Statistics 840 Lecture 21©G. Wahba 2011

Recap, more on tuning for prediction and variable selection.

1. Regularization Class of Statistical Methods, cost functionals, penalty functionals.
2. Ridge Regression, Penalized Least Squares, Relation to Bayes methods.
3. Geometry and inner products based on positive definite functions
4. Reproducing Kernel Hilbert Spaces, Bounded linear functionals and the Riesz representation theorem.
5. First and Second variational problems, the (Kimeldorf-wahba) representer theorem.
6. Univariate cubic and higher order splines, thin plate splines, splines on the sphere, Smoothing Spline ANOVA(SS-ANOVA)

7. Choosing the smoothing parameter, the leaving-out-one lemma.
8. Unbiassed Risk, GML, GCV, GACV, AIC, BGACV, BIC. Degrees of freedom for signal. train-tune-test, 10-fold cross valication
9. The degrees of freedom for signal, the randomized trace method.
10. Radial basis functions, Gaussian, Matern
11. Properties of GCV. Convergence rates for smoothing splines tuned by GCV
12. Data from exponential families, Bernoulli, Poisson. The Comparative Kullback-Distance criteria for tuning, as a generalization of least squares.
13. Bayesian "Confidence Intervals".
14. Standard and Non Standard Support Vector Machines (SVM's)

SVM's and penalized likelihood for Bernoulli data compared. Optimal classification.

15. The Multicategory SVM

16. The LASSO, the LASSO-Patternsearch Algorithm, Variable and pattern selection. GACV and BGACV for Bernoulli responses.

17. Early stopping as a regularization method.

18. Regularized Kernel Estimation (RKE), Regularized Manifolding Unfolding (RMU)

19. Using difference data (pedigrees) with Spline ANOVA models using RKE

Degrees of Freedom for Signal

Recall that the degrees of freedom for penalized likelihood estimation for Gaussian data with a quadratic norm penalty functional generalizes the ordinary parametric notion of degrees of freedom.

Parametric least squares regression:

$$y_{n \times p} = X\beta + \epsilon$$

where $\epsilon \sim N(0, \sigma^2 I)$ and X is of full column rank. Find β to

$$min\|y - X\beta\|^2, \ \ \hat{\beta} = X(X^T X)^{-1} X^T y.$$

$\hat{y}$, the predicted y is given by

$$\hat{y} = X\hat{\beta} = Ay$$

where

$$A = X(X^TX)^{-1}X^T.$$

A (the ”influence matrix”) is an orthogonal projectiom onto the p dimensional columm space of X with trace p. Notice that

$$\frac{d\hat{y}}{dy} = a_{ii},$$

the iith entry of A.

Nonparametric Regression

Let $y_i = f(x_i) + \epsilon_i, i = 1, \cdots, n$ or, more compactly,

$$y = f + \epsilon$$

where $f \in \mathcal{H}_K$, an RKHS with RK K, and some domain $\mathcal{X}$. Find $f \in \mathcal{H}_K$, to

$$min\|y - f\|^2 + \lambda\|f\|^2_{\mathcal{H}_K}.$$

Letting f_λ be the minimizer, then $\hat{y} \equiv f_\lambda$ depends linearly on y and has the property that

$$f_\lambda = A(\lambda)y.$$

$A(\lambda)$ is known as the influence matrix, note that

$$\frac{d\hat{y}_i}{dy_i} = a_{ii},$$

the iith entry of A. $A(\lambda)$ is a smoother matrix, that is, it is symmetric non-negative definite with all its eigenvalues in $[0, 1]$.

Trace $A(\lambda)$ was called the Equivalent Degrees of Freedom for signal in `wahba.ci.83.pdf` p 139, (1983). by analogy with regression.

Methods for Choosing λ

The unbiased risk estimate (UBR). Need to know σ^2, the variance of the Gaussian noise. Choose λ to min

$$U(\lambda) = \|(I - A(\lambda))y\|^2 + 2\sigma^2 tr(A(\lambda)).$$

The expected value of $U(\lambda)$ is, up to a constant, an unbiased estimate of $\|f_\lambda - f\|^2$.

The generalized cross validation estimate (GCV). Do not need to know σ^2. Signal to noise ratio needs to satisfy some conditions. Choose λ to min

$$V(\lambda) = \frac{\|(I - A(\lambda))y\|^2}{[tr(I - A(\lambda))]^2}$$

drived from a leaving-out-one argument. Optimality properties have been discussed and some references are in lect7.

