Improved kNN Rule for Small Training Sets
Sunsern Cheamanunkul and Yoav Freund
{scsheaman, yfreund}@eng.ucsd.edu
Department of Computer Science and Engineering
University of California, San Diego

Introduction
The traditional k-NN classification rule predicts a label based on the most common label of the k nearest neighbors, namely the plurality rule. It is known that the plurality rule is optimal when the number of examples tends to infinity [1,2]. However, real-world datasets are always finite and often small. In such cases, there might be other rules that might perform significantly better than the plurality rule.

Our approach is motivated by the fact that the plurality rule ignores information present in the counts of labels other than the most common label. Consider the following example from letter recognition task. The most common label in the neighborhood is either 'h' or 'm'. However, when we take into account the training data, we find that h's and m's rarely appear in the same neighborhood while k's are often found with both h's and m's. Therefore, in this case, 'k' is a better prediction than the more common labels 'm' and 'h'.

Minimizing KL-Divergence Rule

Training
- Require: Training set $S$ and $k$
- Output: The center distributions $Q_j$ for all $j \in Y$
  1. $Q_j \leftarrow \phi$ for $j \in Y$
  2. for each example $(x, y) \in S$
  3. $Q_j \leftarrow Q_j + p(x, y)$
  4. end for
  5. $Q_j \leftarrow Q_j / |S|$ for all $j \in Y$

Prediction
- Require: Training set $S$, a test example $x$
- The center distributions $Q_j$ for all $j \in Y$
- Output: Predicted label $\hat{y}$
  1. $\hat{y} = \arg \min_{j \in Y} D_{KL}(p(x, y) | Q_j)$

Experiments

(A) SYN1 (10 classes, 1600 examples)

Overlapping normal distributions
Error rate of MinKL and Majority when n = 1600 and k is varied.
Error rate of MinKL and Majority when n is varied and k is fixed by CV.

(B) uRight (26 classes, 9945 examples)

Error rate of letter recognition task using 5-NN with MinKL and Majority rule.
Examples misclassified by Majority but not MinKL.

(C) MNIST[3] and SVHN[4]

MNIST (10 classes, 60000 examples)
SVHN (10 classes, 73257 examples)

Error rate as a function of neighborhood size (k), comparing MinKL and Majority.

Conclusion and Future Work
In this work, we provided a justification for the minimizing KL-divergence rule in a finite sample setting under certain assumptions about the data. Additionally we showed experimentally that the minimizing KL-divergence rule is able to outperform the majority rule on various synthetic datasets and real-world datasets. Future directions include experimenting on more real-world datasets and investigating more complex models for representing the center distributions.

References