Module 1: Nonparametric Preliminaries

LASSO cont’d

STAT/BIOSTAT 527, University of Washington
Emily Fox
April 7th, 2014

©Emily Fox 2014

fMRI Prediction Subtask

Goal: Predict semantic features from fMRI image

Features of word

\[ \hat{\beta} = \left( X^T X \right)^{-1} X^T y \]

\[ n \]

\[ P \gg n \]

\[ p = \# \text{voxels} = 70,000 \]

\[ \text{rank deficient} \]
Regularization in Linear Regression

- Overfitting usually leads to very large parameter choices, e.g.:
  
  \[ -2.2 + 3.1 X - 0.30 X^2 \quad \text{and} \quad -1.1 + 4,700,910.7 X - 8,585,638.4 X^2 + \ldots \]

- **Regularized or penalized** regression aims to impose a “complexity” penalty by penalizing large weights
  
  - “Shrinkage” method

Ridge Regression

- Ameliorating issues with overfitting:
  
  - New objective:
Variable Selection

- Ridge regression: Penalizes large weights

- What if we want to perform “feature selection”?
  - E.g., Which regions of the brain are important for word prediction?
  - Can’t simply choose predictors with largest coefficients in ridge solution
  - Computationally impossible to perform “all subsets” regression

- Try new penalty: Penalize non-zero weights
  - Penalty: $L_1 \| \beta \|_1 = \sum_j |\beta_j|$ for $j \neq 0$
  - Leads to sparse solutions
  - Just like ridge regression, solution is indexed by a continuous param $\lambda$

LASSO Regression

- **LASSO**: least absolute shrinkage and selection operator

- New objective:

  \[
  \min_{\beta} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 + \lambda \| \beta \|_1
  \]

  \[
  = \min_{\beta} \text{RSS}(\beta) \quad \text{s.t.} \quad \| \beta \|_1 \leq B
  \]
Geometric Intuition for Sparsity

From Rob Tibshirani slides

Soft Thresholding

- To see why LASSO results in sparse solutions, look at conditions that must hold at optimum
  - $L_1$ penalty $||\beta||_1$ is not differentiable whenever $\beta_j = 0$
  - Look at subgradient...
Subgradients of Convex Functions

- Gradients lower bound convex functions:
  - Gradients are unique at \( x \) if function differentiable at \( x \)
  - Subgradients: Generalize gradients to non-differentiable points:
    - Any plane that lower bounds function:

Soft Threshholding

- Gradient of RSS term:
  - Subgradient of full objective:
Soft Threshholding

- Set subgradient = 0:
  \[
  \partial_{\beta_j} F(\beta) = \begin{cases} 
  a_j \beta_j - c_j - \lambda & \beta_j < 0 \\
  -c_j - \lambda, -c_j + \lambda & \beta_j = 0 \\
  a_j \beta_j - c_j + \lambda & \beta_j > 0 
  \end{cases}
  \]

- The value of \( c_j = 2 \sum_{i=1}^{N} x_{ij}^t (y_i - \beta_{-j} x_{ij}) \) constrains \( \beta_j \)

---

Soft Threshholding

\[
\hat{\beta}_j = \begin{cases} 
(c_j + \lambda)/a_j & c_j < -\lambda \\
0 & c_j \in [-\lambda, \lambda] \\
(c_j - \lambda)/a_j & c_j > \lambda 
\end{cases}
\]

From Kevin Murphy textbook
Coordinate Descent

- Given a function $F$ 
  - Want to find minimum

- Often, hard to find minimum for all coordinates, but easy for one coordinate

- Coordinate descent:
  - How do we pick a coordinate?
  - When does this converge to optimum?

Stochastic Coordinate Descent for LASSO (aka Shooting Algorithm)

- Repeat until convergence
  - Pick a coordinate $j$ at random
    - Set: 
      $$\hat{\beta}_j = \begin{cases} 
      \frac{(c_j + \lambda)}{a_j} & c_j < -\lambda \\
      0 & c_j \in [-\lambda, \lambda] \\
      \frac{(c_j - \lambda)}{a_j} & c_j > \lambda 
      \end{cases}$$

    - Where:
      $$a_j = 2 \sum_{i=1}^{N} (x_i^j)^2$$
      $$c_j = 2 \sum_{i=1}^{N} x_i^j(y_i - \hat{\beta}' - x_i^\prime \hat{\beta}'_j)$$

