Lloyd Beaufils, Jerrry Bonnell, Randall Naar, Gururaj Shriram

ECE-648

18 October 2017

Project 1 - Group 5

The Effect of Irrelevant Attributes on k-NN Classifiers

The k-NN classifier is a version of the nearest neighbor classifier that looks at k neighbors as opposed to simply one. For any given example, the classifier finds its k nearest neighbors and chooses a class label based upon those of those neighbors. Since this classifier makes use of geometric distances to other examples when deciding class labels, it is essential that those examples be relevant. Each attribute contributes equally to the distance, so attributes that should have no bearing on the class label can quickly overpower the relevant attributes and lead to improper classification. With the irrelevant attributes adding essentially random values to and dominating the distance, the classifier degenerates until it does little better than label classes at random. This behavior is dependent upon the ratio of relevant attributes to irrelevant attributes. The higher that ratio, the less of a negative impact the irrelevant attributes will have. The more irrelevant attributes, then the more terms negatively impacting the distance formula and the worse the classifier will perform, on average. Our group has run some tests to obtain evidence of these claims, and in this paper, we will outline our findings.

For this project, DEBUG: talk about 2 datasets & 2 classifiers

IRIS DATASET: regular classification, then classification with irrelevant attributes (and then classification with more relevant attributes?)

GURU DATASET: same as above, fill up a page but also explain how the data was generated