COS 424: Interacting with Data

Lecturer: Dave Blei
Scribe: Josh Herbach, Lindsay Gorman

Lecture #17
April 8, 2008

1 Previously on 424

Ridge regression (an adjustment to linear regression which introduces bias in exchange for a reduction in variance) has the effect of constraining the parameters to lie within a hypersphere centered at the origin. The size of this sphere is based on λ which is usually chosen by cross-validation.

By constraining the parameters to lie within the hypersphere, we produce a model with smaller coefficients. This model is then potentially simpler and more interpretable. This model is also potentially better because if the reduction in variance is significant enough, we can reduce test error (see previous lecture for more on the bias-variance trade-off, and the parameters that determine test error).

2 An Aside on Bayesian Statistics

2.1 Background

In bayesian statistics as in the linear models we have discussed, $y_n \sim F(\theta)$ (where θ denotes the set of all parameters). Unlike the models we have been discussing, in the bayesian setting, θ is also a random variable, and $\theta \sim G_0(\alpha)$ where G_0 is the prior distribution of θ and the parameter to that prior distribution α is called the 'hyperparameter.' In this context, to determine θ we perform posterior inference $p(\theta|y_{1:N}, \alpha)$.

Some of the possible strategies for posterior inference are:

MAP (maximum a posteriori estimation): $\theta^{MAP} = \arg\max_\theta p(\theta|y_{1:N}, \alpha)$

Mean (posterior mean estimation): $\theta^{mean} = E[\theta|y_{1:N}, \alpha]$

Note that the above strategies are different from MLE in that they include the prior.

In general, there is still some lingering animosity from old men in frequentist statistics who contest the validity of the Bayesian approach, but as we'll see with ridge regression and its Bayesian interpretation, Bayesians and non-Bayesians are often doing similar things. Here Professor Blei notes his red-eye flight and resulting "loopiness."

2.2 Connection to Ridge Regression

Ridge regression is a form of MAP in which $\beta_i \sim gsn(0, 1/\lambda)$ and $y_n|x_n, \beta \sim gsn\left(\beta^\top x_n, \sigma^2\right)$.

The MAP estimate of β:

$$\begin{aligned}
\arg\max_\beta p(\beta|y_{1:N}, x_{1:N}, \lambda) &= \arg\max_\beta \log p(\beta|y_{1:N}, x_{1:N}, \lambda) \\
&= \arg\max_\beta \log p(\beta, y_{1:N}|x_{1:N}, \lambda) \\
&= \arg\max_\beta \log \left(p(y_{1:N}|x_{1:N}, \beta, \lambda) \prod_{i=1}^P p(\beta_i|x_{1:N}, \lambda)\right) \\
&= \arg\max_\beta \log p(y_{1:N}|x_{1:N}, \beta) + \sum_{i=1}^P \log p(\beta_i|\lambda) \\
&= \arg\max_\beta -RSS(\beta; y_{1:N}, x_{1:N}) - \sum_{i=1}^P \lambda\beta_i^2
\end{aligned}$$

Note that the final equality above arises from two facts. $p(\beta_i|\lambda)$ is $gsn(0, 1/\lambda)$ so the log of it is a constant independant of β and $\lambda\beta_i^2$. Also, $\log p(y_{1:N}|x_{1:N}, \beta)$ is just the likelihood and maximizing with respect to β means this is equivalent to the MLE and as was established previously, the MLE is equivalent to maximizing -RSS. Thus this final equality is precisely the objective function for ridge regression we found earlier in a non-Bayesian context.

2.3 Summary

Bayesian statistics give an alternative way to think about the trade off between λ and bias as a tradeoff between the prior idea of what the data should be and how it falls. This information is encapsulated in the choice of hyperparameter λ which controls how far away from the MLE the estimate will be. For small λ, there is larger variance, and the MLE will be chosen (i.e. the data itself will totally determine the estimate), whereas as λ gets larger, the prior $E[\beta] = 0$ becomes more influential, introducing more bias by moving the estimate further from the MLE. A philosophical aside: ideally, the prior selected should reflect some 'prior' knowledge one has about the parameters; however, this is not always the case in practice (otherwise it wouldn't be philosophical, now would it).

This restriction of parameters to be in the space around the origin and deviation from the true $\hat{\beta}$ is also referred to as 'shrinkage'. Some applications of this technique are in movie recommendation and gene regulation.

3 The Lasso

or: 20 minutes of digression and diagrams and this is all we get?

