Extended abstract for NIPS*2000 Workshop on "Cross-Validation, Bootstrap and Model Selection"

Empirical Model Comparison: Bayesian Analysis of Disjoint Cross Validation for Continuous Losses

Carl Edward Rasmussen & Pedro A. d. F. R. Højen-Sørensen
Department of Mathematical Modelling, Building 321,
Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark
carl,phs@imm.dtu.dk

The ability to form reliable estimates of generalization error based on limited samples of data is central to research in statistical modeling. As pointed out in the taxonomy of Diettrich [1998] it is important to distinguish between the 1) application oriented and 2) algorithm oriented evaluation problems. Whereas in the application oriented problem we are concerned exclusively with estimating the *average* performance, we have to estimate both the *averages and uncertainties* in the algorithm oriented setting (since it is crucial to know whether empirically measured performance differences are statistically significant).

Due to the limited data at hand, methods relying on resampling are often used to estimate generalization errors; however the reuse of data makes reliable estimation of the required uncertainties difficult. We suggest a new scheme based on n-way cross validation with the modification that one trains only on a single subset at a time and tests on the remaining $n-1$ subsets in each rotation. This has the advantage that the training sets are non-overlapping, avoiding the need to hypothesize about the correlations between models trained on similar datasets. The overlapping test sets are easily handled since test errors are measured on a case by case basis. We model the observed losses (or difference in losses between two methods corresponding to a *paired test*) by an additive model:

$$y_{ic} = \mu + a_i + b_c + \epsilon_{ic},$$

where y_{ic} is the loss, a_i is the training set effect, b_c is the test case effect and ϵ_{ic} is a noise term; index $1 \leq i \leq n$ refers to training subset and $1 \leq c \leq N$ to test case number. Under Gaussian assumptions for the training set effects, test set effects and noise, with hyperparameters controlling their magnitude and the correlation between training and test set effects (since the same data is used to estimate these) we can efficiently integrate out the effects, compute the posterior over the hyperparameters and sample from it with a simple Markov Chain. Sampling is done using the Metropolis procedure which converges rapidly on this problem (less that 2 seconds of CPU time – the running time is (close to) independent of the total number of cases N and number of subsets n).

This model can be seen as an extension to the DELVE [Rasmussen et al, 1996] scheme for comparing learning algorithms, where data is recycled compared to the original scheme. We demonstrate that the new model is more powerful in terms of finding small differences than DELVE. We also discuss the implications of the Gaussian assumptions and the effect of violations of the assumptions (eg. of outliers).

T. G. Diettrich (1998) Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1924.

C. E. Rasmussen, R. M. Neal, G. E. Hinton, D. van Camp, M. Revow, Z. Ghahramani, R. Kustra & R. Tibshirani (1996) DELVE – Data for Evaluating Learning in Valid Experiments, http://www.cs.utoronto.ca/~delve.