Sparse linear models

Jiali Lin

Virginia Tech

January 15, 2017

Outline

Introduction

Bayesian Variable Selection

ℓ_1 regularization: basics

ℓ_1 regularization: algorithms

ℓ_1 regularization: extensions

Non-convex regularizers

Introduction

- Consider a generalized linear model, $p(y|\boldsymbol{x}) = p(y|f(\boldsymbol{w}^T\boldsymbol{x}))$ for some link function f.
- **Goal**: perform feature selection by encouraging the weight vector $\boldsymbol{w}$ to be **sparse**, i.e., to have lots of zeros.
- When p is large, it becomes unrealistic to go through all possible choices and determine the best subset of variables based some selection criterion such as, AIC or BIC.

Outline

Introduction

Bayesian Variable Selection

ℓ_1 regularization: basics

ℓ_1 regularization: algorithms

ℓ_1 regularization: extensions

Non-convex regularizers

Bayesian Variable Selection

- A natural way to pose the variable selection problem is to introduce a hyper-parameter γ_j to the prior w_j, where

$$\gamma_j = \begin{cases} 1, & \text{feature } j \text{ is in} \\ 0, & \text{feature } j \text{ is out} \end{cases}$$

- We will seek various summary statistics. A natural one is the posterior mode, or **MAP** estimate

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}}{\operatorname{argmax}}\, p(\boldsymbol{\gamma}|D) = \operatorname{argmin} f(\boldsymbol{\gamma})$$

- Drawbacks:
  - Still need to search over all possible $\boldsymbol{\gamma}$.
  - The mode of the posterior distribution does not necessarily represent the full posterior distribution well.
  - Alternative: median of the marginal inclusion probabilities. Then we have $\hat{\boldsymbol{\gamma}} = \{j : p(\gamma_j = 1|D) > .5\}$.

Case I: Spike and slab model

- **Main idea**: Find a prior has a mixture of a point mass at 0 (forcing $w_j = 0$, and excluding that covariate j) and a flat prior (Gaussian, often) on the included variables.
- A common prior on the feature inclusion vector

$$p(\boldsymbol{\gamma}|\pi_0) = \prod_{j=1}^{n} \mathsf{Bern}(\gamma_j|\pi_0) = \pi_0^{\|\gamma\|_0}(1-\pi_0)^{(p-\|\gamma\|_0)}$$

- Spike and slab prior

$$w_j|\sigma^2, \gamma_j \sim \begin{cases} \delta_0(w_j), & \text{if } \gamma_j = 0 \\ N(w_j|0, \sigma^2\sigma_w^2), & \text{if } \gamma_j = 1 \end{cases}$$

- The interpretation of the two mixture components: clustering each predictor as noise (the spike at 0; excluded) and signal (the slab; included).

Case II: Bernoulli-Gaussian model

- The prior distribution of w_j

$$w_j|\gamma_j \sim \gamma_j N(0, v_{1j}^2) + (1-\gamma_j) N(0, v_{0j}^2)$$

- v_{1j} is far from zero but v_{0j} is close to zero, $v_{1j} \geq v_{0j} > 0$.
- This prior is a normal with variance either large or close to zero depending on the value of γ_j.
- When $\gamma_j = 0$, w_j has a normal prior with small variance v_{0j}. Since v_{0j} is close to zero, w_j can be a priori excluded from the subset.
- We update $\boldsymbol{\gamma}$ using a Gibbs sampler. See demo `BvsGibbsDemo`.

Case III: Revise the prior

- Now, assume
$$\boldsymbol{\beta}|\boldsymbol{\gamma}, \sigma^2 \sim N(0, \sigma^2 \boldsymbol{\Sigma}_{\boldsymbol{\gamma}})$$
- This makes $(\boldsymbol{\beta}, \sigma^2)$ conjugate prior. Therefore we can integrate out them analytically from the joint posterior to get $\pi(\boldsymbol{\gamma}|\boldsymbol{Y})$.
- Given the marginal posterior $\pi(\boldsymbol{\gamma}|\boldsymbol{Y})$, we can also design a MH sampler to get posterior samples of $\boldsymbol{\gamma}$.