  - For convergence rates, see Shalev-Shwartz and Tewari 2009

- Other common technique = LARS
  - Least angle regression and shrinkage, Efron et al. 2004
Recall: *Ridge Coefficient Path*

- Typical approach: select \( \lambda \) using cross validation

Now: *LASSO Coefficient Path*

- From Kevin Murphy textbook
LASSO Example

<table>
<thead>
<tr>
<th>Term</th>
<th>Least Squares</th>
<th>Ridge</th>
<th>Lasso</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.465</td>
<td>2.452</td>
<td>2.468</td>
</tr>
<tr>
<td>lcavol</td>
<td>0.680</td>
<td>0.420</td>
<td>0.533</td>
</tr>
<tr>
<td>lweight</td>
<td>0.263</td>
<td>0.238</td>
<td>0.169</td>
</tr>
<tr>
<td>age</td>
<td>−0.141</td>
<td>−0.046</td>
<td></td>
</tr>
<tr>
<td>lbph</td>
<td>0.210</td>
<td>0.162</td>
<td>0.002</td>
</tr>
<tr>
<td>svi</td>
<td>0.305</td>
<td>0.227</td>
<td>0.094</td>
</tr>
<tr>
<td>lcp</td>
<td>−0.288</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>gleason</td>
<td>−0.021</td>
<td>0.040</td>
<td></td>
</tr>
<tr>
<td>pgg45</td>
<td>0.267</td>
<td>0.133</td>
<td></td>
</tr>
</tbody>
</table>

From Rob Tibshirani slides

Sparsistency

- Typical Statistical Consistency Analysis:
  - Holding model size \((p)\) fixed, as number of samples \((n)\) goes to infinity, estimated parameter goes to true parameter

- Here we want to examine \(p >> n\) domains
- Let both model size \(p\) and sample size \(n\) go to infinity!
  - Hard case: \(n = k \log p\)
Sparsistency

- Rescale LASSO objective by $n$:

- Theorem (Wainwright 2008, Zhao and Yu 2006, ...):
  - Under some constraints on the design matrix $X$, if we solve the LASSO regression using

  Then for some $c_1 > 0$, the following holds with at least probability

  - The LASSO problem has a unique solution with support contained within the true support
  - If $\min_{j \in S(\beta^*)} |\beta_j^*| > c_2 \lambda_n$ for some $c_2 > 0$, then $S(\hat{\beta}) = S(\beta^*)$

Comments

- In general, can’t solve analytically for GLM (e.g., logistic reg.)
  - Gradually decrease $\lambda$ and use efficiency of computing $\hat{\beta}(\lambda_k)$ from $\hat{\beta}(\lambda_{k-1})$ = warm-start strategy
  - See Friedman et al. 2010 for coordinate ascent + warm-starting strategy

- If $n > p$, but variables are correlated, ridge regression tends to have better predictive performance than LASSO (Zou & Hastie 2005)
  - Elastic net is hybrid between LASSO and ridge regression
Fused LASSO

- Might want coefficients of neighboring voxels to be similar
- How to modify LASSO penalty to account for this?
- Graph-guided fused LASSO
  - Assume a 2d lattice graph connecting neighboring pixels in the fMRI image
  - Penalty:

A Bayesian Formulation

- Consider a model with likelihood
  \[ y_i \mid \beta \sim N(\beta_0 + x_i^T \beta, \sigma^2) \]
  and prior
  \[ \beta_j \sim \text{Lap}(\beta_j; \lambda) \]
  where
  \[ \text{Lap}(\beta_j; \lambda) = \frac{\lambda}{2} e^{-\lambda |\beta_j|} \]
- For large \( \lambda \)
  - LASSO solution is equivalent to the **mode** of the posterior
  - Note: posterior mode ≠ posterior mean in this case
  - There is no closed-form for the posterior. Rely on approx. methods.
Reading

- Hastie, Tibshirani, Friedman: 3.4, 3.8.6

What you should know

- LASSO objective
- Geometric intuition for differences between ridge and LASSO solns
- How LASSO performs soft thresholding
- Shooting algorithm
- Idea of sparsistency
- Ways in which other L1 and L1-Lp objectives can be encoded
  - Elastic net
  - Fused LASSO