Complete Cross-Validation for Nearest Neighbor Classifiers

Matthew Mullin MDM@WHIZBANG.COM
Whizbang! Labs, 4616 Henry Street, Pittsburgh, PA 15213 USA

Rahul Sukthankar RAHULS@CS.CMU.EDU
Compaq CRL, Cambridge, MA, 02142
The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA

In supervised learning (Mitchell, 1997), a pool of labeled data, S, is used to predict the labels of unseen data. Since the distribution of data is unknown, the generalization accuracy (the probability that an unseen data point will be correctly classified) cannot be directly determined. Using S, we can calculate an empirical accuracy and use this as an estimate of the generalization accuracy.

Two accepted techniques for estimating the generalization accuracy are *subsampling* and *k-fold cross-validation*. In the former, S is partitioned into a training set, T, and a test set $S \backslash T$. The empirical accuracy is given by the fraction of test set items labeled correctly by a classifier trained on T. In the latter, S is partitioned into k equally-sized subsets. Each subset is used as a test set for a classifier trained on the remaining $k - 1$ subsets. The empirical accuracy is given by the average of the accuracies of these k classifiers. Both techniques may employ a stratified partitioning in which the subsets contain approximately the same proportion of classes as S.

The empirical accuracy obtained using either of these techniques depends on the items from S that appeared in a given subset, and is therefore a random variable (different for each technique), and any given trial of subsampling or k-fold cross-validation is equivalent to a single observation. The best estimate of the generalization accuracy is given by the expectation of this random variable, termed *complete cross-validation* (Kohavi, 1995). However, directly calculating this expectation requires averaging over all permissible partitions of S and is generally impractical since the number of such partitions grows exponentially with $|S|$. Therefore, it is customary to create an estimate by averaging the accuracies from a manageable number of partitions (Mitchell, 1997). Unfortunately, the variance of the random variable can be significant and obtaining an estimate with sufficiently low variance can require a large number of time-consuming trials.

We present a technique for calculating the complete cross-validation (CCV) for the nearest neighbor family of classifiers (including the popular K-nearest-neighbor algorithm). Nearest neighbor methods (Dasarathy, 1991) frequently appear at the core of sophisticated pattern recognition and information retrieval systems. There are examples in face recognition, image retrieval and text retrieval.

We present formulae for

- 1-nearest neighbor
- stratified 1-nearest neighbor
- K-nearest neighbor for 2 class problems
- stratified K-nearest neighbor
- 1-nearest neighbor with R-rank-scoring

with complexity analysis and a brief statistical analysis.

References

Dasarathy, B. (1991). *Nearest neighbor pattern classification techniques*. Los Alamitos, CA: IEEE Computer Society Press.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence* (pp. 1137–1143).

Mitchell, T. (1997). *Machine learning*. New York: McGraw-Hill.