Gov 2002 - Section 2:
The Bayesian Lasso

Connor Jerzak

September 21, 2016

Harvard University, Department of Government

Warmup!

Joint Distributions

Often, we are interested in two or more random variables defined on the same sample space. The distribution of these variables is called a joint distribution. Joint distributions can be made up of any combination of discrete and continuous random variables.

- **Joint Probability Distribution**: If both X and Y are random variable, their joint probability mass/density function assigns probabilities to each pair of outcomes
  - Discrete: $p(x, y) = \Pr(X = x, Y = y)$
    s.t. $p(x, y) \in [0, 1]$ and $\sum \sum p(x, y) = 1$
  - Continuous: $f(x, y); \Pr((X, Y) \in A) = \iint_A f(x, y) dxdy$
    s.t. $f(x, y) \geq 0$ and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dxdy = 1$
  - If X and Y are independent, then $P(X = x, Y = y) = P(X = x)P(Y = y)$ and $f(x, y) = f(x)f(y)$

Marginal & Conditional Distributions

- **Marginal Probability Distribution**: probability distribution of only one of the two variables (ignoring information about the other variable), we can obtain the marginal distribution by summing/integrating across the variable that we don't care about:

  Discrete: $p_X(x) = \sum_i p(x, y_i)$
  Continuous: $f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy$

- **Conditional Probability Distribution**: probability distribution for one variable, holding the other variable fixed. Recalling from the previous session that $\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$, we can write the conditional distribution as

  Discrete: $p_{Y|X}(y|x) = \frac{p(x,y)}{p_X(x)}, \quad p_X(x) > 0$
  Continuous: $f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}, \quad f_X(x) > 0$

Bayesian Statistics, recap

In the Bayesian approach to statistics, parameters are viewed as random variables to reflect our uncertainty about them.

- The prior for a parameter is its distribution before observing data.
- The likelihood is the distribution of the data, given the parameters.
- The posterior is the distribution for the parameter after observing data.

The Bayesian Lasso

Motivation for Bayesian Statistics

The Bayesian Lasso attempts to solve 2 related problems.

1. **The model selection problem**. Say I have p covariates. Which of those covariates should I include in a regression model for Y?

   > Example: I might have 1000 documents, and I want to identify a good statistical model relating the word combinations to the category the documents belong to. In most text datasets, there are likely 1000s of word combinations I could use to try to predict category membership. Which combinations should I choose?

2 **The overfitting problem**. Say I have p covariates. How do I prevent my statistical model from describing noise in my data, as opposed to genuine relationships?

- Example: Let's go back to the earlier example. If I have 1000s of word combinations at my disposal, I might be running statistical models with so many covariates that I might actually perfectly predict some categories in my training set. However, it is likely that my model is picking up not only on genuine patterns but also randomness. By chance alone, some of the word combinations might be related to category membership. I will want to "shrink" my regression coefficients to take into account this possibility.

Probabalistic Setup

Lasso estimates can be interpreted as posterior mode estimates when the regression parameters have independent and identical Laplace (i.e. double-exponential) priors.

- $y|X, \beta, \sigma^2 \sim N(X\beta, \sigma^2)$.
- Put a Laplace prior on β_i:

$$\beta_i|\sigma^2 \sim \frac{\lambda}{2\sigma} \exp(-\lambda|\beta_i|/\sigma)$$

We can rewrite the conditional distribution of β_i as a hierarchical mixture:

$$\beta_i|\sigma^2, \gamma_i^2 \sim N(0, \sigma^2\gamma_i^2)$$
$$\gamma_i^2|\sigma^2 \sim \mathrm{Expo}(\lambda^2/2)$$

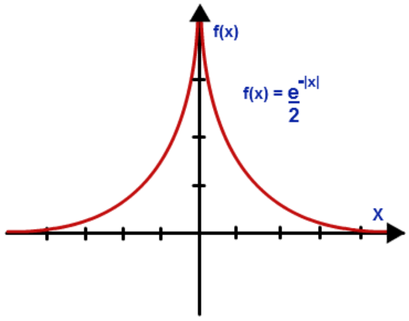
- Assume $\pi(\sigma^2) = 1/\sigma^2$.

Laplace Prior for β

Why the Laplace Prior?

- The Laplace distribution places most of its probability density at 0. However, the distribution also has a long tail.
- As a result, the regression estimates for the β's will often be exactly 0. However, there will still be some moderate or even large β estimates (due to the long tail).
- **Normal comparison**. We could place a Normal prior on the β's. That would give moderately-sized coefficients, with none being exactly zero b.c. the Normal is diffuse around 0.

The Laplace (or Double Exponential) Distribution



The Conditional Distributions

- $\beta \sim N(Q^{-1}X^Ty, \;\; \sigma^2 A^{-1}), Q = X^TX + D^{-1}$, where $D = \mathrm{diag}(\gamma_1^2, ..., \gamma_p^2)$.
- $\sigma^2 \sim \mathrm{InverseGamma}(a, b)$, where

  $a = (n+p)/2$ (scale parameter)

  $b = (y - X\beta)^T(y - X\beta)/2 + \beta^T D^{-1}\beta/2$ (scale parameter)
- $1/\gamma_i^2 \sim \mathrm{InverseGaussian}(c, d)$ where

  $a = \sqrt{\lambda^2\sigma^2/\beta_j^2}$ (mean parameter)

  $b = \lambda$ (shape parameter)

Alternative interpretation

The Lasso can also be obtained in an optimization framework.