Case III: Revise the prior (Cont'd)

- Generate a candidate sample γ^* from a transition kernel (proposal distribution), $f(\gamma^*|\boldsymbol{\gamma})$, then update $\boldsymbol{\gamma}$ by γ^* with probability

$$\min\{\frac{\pi(\gamma^*|Y)f(\boldsymbol{\gamma}|\gamma^*)}{\pi(\boldsymbol{\gamma}|Y)f(\gamma^*|\boldsymbol{\gamma})}, 1\}$$

- For convenience, the transition kernel can be chosen to be symmetric so that the $f(\boldsymbol{\gamma}|\gamma^*)$ term and $f(\gamma^*|\boldsymbol{\gamma})$ term in the proposal ratio are canceled.
- The candidate sample γ^* is typically generated:
  - With probability ϕ, randomly change one component of $\boldsymbol{\gamma}$;
  - With probability $1-\phi$, randomly choose two components with 0 and 1 and swap them, known as **switch and swap proposal**.
- Based on the marginal posterior developed in previous example, we can design a MH using switch-swap proposal. See demo `BvsMHDemo`.

Outline

Introduction

Bayesian Variable Selection

ℓ_1 regularization: basics

ℓ_1 regularization: algorithms

ℓ_1 regularization: extensions

Non-convex regularizers

ℓ_1 regularization

- The ideal approach to introducing sparsity is to use the ℓ_0 norm (number of non-zero elements) for coefficients $\boldsymbol{w}$.
- In practice, ℓ_1 norm is often used since it is a convex approximation of the ℓ_0 norm, and thus makes computation much easier.
- This amounts to introducing a **Laplace prior** (or a **double exponential prior**) on $\boldsymbol{w}$

$$p(\boldsymbol{w}|\lambda) = \prod_{j=1}^{p} \mathsf{Lap}(w_j|0, 1/\lambda) \propto \prod_{j=1}^{p} \exp\{-\lambda|w_j|\}$$

- Then, the penalized negative log likelihood has the form

$$-\log p(\boldsymbol{w}|D) = -\log p(D|\boldsymbol{w}) - \log p(\boldsymbol{w}|\lambda) = \mathsf{NLL} + \lambda\|\boldsymbol{w}\|_1$$

Why does ℓ_1 regularization yield sparse solutions?

The MAP estimator $\hat{\boldsymbol{w}}_{\text{MAP}}$ is obtained by solving the following optimization problem,

$$\min_{\boldsymbol{w}} \text{RSS}(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_1.$$

Or equivalently,

$$\min_{\boldsymbol{w}} \text{RSS}(\boldsymbol{w}) \quad \text{s.t.} \|\boldsymbol{w}\|_1 \leq B$$

where B is a given upper bound of the ℓ_1 norm, λ dictates the sparsity weight. This optimization problem is called **Lasso**.

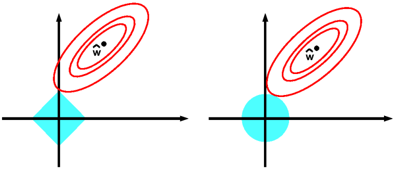
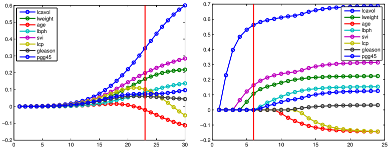


Figure: Illustration of ℓ_1 (left) vs ℓ_2 (right) regularization of a least squares problem. Based on Figure 3.12 of (Hastie et al. 2001).

Regularization path

- **Regularization path**: as we increase λ, the solution vector $\hat{\boldsymbol{w}}(\lambda)$ will tend to get sparser, although not necessarily monotonically. We can plot the values $\hat{w}_j(\lambda)$ vs λ for each feature j.
- **Ridge regression**: for any finite value of λ, all coefficients are non-zero; furthermore, they increase in magnitude as λ is decreased.
- **Lasso**: as B increases, the coefficients gradually "turn on". But for any value between 0 and $B_{\max} = \|\hat{\boldsymbol{w}}_{OLS}\|_1$, the solution is sparse.



(a) (b)

Figure: (a) Based on Figure 3.8 of (Hastie et al. 2009). Figure generated by `RidgePathProstate`. (b) Based on Figure 3.10 of (Hastie et al. 2009). Figure generated by `LassoPathProstate`.

