STATISTICAL HORIZONS

Machine Learning

Stephen Vardeman, Ph.D.

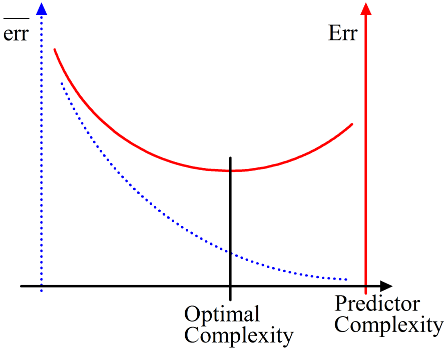
Upcoming Seminar:
June 8-9, 2017, Philadelphia, Pennsylvania

The following pages are a random assortment of slides from the 20 modules of the course.

Performance and Complexity in Supervised Learning

Training Error and Err

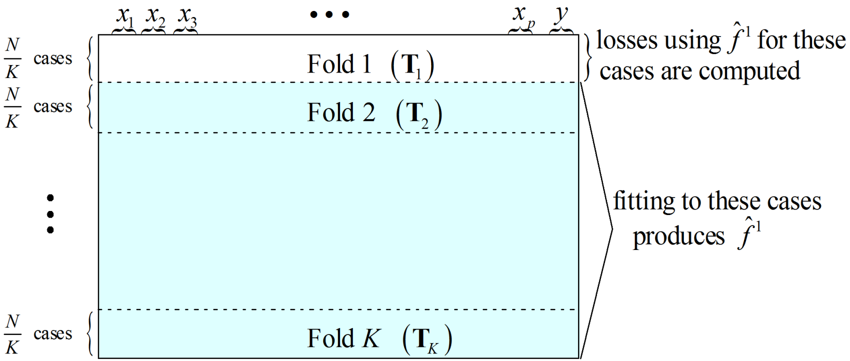
The cartoon below illustrates the general issue faced in choosing a predictor based on training data. $\overline{\text{err}}$ decreases with increased complexity ("low bias" in SEL problems) while Err decreases and then increases. One must try to somehow find a predictor with approximately optimal complexity (e.g. in light of the "variance-bias trade-off" in SEL problems).



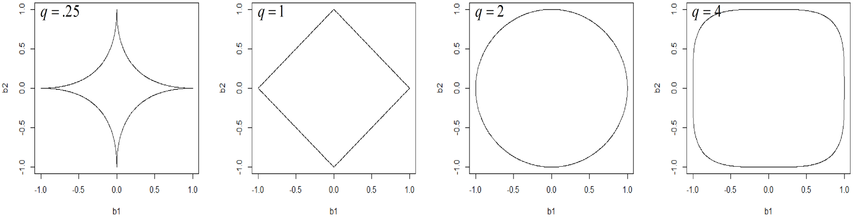
Predicting Predictor Performance

Cross-Validation

Below is a graphic suggesting roughly how (after putting the N cases into a random order) one breaks $\mathbf{T}$ into folds and computes the part of sum defining $CV\left(\hat{f}\right)$ for cases in the first fold. (Of course, slightly different pictures are needed for the sums from the other $K-1$ folds.)



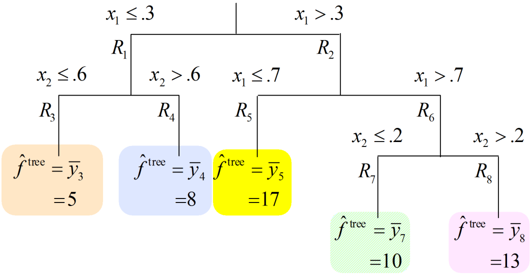
For comparison purposes, here are representations of $p = 2$ bridge regression constraint regions for $t = 1$. For $q < 1$ the regions not only have "corners," but are not convex.

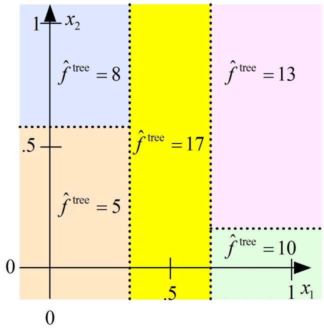


Regression Trees

Representations of the Small p=2 Hypothetical Case

Below are two representations of the hypothetical $p = 2$ tree predictor.

