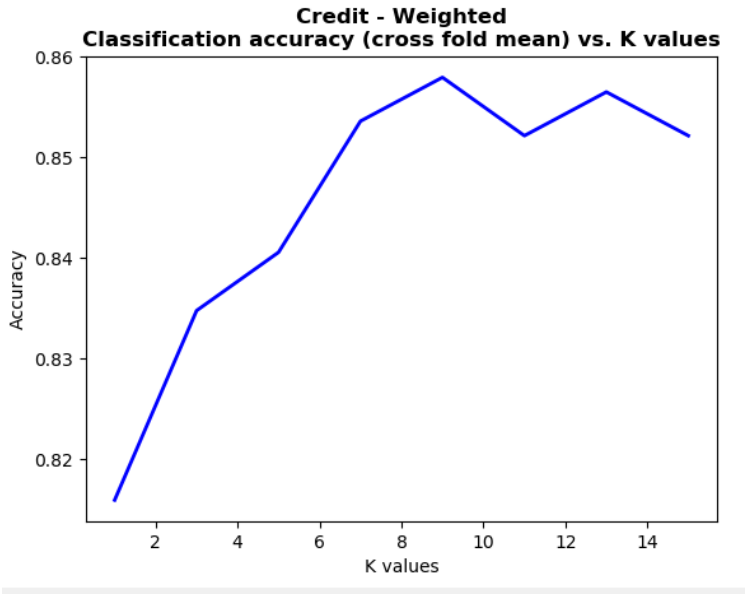
The following data was attained by training on the Housing dataset with normalization, distance weighting, and varying values of k. The accuracy is given in terms of MSE.



The effect of distance weighting is seen in the graph. As k increased, i.e. more neighbors were consulted on their outputs, the MSE only increased slightly, because these neighbors who were far away from the instance had a lower vote overall because of their distance. The lowest MSE was attained with a k value of 7, resulting in an MSE of 11.07. This was lower than when not using distance weighting and it also had the added perk of not skyrocketing in MSE when k increased.

5. Credit approval

For this dataset, I tested different values of k and used normalization and distance weighting. When calculating the distance between two instances, I first checked if either was unknown, if so, then the distance was set to 1, the highest distance possible for normalized data. If the attributes were nominal and didn't match, I also set the distance to 1. When they were nominal and matched, the distance was 0, and when they were continuous, the distance was calculated by using the Euclidean distance formula. This distance metric is justified because when you have an unknown attribute value, nothing can be assumed. It may be close to the average of that attribute or could be a noisy instance, to be safe however, setting the distance to 1 allows for this unknown to not have much say in what the output is of the instance. Similarly with nominal data, it is difficult to know just how close one nominal value is to another, so to be safe, they are treated as completely different values, hence a distance of 1. Below is a graph displaying the classification accuracy using cross fold validation with 10 folds and different k values.



The highest accuracy of 0.858 was with a k value of 9. This is a fairly good accuracy, the dataset is large and allows for more consistency when asking lots of neighbors. After asking too many neighbors however the accuracy began to decline, leading me to believe that the data, like all data, can only represent a portion of the population.

6. Experiment

I didn't do an experiment