Here's a very rough intuitive argument why it works. Suppose that $\frac{1}{n}trA(\lambda)$ is small. Noting that the power series expansion of $(1-\rho)^{-2}$ for small ρ is $(1+2\rho+...)$ gives, for $\rho=\frac{1}{n}traceA(\lambda)$ sufficiently small

$$n^2V(\lambda) \approx (\|((I-A(\lambda))y\|^2(1+\frac{2}{n}tr(A(\lambda)+...$$

and so

$$n^2V(\lambda) \approx \|(I-A(\lambda)y\|^2+2\left(\frac{1}{n}\|(I-A(\lambda))y\|^2\right)tr(A(\lambda).$$

(See Charles Stein Ann. Stat 1981, Ker-Chau Li, Ann. Stat 1985)

Linear or Non Linear Estimates

Early Stopping via the Conjugate Gradient Algorithm, other nonlinear estimates, for example, $y_i = \int G(x_i, t, g(t))dt + \epsilon_i$, where it is desired to estimate g. For notational convenience, let $f_\lambda \equiv \hat{y}$ be the estimate of f, however it is obtained. Here f is the 'true' prediction, $f_i = \int G(x_i, t, g(t))dt$. The relationship between tuning for g and tuning for f in the linear case is discussed in `wahba.parzen09.pdf` and references cited there. A simulation of a meteorological data fitting problem where the fit to a time-dependent vector of observations has to approximately solve a partial differential equation is given in `gong.wahba.johnson.tribbia.mwr.pdf`.

Efron `efron.covariance.04.pdf`, Stein (1981), Mallows (1973) result in the following:

$$E\|f - f_\lambda\|^2 = E\|y - f_\lambda\|^2 - n\sigma^2 + 2\sum_{i=1}^{n} cov(f_{\lambda,i}, y_i).$$

Under some mild conditions on the nonlinearity of the estimate, Stein's Lemma (or, more precisely one of Stein's lemmas) says:

Stein's Lemma: Let $A(\lambda, y) = \{\frac{\partial f_{\lambda,i}}{\partial y_j}\}$. Then

$$\sum_{i=1}^{n} cov(f_{\lambda,i}, y_i) = \sigma^2 EtrA(\lambda, y).$$

If the estimate is linear in y then $A(\lambda, y)$ does not depend on y.

This leads to the remarkable result

$$E\|f - f_\lambda\|^2 = E\|y - f_\lambda\|^2 + 2\sigma^2 Etr A(\lambda, y) - n\sigma^2,$$

that is, the expected value of the difference between the true f and the fitted f_λ is, up to a constant, $2\sigma^2$ times expected value of the degrees of freedom even if the estimate is (mildly) nonlinear in the data!

The Randomized Trace Estimate

The randomized trace method may be used to estimate the expected value of the degrees of freedom for signal when $\hat{y}$ depends linearly or (mildly) nonlinearly on the data. Let ξ be an iid random vector with mean zero and component variances σ_ξ^2. Let f_λ^z be the estimate with data z. Then, if f_λ depends linearly on the data,

$$E\frac{1}{\sigma_\xi^2}\xi^T(f_\lambda^{y+\xi} - f_\lambda^y) = \frac{1}{\sigma_\xi^2}E\xi^T A(\lambda)\xi = trA(\lambda)$$

Note that even if $\hat{y}$ depends (mildly) nonlinearly on y, the left hand side above could be considered a divided difference approximation to $\sum_{i=1}^n \frac{\partial f_{\lambda,i}}{\partial y_i}$ which is exactly what you want for $\hat{df}$, according to Stein's Lemma.

The randomized trace estimate was proposed by Hutchinson and Girard independently in 1979, Girard later proved theorems about its properties in Ann. Statist. Note that the same ξ should be used as λ varies. If f_λ depends linearly on y, then the result does not depend on σ_ξ^2. If the relationship is not linear, then the value of σ_ξ^2 can make a difference. It can be shown, given that the variance of the components of ξ are the same, that centered Bernoulli random components for ξ are more efficient, than a Gaussian ξ.

Tuning for Exponential Families:

The general form of the probability density for exponential families with no nuisance parameter is of the form negative log likelihood $= -yf + b(f)$ where f is the so-called canonical link. $\frac{\partial b(f)}{\partial f}$ is the mean and $\frac{\partial^2 b(f)}{\partial^2 f}$ is the variance of the distribution for exponential families. See McCullough and Nelder's (1989) book. A popular criteria for tuning members of the exponential family is the Comparative Kullback- Liebler (CKL) distance of the estimate from the true distribution.

The Kullback-Liebler distance is not a real distance, and is not even symmetric, but it is defined for most tuning purposes as

$$KL(\hat{F}, F) = E_F log \frac{(F, y)}{(\hat{F}, y)}$$

where $\hat{F}$ and F are the two densities to be compared and the expectation is taken with respect to F. The CKL is the KL with terms not dependent on λ deleted, and is

$$\sum_{i=1}^{n} -Ey_i f(x_i) + b(f(x_i)).$$

The residual sum of squares for the Gaussian distribution $N(0, I)$ is an example of the CKL.