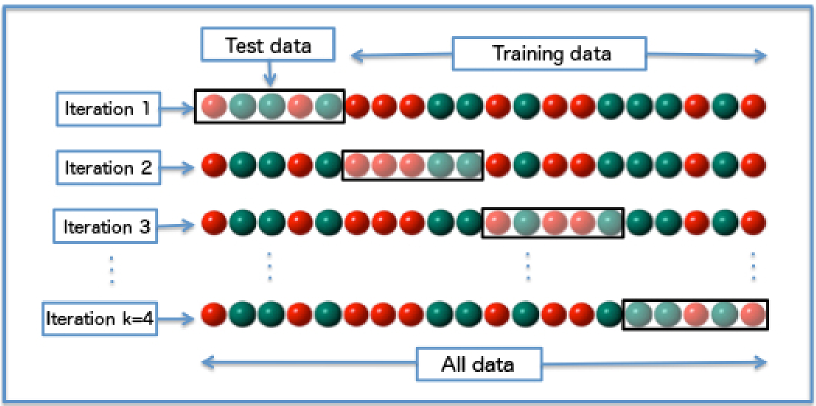
- $\min_{\beta} \quad (y - X\beta)^T(y - X\beta)^T + \lambda \sum_{j=1}^{p} |\beta_j|$.
- Estimates for β are found using an optimization routine.
- λ is a hyperparameter, and is usually found using cross-validation.

Cross validation for choosing λ

- Form a length n_λ sequence of candidate λ values.
- Split data into k equally sized partitions.
- For i in $1, ..., 2, ..., k$:
    - Fit model on the i-th partition for $\lambda = \lambda_j$ for $j = 1, ..., n_\lambda$.
        - Use fitted model to predict outcome values on the other partitions, and store the error associated with this prediction (this error is called the "cross-validation error" associated with λ_j on the i-th fold).
- Average the cross-validation values associated with each λ_j.
- Set $\lambda^* = \lambda_j$ for the λ_j which minimizes the average cross-validation error.

Cross validation for choosing λ, continued

Cross validation



Example: Predicting Views about Healthcare Reform

Example: Predicting Views about the Affordable Healthcare Act

- *Data*: 653 posts from the online debate forum `www.opposingviews.com` (circa 2009).
- *Topic*: the Affordable Healthcare Act.
- *Goal*: to discover the key word-level predictors of support for healthcare reform.
- *Method*: Lasso-penalized logistic regression.

$$\max_{\beta} \quad \ell(\beta) - \lambda \sum_{j=1}^{p} |\beta_j|,$$

where $\ell(\beta)$ corresponds to the traditional likelihood.

The Method, Reviewed

Unit of analysis: the documents.

Outcome variable (Y): an indicator of whether a given post supports the Affordable Healthcare.

Covariates (X): word usage patterns. Specifically, we use a **document-term matrix** as the predictor matrix for Y. The rows correspond to documents in the collection. The columns correspond to words/terms. We only include terms that occur more than 3% of documents.

> Notice that there are 344 predictors (words), but only 653 documents. There are too many predictors relative to the number of observations for traditional MLE.

The document term matrix

Stemming: Notice that the words have been stemmed. This means that words like “acting,” “acted,” and “actor” are reduced to the root word “act.” This is done to make documents more comparable for forming the document term matrix.

```
head(X[,1:10])
     abl access act actual advanc afford agre allow alreadi also
[1,]   0      0   0      0      0      0    0     0       0    0
[2,]   0      0   0      0      0      1    0     0       0    0
[3,]   0      0   0      0      0      0    0     0       0    0
[4,]   0      0   0      0      0      0    0     0       0    0
[5,]   0      0   0      0      0      0    0     0       0    0
[6,]   0      0   0      0      0      0    0     0       0    1
```

The non-zero predictors summarize the debate well

```
#only about 20 of the 344 words are non-zero at the optimal solution.

#the non-zero words predictive of a post opposing healthcare reform
    someon      govern        forc     freedom        will       creat
-0.55690254 -0.38756747 -0.36659896 -0.27802506 -0.24861069 -0.15729415
       els     respons       never      social      result
-0.12153437 -0.08180209 -0.05637216 -0.02378276 -0.01390274


#the non-zero words predictive of a post supporting healthcare reform
   reform     profit        ill   interest   wouldnt        lot
0.46018680 0.32690215 0.27945714 0.19946746 0.19197578 0.18234222
       per   institut     health      world      insur         us
0.16948682 0.16683929 0.14739395 0.11295995 0.10625697 0.07280756
      good    compani      often    citizen       part       deni
0.05521403 0.05022210 0.04710222 0.03281935 0.02422000 0.01290329
```

The upshot

- The Lasso is a powerful tool for solving prediction problems and for reducing *over-fitting problems*.
- The Lasso also has nice *model selection properties*. In fact, given certain assumptions, the Lasso possesses the **oracle property**. That is, there exists a λ value such that the probability that the Lasso recovers the true non-zero subset of coefficients is 1 as n grows.
- The properties of the Lasso are particularly attractive for text data, where the “true” model is often hard to find, and where there are often more predictors (e.g. word combinations) than documents.

Acknowledgements

These slides build from Park & Casella, 2008.