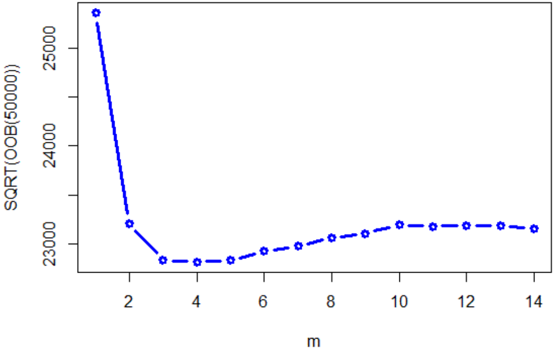




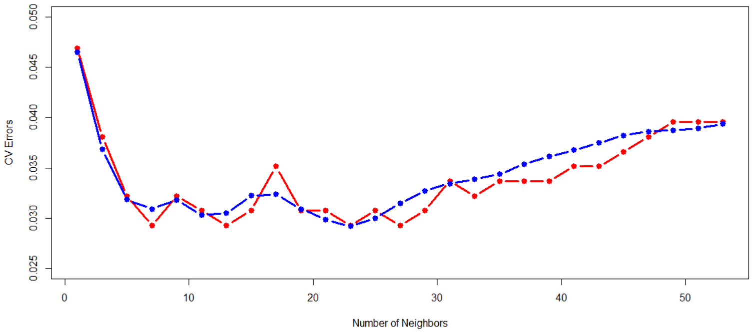
Random Forests

Ames House Price Example: Choice of m

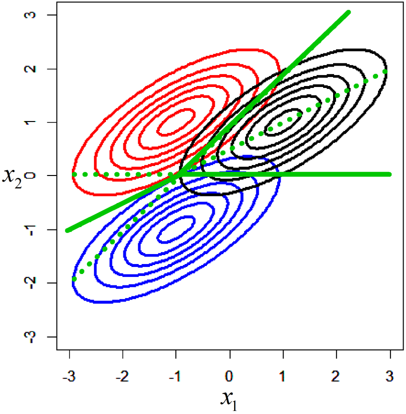
The `randomForest` package was used to fit random forests to the Ames House Price data for $m = 1, 2, \ldots, 14$ (with all other parameters at their default values). A plot of the square root of the OOB error based on $B = 50000$ trees is below. The best value of m is 4 with $\sqrt{\text{OOB}\,(50000)} \approx 22813$.



A more careful cross validation exercise done (with the `tune` routine in the `caret` package and 100 repeats of 10-fold cross validation) *restandardizing after removing each fold* produces essentially the same conclusions about k in this problem. This is evident in the plot below of both the earlier LOOCV error (in red) and the more carefully made CV error (in blue).

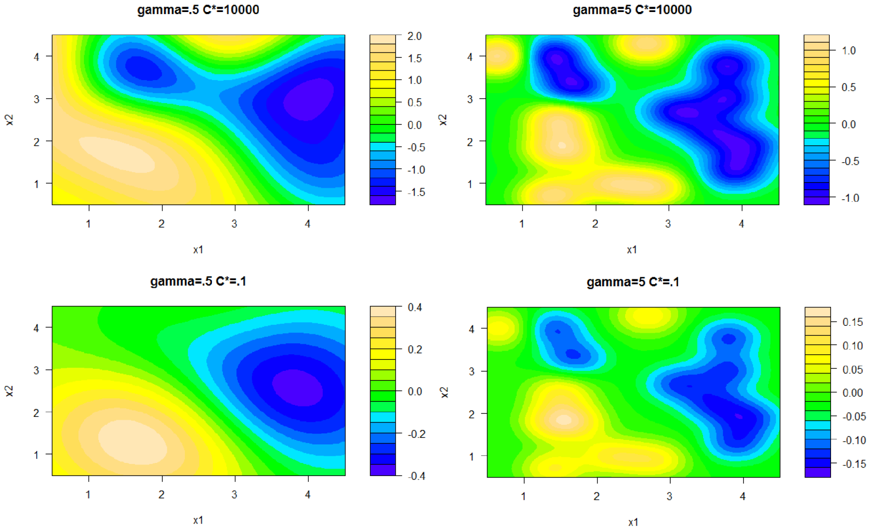


Below are contour plots for $K = 3$ bivariate normal densities with a common covariance matrix and the linear classification boundaries corresponding to equal class probabilities $\pi_1 = \pi_2 = \pi_3$.

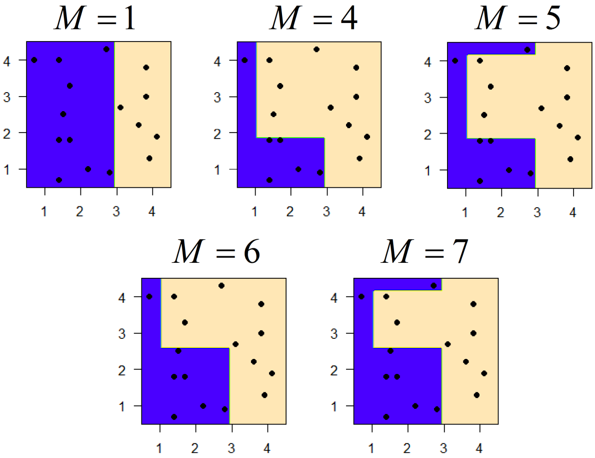


Support Vector Machines (SVMs)

A Toy p=2 Example



Below are graphics portraying the 5 different classfiers met in the development of the 0 training error rate $M = 7$ AdaBoostM.1 classfier.



Associated with the principal component directions are so-called "singular values." These are non-negative values d_l that decrease as the index on the principal component direction $\mathbf{v}_l$ increases. The sum of the squared entries of the L-dimensional approximation to $\mathbf{X}$ in display (1) is $\sum_{l=1}^{L} d_l^2$. So these can be thought of as related to the portion of the variance of the (centered) values in $\mathbf{X}$ accounted for by the first L principal components.

As it turns out, the principal component directions are also so-called eigenvectors of the $p \times p$ matrix of inner products for columns of $\mathbf{X}$ and the singular values are the square roots of the so-called eigenvalues of that matrix. (Those are in turn a multiple of the eigenvalues for the sample covariance or correlation matrices for $\mathbf{X}$). Principal components analysis is then sometimes based not on the singular value decomposition of $\mathbf{X}$, but on an eigen analysis of a covariance or correlation matrix.

Partitioning and Hierarchical Methods

Another Toy Example

Below are plots of $p = 2$ data pairs. The left indicates (by both color and symbol) how the pairs were generated from 6 bivariate normal distributions. The center indicates the result of 6-means clustering. The right shows complete linkage agglomerative clustering cut at 6 clusters. The latter 2 are clearly different.