Julian Benitez Mages
DS 2500
Prof. Laney Strange
4 April 2021

Homework 5 Write-up: kNN Classifiers

Assignment Interpretation: Create your own classifier from scratch, cycling through each instance of data. Each instance will become a singular test, while all the others will be used as training data (k nearest neighbors) for the test instance. Report back the accuracy of your home-grown classifier and use sci-kit learn metrics to generate the precision, recall, and f1 score for each k value. Include visualizations to show the effects of a different k value on the metrics.

Data Used: Results from a happiness survey in Somerville.

https://archive.ics.uci.edu/ml/datasets/Somerville+Happiness+Survey

My Process:

Out of 6 questions, I chose two to use as attributes to try to predict the answer to the final question: are you happy or unhappy? Because this target value is binary (0 for no and 1 for yes), I was able to generate a confusion matrix for each k, showing true as well as false positives and negatives.

The chosen questions were X1: “the availability of information about the city services,” and X4: “your trust in the local police,” with the results to each question as integers ranging from zero to five. Using Euclidian distance, I was able to find the k closest results for each instance when used as the test set. I found the majority target result of the k nearest neighbors to compare it to the actual target of the test instance, which contributed to the accuracy of the entire dataset for each k. I created additional functions to count up true and false positives and negatives, and created a heatmap for the confusion matrix for each k. I also counted up all of the predicted and actual results, and plugged these into the sci-kit metrics functions to calculate precision, accuracy and f-1 score for each k. Finally, I graphed all of the metrics results as a function of k, to see the effect of different k values on the metrics.

Results:

As seen in the first visual, a k value of 7 leads to the best accuracy. The second visual is the confusion matrix for k = 7. With the confusion matrix, the x axis indicates actual values and the y axis indicates predicted values.