APPENDIX VI

Fractional K-nearest neighbors: Dealing with non-uniform sampling

Summary

This appendix briefly describes a novel variation of the well-known k-nearest neighbor (KNN) metric for classification. The new variation is often more useful than its predecessor when the number of exemplars of each class is not the same.

The problem

KNN (Duda, Hart and Storm, 2000) assigns a data point X to the cluster or class with the greatest number of points among the k-nearest neighbors of X. When our sample has a different number of points for each class, and when the number of points we have for each class does not reflect the actual density or probability of each class in the underlying distribution, but rather derives from sampling biases, for example, then the outcome of KNN will be biased toward classes with the largest number of samples, an undesired effect.

Fractional K-Nearest Neighbors

This problem is solved by simply assigning a point X, not to the class with the greatest *number* of points among its k-nearest neighbors, but rather to the class with the greatest *fraction of its members* among X's k-nearest neighbors.

Acknowledgment

Thanks to Pietro Perona for a discussion on this subject.