Statistics 191: Introduction to Applied Statistics

Bias-Variance tradeoff: penalized techniques

Jonathan Taylor
Department of Statistics
Stanford University

March 10, 2010

Penalized regression

Topics

- Bias-Variance tradeoff.
- Penalized regression.
- Cross-validation.

Penalized regression

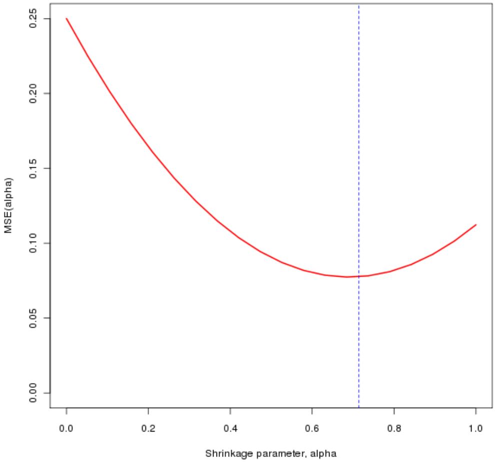
Jonathan Taylor
Department of Statistics
Stanford University

Bias-variance tradeoff

- Arguably, the goal of a regression analysis is to "build" a model that predicts well.
- What does "predict well" mean?

$$
\begin{aligned}
MSE_{pop}(\mathcal{M}) &= \mathbb{E}\left((Y_{new} - \widehat{Y}_{new,\mathcal{M}})^2\right) \\
&= \mathrm{Var}(Y) + \mathrm{Var}(\widehat{Y}_{new,\mathcal{M}}) + \\
&\qquad \mathrm{Bias}(\widehat{Y}_{new,\mathcal{M}})^2.
\end{aligned}
$$

Bias-Variance Tradeoff



Penalized regression

Jonathan Taylor
Department of Statistics
Stanford University

Shrinkage & Penalties

- Shrinkage can be thought of as "constrained" minimization.
- Minimize

$$\sum_{i=1}^n (Y_i - \mu)^2 \quad \text{subject to } \mu^2 \leq C$$

- Lagrange: equivalent to minimizing

$$\sum_{i=1}^n (Y_i - \mu)^2 + \lambda \mu^2$$

for some $\lambda = \lambda_C$

Penalized regression

Shrinkage & Penalties

- Differentiating:

$$-2\sum_{i=1}^{n}(Y_i - \widehat{\lambda}_C) + 2\lambda\widehat{\mu}_\lambda = 0$$

- Solving

$$\widehat{\mu}_\lambda = \frac{\sum_{i=1}^n Y_i}{n+\lambda} = \frac{n}{n+\lambda}\overline{Y}.$$

- As $\lambda \to 0$,

$$\widehat{\mu}_\lambda \to \overline{Y}.$$

- As $\lambda \to \infty$

$$\widehat{\mu}_\lambda \to 0.$$

Penalized regression

Jonathan Taylor
Department of Statistics
Stanford University

Penalties & Priors

- Minimizing

$$\sum_{i=1}^n (Y_i - \mu)^2 + \lambda \mu^2$$

is similar to computing "MLE" of μ if the likelihood was proportional to

$$\exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n (Y_i - \mu)^2 + \lambda \mu^2\right)\right).$$

- If $\lambda = m$, an integer, then $\widehat{\mu}_\lambda$ is the sample mean of $(Y_1, \dots, Y_n, 0, \dots, 0) \in \mathbb{R}^{n+m}$.
- This is equivalent to adding some data with $Y = 0$ ...

Biased regression: penalties

- Not all biased models are better – we need a way to find "good" biased models.
- Inference (F, χ^2 tests, etc) is not quite exact for biased models.
- Generalized one sample problem: penalize large values of β. This should lead to "multivariate" shrinkage of the vector β.
- Heuristically, "large β" is interpreted as "complex model". Goal is really to penalize "complex" models, i.e. Occam's razor.
- If truth really is complex, this may not work! (But, it will then be hard to build a good model anyways ... (statistical lore))

How much to shrink, choosing λ

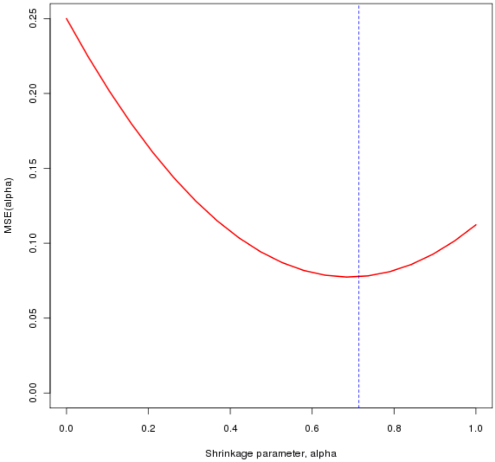
- In our one-sample example,

$$\begin{aligned} MSE_{pop}(\lambda) &= \mathrm{Var}(\lambda \bar{Y}) + \mathrm{Bias}(\lambda \bar{Y})^2 \\ &= \frac{\lambda^2 \sigma^2}{n} + \mu^2 (1-\lambda)^2 \end{aligned}$$

- Differentiating:

$$\begin{aligned} 0 &= -2\mu^2(1-\lambda^*) + 2\frac{\lambda^* \sigma^2}{n} \\ \lambda^* &= \frac{\mu^2}{\mu^2 + \sigma^2/n} \end{aligned}$$

Bias-Variance Tradeoff



Penalized regression

Ridge regression

- Assume that columns $(X_j)_{1 \leq j \leq p-1}$ have zero mean, and length 1 and Y has zero mean.
- This is called the *standardized model*.
- Minimize

$$SSE_\lambda(\beta) = \sum_{i=1}^n \left(Y_i - \sum_{j=1}^{p-1} X_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p-1} \beta_j^2.$$

- Corresponds (through Lagrange multiplier) to a quadratic constraint on β's.

