Improved kNN Rule for Small Training Sets

Sunsern Cheamanunkul and Yoav Freund

Department of Computer Science and Engineering
University of California — San Diego

ICMLA, Detroit, MI
Dec 3-5, 2014
K-nearest neighbors

- Classify using **the majority vote** among $K$ closest examples in the training set.
K-nearest neighbors

• When $N \to \infty$, KNN classifier is Bayes optimal when choosing $K$ s.t. $K \to \infty$ and $K/N \to 0$ [Fix and Hodges, 1951].

• What about when $N$ is small?
  • Limited number of labeled examples.
  • Theoretical results give us little guidance.
Issue with the majority vote

• The majority vote only considers the most common class in the neighborhood and discards the rest.

• “The rest” can give us information too!
Toy example

5-nearest neighbors
Prediction?
Minimizing KL-divergence rule

**Training:**
- For each class, compute a **center distribution** (a.k.a average neighborhood).

**Prediction:**
- Predict a class such that the KL divergence from the empirical neighborhood distribution to the class’ center distribution is minimized.
Minimizing KL-divergence rule

![Graph showing center distributions for 'h', 'k', and 'm' with KL divergence values]

Center distribution for ‘h’

Center distribution for ‘k’

Center distribution for ‘m’

\[ D_{KL} = 494 \]

\[ D_{KL} = 381 \]

\[ D_{KL} = 519 \]

Label distribution of 5 nearest neighbors
Theoretical analysis

• Let $E_1, E_2, \ldots, E_m$ be sets of center distributions such that $\bigcap E_i = \emptyset$

• Suppose each sample $x^k$ is generated by the following process:
  1. Class $i^*$ is chosen with probability $\pi_i$
  2. A distribution $p$ is chosen s.t. $\Pr\{p \in E_{i^*}\} \geq 1 - \delta$
  3. $x^k$ is sampled IID from $p$. 
Theoretical analysis

• Let $\hat{p}$ denote the empirical distribution induced by $x^k$.

• We show that

$$\Pr\{D_{KL}(\hat{p}||E_{i*}) > \min_{j \neq i*} D_{KL}(\hat{p}||E_j) \mid p \in E_{i*}\} \leq (m - 1)(k + 1)^m e^{-k\Delta}$$

where

$$\Delta \doteq \min_{i,j; j \neq i} \min_{q \in \mathcal{P}} \max(D_{KL}(q||E_j), D_{KL}(q||E_i))$$
Experiments: synthetic data

- # of classes: 10
- # of examples: up to 1600
Experiments: synthetic data

- # of classes: 64
- # of examples: up to 2560
Handwritten letter recognition

- Number of classes: 26
- Number of examples: ~520 / user
- Number of users: 15
MNIST and SVHN

**MNIST**
- # of classes: 10
- # of examples: 60000

**SVHN**
- # of classes: 10
- # of examples: 73257
Conclusion

• We introduced the minimizing KL-divergence rule for KNN that consider the entire neighborhood distribution rather than just the majority.

• We provided a theoretical justification of the rule under certain data generation assumptions.

• We demonstrated the benefit of the MinKL rule over the majority vote rule on synthetic datasets and real-world datasets.

• Future directions include:
  - experimenting on more datasets.
  - investigating more complex model for representing center distributions.
Thank you!