The two most common cases are the Bernoulli distribution, where f is the log odds ratio and $b(f) = log(1 + e^f)$, and the Poisson distribution, where $f = log\Lambda$ and $b(f) = e^f$. Λ is mean of the Poisson distribution. The Poisson distribution has an exact unbiased risk estimate for f, `wong.bickelvol.loss.pdf`, with the CKL criteria, but it involves solving an optimization problem solved n times. Readily computable approximations can be found in `yuan.poisson.pdf`.

For the Bernoulli distribution, it is known that no unbiased risk estimate is possible. The Generalized Approximate Cross Validation (GACV) estimate `xiang.wahba.sinica.pdf`. is an approximte unbiased risk estimate for the CKL for the Bernoilli distributon. Letting $OBS(\lambda)$ be the observed (sample) CKL

$$OBS(\lambda) \sum_{i=1}^{n} -y_i f_\lambda(x_i) + b(f_\lambda(x_i).$$

the GACV becomes

$$GACV(\lambda) = OBS(\lambda) + \frac{\sum_{i=1}^{n} y_i(y_i - p_\lambda(x_i))}{tr(I - W^{1/2} H W^{1/2})} trH$$

where p_λ is the fitted mean, W is the diagonal matrix with diagonal entries the fitted variances and H is the inverse hessian of the optimization problem and plays the role of the influence matrix.

Tuning the LASSO with Gaussian Data.

Let $y_i = x(i)^T \beta + \epsilon_i, i = 1, ..., n$, where ϵ is $N(0, \sigma^2 I)$ and β is a p dimensional vector of possible coefficients, and $x(i)$ is the ith design vector. The LASSO finds β to min

$$\|y - X\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

The LASSO penalty, also known as the ℓ_1 penalty is known to give a sparse solution, that is, depending on λ, the larger the λ the fewer non-zero βs will appear in the solution.

Zou, Hastie and Tibshirani `zou.hastie.tibshirani.lasso09.pdf` show that in Gaussian LASSO setting, the appropriate choice of degrees of freedom for signal is the number of non-zero basis functions, leading to the unbiased risk-type estimate for λ as the minimizer of

$$U_{LASSO}(\lambda) = \|y - X\hat{\beta}\|^2 + 2\sigma^2 \hat{df}$$

where $\hat{df}$ is the number of non zero βs. See `zou.hastie.tibshirani.lasso09.pdf` for details. If $\frac{\hat{df}}{n}$ is small and σ^2 is unknown, then GCV can be used.

Tuning the LASSO with Bernoulli Data

The LASSO with Bernoulli data can be tuned with the GACV, see `shi.wahba.wright.lee.08.pdf, tr1166.pdf`. The GACV becomes

$$GACV(\lambda) = OBS(\lambda) + \frac{\sum_{i=1}^{n} y_i(y_i - p_\lambda(x_i))}{n - N_{\beta_0}} trH$$

where N_{β_0} is the number of non-zero coefficients in the model.

Tuning the LASSO: Prediction vs Variable Selection.

All of the tuning methods so far have been based on a prediction criteria, either the residual sum of squares or the CKL. When the LASSO is used, typically it is believed that the model is sparse, that is, if a tentative model is given as $f(x) = \sum_{j=1}^{p} c_j B_j(x)$, for large p and some basis functions B_j, and some c, the number of true non-zero c's is believed to be small. The difference between tuning for prediction and tuning for sparsity can be seen by comparing the tuning criteria AIC and BIC in their simplest forms: (see Wikipedia and references there) $AIC = -2loglikelihood + 2k$ and $BIC = -2loglikelihood + klogn$, where k is the number of terms in the model, according to the descriptions given in Wikipedia.

For the present argument, think of k as the degrees of freedom. Then AIC is essentially a UBR method, while BIC was proposed as a variable selection method. To get from AIC to BIC you just replace $2k$ by $klogn$. BIC ("Bayesian Information Criteria") was proposed by Schwartz by assuming that $p < n$ and that all of the p coefficients were *a priori* equally likely to appear. For $logn > 2$ BIC will give a model that is no larger than, and generally smaller than AIC. For tuning the LASSO with Bernoulli data, GACV was replaced by BGACV, by the replacement suggested above. However, simulation experience has shown that when $p >> n$, BGACV is not strong enough to return a small model when one is warranted. See `shi.wahba.wright.lee.08.pdf, tr1166.pdf`. When the true model is small, prediction criteria will tend to give models that include the true model but are bigger. Open questions remain as to appropriate tuning for variable selection.

Finally, Tuning the Support Vector Machine

Often the SVM is applied to very large data sets, then the luxury of dividing the observational data set into train, tune and test subsets can be carried out. Ten fold cross validation is also popular. Regarding internal tuning, a GACV based method for the SVM can be found in `tr1039.pdf`, a similar method was earlier given by Thorsten Joachims, called the ξ/α method.

Have a Happy Thanksgiving!