The Cross Validation Problem

John Langford (jl@hunch.net)

TTI-Chicago

1 The Method

K-fold cross validation is a commonly used technique which takes a set of m examples and partitions them into K equal-size sets (folds) of size m/K. For each set, a classifier is trained on the other sets.

2 The Problem

Assume only that samples are drawn iid from a distribution D on classification example $X \times \{0,1\}$ where X is a feature and $\{0,1\}$ is a label. Derive a classifier from the K classifiers with a lower true error rate bound than other approaches (discussed in section 3). The bound should have the form: For all learning algorithms A, number of sets K,

$$\Pr_{S \sim D^m}(e(D,h) \leq f(S,h,h_1,...,h_K)) \geq 1-\delta$$

where h is the classifier derived from the cross validation process on m examples producing classifiers $h_1,...,h_K$, f is the bound (a computable function), and $e(D,h) = \Pr_{(x,y)\sim D}(h(x) \neq y)$ is the true error rate of h.

3 Past Work

1. Devroye, Rogers, and Wagner analyzed cross validation and found algorithm specific bounds for nearest neighbor and space partitioning algorithms. Much of this is documented here [1].
2. Michael Kearns and Dana Ron [2] analyzed cross validation and found that under additional stability assumptions the bound for the classifier which learns on all the data is not much worse than for a test set of size m/K.
3. Avrim Blum, Adam Kalai, and John Langford [3] analyzed cross validation and found that you can do at least as well as a test set of size m/K using the randomized classifier which draws uniformly from the K learned classifiers.

4. Yoshua Bengio and Yves Grandvalet [4] analyzed cross validation and concluded that there was no unbiased estimator of variance.
5. Matti Kaariainen [5] noted that you can safely derandomize a stochastic classifier (such as one that randomizes over the K sets) using unlabeled data without additional assumptions.

4 Some Extreme Cases to Sharpen Intuition

1. Suppose on every set the learned classifier is the same. Then, the cross-validation error should behave something like a test set of size m. This is radically superior to a test set of size m/K. Behavior like this is often observed in practice.
2. Consider leave-one-out cross validation. Suppose we have a "learning" algorithm that uses the classification rule "always predict the parity of the labels on the training set". Suppose the learning problem is defined by a distribution which picks y=1 with probability 0.5. Then, with probability 0.5, all leave-one-out errors will be 0 and otherwise 1 (like a single coin flip). This example suggests that incorporating information about the stochastic difference between the classifier may yield a tighter bound.

5 Impact

On any individual problem, solving this might have only have a small impact due to slightly improved judgement of success. But, because cross validation is used very extensively, the overall impact of a good solution might be very significant.

References

1. Luc Devroye, Laszlo Gyorfi, and Gabor Lugosi, "A Proabibilistic Theory of Pattern Recognition", Springer, 1996.
2. Michael Kearns and Dana Ron, "Algorithmic Stability and Sanity-Check Bounds for Leave-One-Out Cross-Validation", Neural Computation 11(6), pages 1427-1453. 1999.
3. Avrim Blum, Adam Kalai, and John Langford, "Beating the Holdout: Bounds for KFold and Progressive Cross-Validation", COLT 1999, pages 203-208.
4. Yoshua Bengio and Yves Grandvalet "No unbiased estimator of the variance of K-fold cross-validation", Journal of Machine Learning Research, 5, 1089-1105, 2004.
5. Matti Kaariainen, "Generalization Error Bounds Using Unlabeled Data", COLT 2005.