Penalized regression

Solving the normal equations

- Normal equations

$$\frac{\partial}{\partial \beta_l} SSE_{\lambda}(\beta) = -2\langle Y - X\beta, X_l \rangle + 2\lambda \beta_l$$

-

$$-2\langle Y - X\widehat{\beta}_{\lambda}, X_l \rangle + 2\lambda \widehat{\beta}_{l,\lambda} = 0, \qquad 1 \leq l \leq p-1$$

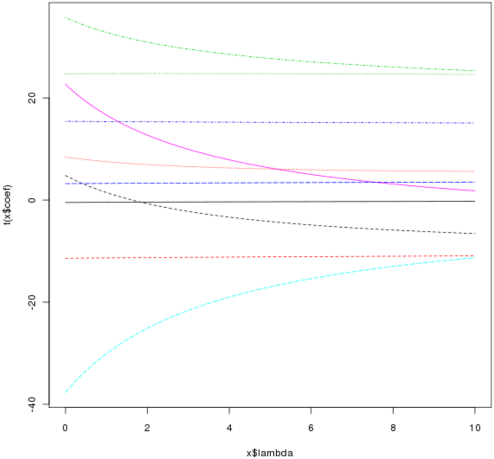
- In matrix form

$$-Y^tX + \widehat{\beta}_{\lambda}^t(X^tX + \lambda I) = 0$$

- Or

$$\widehat{\beta}_{\lambda} = (X^tX + \lambda I)^{-1}X^tY.$$

Ridge regression



Penalized regression

LASSO regression

- Another popular penalized regression technique.
- Use the standardized model.
- Minimize

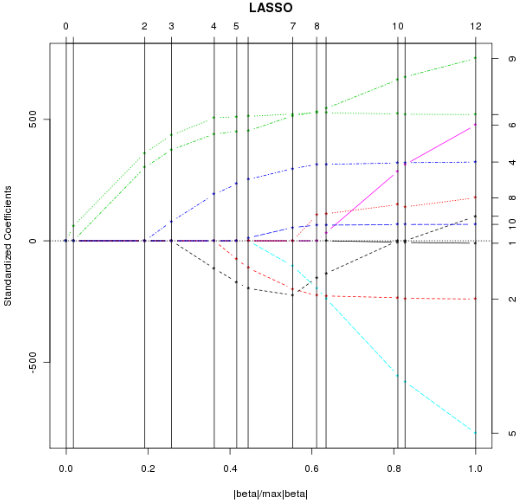
$$SSE_{\lambda}(\beta) = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p-1} X_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

- Corresponds (through Lagrange multiplier) to an ℓ^1 constraint on β's. In theory, it works well when many β_j's are 0 and gives "sparse" solutions unlike ridge.

LASSO

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University



Penalized regression

Jonathan Taylor
Department of Statistics
Stanford University

Choosing λ: cross-validation

- If we knew *MSE* as a function of λ then we would simply choose the λ that minimizes *MSE*.
- To do this, we need to estimate *MSE*.
- A popular method is "cross-validation." Breaks the data up into smaller groups and uses part of the data to predict the rest.
- We saw this in diagnostics: i.e. Cook's distance measured the fit with and without each point in the data set.

Penalized regression

K-fold cross-validation

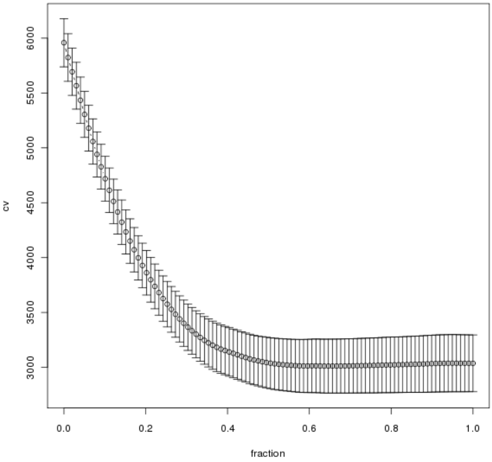
- Fix a model (i.e. fix λ). Break data set into K approximately equal sized groups $(G_1, \dots, G_K)$.
- `for (i in 1:K)` Use all groups except G_i to fit model, predict outcome in group G_i based on this model $\widehat{Y}_{j(i),\lambda}, j \in G_i$.
- Estimate

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^{K} \sum_{j \in G_i} (Y_j - \widehat{Y}_{j(i),\lambda})^2.$$

K-fold cross-validation (continued)

- It is a general principle that can be used in other situations, not just for Ridge.
- Pros (partial list): "objective" measure of a model.
- Cons (partial list): inference is, strictly speaking, "out the window" (also true for other model selection procedures in theory).
- If goal is not really inference about certain specific parameters, it is a reasonable way to compare models.

LASSO: cross-validation



Penalized regression

Generalized Cross Validation

- A computational shortcut for n-fold cross-validation (also known as leave-one out cross-validation).
- Let
$$S_\lambda = (X^tX + \lambda I)^{-1}X^t$$
be the matrix in ridge regression.
- Then
$$GCV(\lambda) = \frac{\|Y - S_\lambda Y\|^2}{n - \mathrm{Tr}(S_\lambda)}.$$
- The quantity $\mathrm{Tr}(S_\lambda)$ is the *effective degrees of freedom*.

Ridge regression