Like ridge regression, the lasso is a regularization method. It is properly pronounced Lass-oh (as in the scottish 'lass', and the exclamation 'OH', not to be confused with la-sue). It is a relatively new method (developed in the mid 90's) and is quite powerful. It attempts to optimize the same function as ridge regression ($\sum_{n=1}^{N} \frac{1}{2}\left(y_n - \beta^\top x_n\right)$). Unlike ridge regression, which is subject to the constraint $\sum_{i=1}^{P} \beta_i^2 \leq s$, it is subject to the constraint $\sum_{i=1}^{P} |\beta_i| \leq s$. Whereas ridge regression kept β locked within a sphere around the origin, the lasso keeps β within a hypercube with its diagonals axis aligned. There exists some $s' > 0$ such that for all $s < s'$ the point of minimal error on the hypercube will be at one of its corners. [1] This means that for small enough s, only one of the coefficients β_i will be non-zero. Thus, the lasso acts as a form of feature selection, choosing the most important features for the given s (as s grows more parameters may become non-zero).

The lasso is equivalent to $\hat{\beta}^{lasso} = \arg\min_\beta \sum_{n=1}^{N} \frac{1}{2}(y_n - \beta x_n)^2 + \lambda \sum_{i=1}^{P} |\beta_i|$. Like ridge regression there is a 1-1 mapping between λ and s, and like ridge regression, this object is convex. "'This is really exciting."' (Blei, lecture, April 8, 2008). Before the lasso was invented, the best method to find a sparse set of features was 'subset selection.' Sadly, subset selection was computationally expensive. As a result, the prospect of simply performing a convex optimization to find a sparse solution was (and still is) a great boon.[2]

[1] There is one small exception to this statement. If the parameters β_i are perfectly correlated, then there is no $s > 0$ for which the minimum error point on the lasso will be on a corner. In general this is a moot case because if one has perfectly correlated parameters, then all but one should be thrown out since they offer no information.

[2] The lasso isn't perfect and it will not explore all possible subsets of features, but it seems to work pretty darn well. At least the yokels at Google plus folk like Professor Blei think so.

With ridge regression, λ was chosen using cross-validation. With the lasso, there is an alternative algorithm (the LARS algorithm[3]) which can eficiently explore the space of λ.

Like ridge regression, the lasso also has a bayesian interpretation. In particular, lass regression corresponds to MAP estimation with the model $\beta_i \sim \text{Laplace}(\lambda)$ and $y_n|x_n, \beta \sim gsn\left(\beta^\top x_n, \sigma^2\right)$. The coefficients come from a Laplace distribution $p(\beta_i|\lambda) = \frac{1}{2}exp\{-\lambda\,|\beta_i|\}$

3.1 Nathan's e-mail

When the diamond is fairly near the MLE, the intersection of the diamond with the "best" contour will be a point of tangency. Plotting all points closest to the origin for each contour will yield a straight line if the error is the RSS, as the contours will be a circle or ellipse. If the diamond intersects this line, the best estimate will lie on this line. If not, the diamond is smaller than the intercept of this line, and the best point will be at one the diamond's extreme points. Then the beta estimate will be a piecewise linear segment from the origin, along one axis, then to the line above, and to the MLE. The degenerate case described corresponds to the case where this line passes through the origin, reducing the problem to one parameter.

3.2 Professor Blei's e-mail

In two dimensions, if the covariates are perfectly correlated, the lasso will not find a sparse solution, but as there is perfect correlation, there is really only one covariate and the lasso makes the right choice. In higher dimensions, this issue arises when all the covariates are perfectly correlated, a very bizarre setting. In general, if two covariates are correlated, one is redundant and is thrown away.

4 Generalized Regulation

Ridge regression and the lasso (and subset selection and others) can all be considered instances of generalized regularization. In general, regularization can be seen as minimizing the RSS with a constraint on a q-norm: $\|\beta\|_q \leq s$. Subset selection corresponds to $q = 0$, the lasso corresponds to $q = 1$, ridge regression to $q = 2$, and MLE to $q = \infty$). All of the different possible choices of q are akin to different bayesian solutions (different priors) and each offers a different bias-variance trade-off (q controls the shape of the region and larger q explore more area for a given λ). If $0 \leq q \leq 1$, the regulation provides sparsity in addition to the usual bias-variance trade-off. The only value of q that offers both sparsity and a convex optimization problem is 1 (the lasso).

[3]Efron et al. 2004