Outline

Introduction

Bayesian Variable Selection

ℓ_1 regularization: basics

ℓ_1 regularization: algorithms

ℓ_1 regularization: extensions

Non-convex regularizers

ℓ_1 regularization: algorithms

(See scribes for details)

1. **Coordinate descent**: optimize variables one by one. We can choose to update the coordinate for which the gradient is steepest.
2. **Least-angle regression (LARS)**: similar to forward stepwise regression, but instead of including variables at each step, the estimated parameters are increased in a direction equiangular to each one's correlations with the residual.
3. **Proximal and gradient projection methods**: solve large scale convex optimization problems that has a form

$$f(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) + R(\boldsymbol{\theta})$$

   where $L(\boldsymbol{\theta})$ (loss) is convex and differentiable, and $R(\boldsymbol{\theta})$ (regularizer) is convex but not differentiable.

EM for lasso

4. We can solve the lasso problem using **EM**.
   – **Key:** Use the Laplace distribution as a **Gaussian scale mixture (GSM)**

$$\mathsf{Lap}(w_j|0, 1/\gamma) = \frac{\gamma}{2}e^{-\gamma|w_j|} = \int N(w_j|0, \tau_j^2)\mathsf{Ga}(\tau_j^2|1, \frac{\gamma}{2})d\tau_j^2$$

   – Laplace is a GSM where the mixing distribution on the variances is the exponential distribution.
   – The corresponding joint distribution has the form

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{\tau}, \sigma^2|\boldsymbol{X}) =& N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{w}, \sigma^2\boldsymbol{I}_N)N(\boldsymbol{w}|0, \boldsymbol{D}_\tau) \\ & \mathsf{IG}(\sigma^2|a_\sigma, b_\sigma)[\prod_j \mathsf{Ga}(\tau_j^2|1, \gamma^2/2)] \end{aligned}$$

EM for lasso (cont'd)

- In the E step, infer τ_j^2 and σ^2.
- In the M step, estimate $\boldsymbol{w}$.
- The resulting estimate $\hat{\boldsymbol{w}}$ is the same as the lasso estimator.

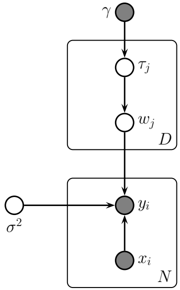


Figure: Representing lasso using a Gaussian scale mixture prior.

EM for lasso (cont'd)

Why EM?

- Can easily derive find ℓ_1-regularized parameter estimates.
- Suggests other priors on the variances besides $\text{Ga}(\tau_j^2|1, \gamma^2/2)$.
- It makes it clear how we can compute the full posterior, $p(\boldsymbol{w}|\mathcal{D})$, rather than just a MAP estimate (**Bayesian lasso**).

Outline

Introduction

Bayesian Variable Selection

ℓ_1 regularization: basics

ℓ_1 regularization: algorithms

ℓ_1 regularization: extensions

Non-convex regularizers

Group Lasso

- **Group Lasso** allows predefined groups of covariates to be selected into or out of a model together.
- Partition the parameter vector into G groups. We now minimize

$$J(\boldsymbol{w}) = \mathsf{NLL}(\boldsymbol{w}) + \sum_{g=1}^{G} \lambda_g \|\boldsymbol{w}_g\|_2 \quad \|\boldsymbol{w}_g\|_2 = \sqrt{\sum_{j \in g} w_j^2}$$

- E.g. if we have groups $\{1, 2\}$ and $\{3, 4, 5\}$, the objective becomes

$$J(\boldsymbol{w}) = NLL(\boldsymbol{w}) + \lambda \left[\sqrt{2}\sqrt{(w_1^2 + w_2^2)} + \sqrt{3}\sqrt{(w_3^2 + w_4^2 + w_5^2)} \right]$$

- Group sparsity: using the square root penalizes the radius of a ball containing the group's weight vector, that is, the only way for the radius to be small is if all elements are small.

GSM interpretation of group lasso

- Group lasso is equivalent to MAP estimation using the following prior

$$p(\boldsymbol{w}|\gamma, \sigma^2) \propto \exp(-\frac{\gamma}{\sigma} \sum_{g=1}^{G} \|\boldsymbol{w}_g\|_2)$$

- Now one can show that this prior can be written as a GSM, as follows

$$\boldsymbol{w}_g|\sigma^2, \tau_g^2 \sim N(0, \sigma^2 \tau_g^2 \boldsymbol{I}_{d_g}) \quad \tau_g^2|\gamma \sim \mathsf{Ga}(\frac{d_g + 1}{2}, \frac{\gamma}{2})$$

  where d_g is the size of group g.
- There is one variance term per group, each of which comes from a Gamma prior, whose shape parameter depends on the group size, and whose rate parameter is controlled by γ.

Sparse group lasso

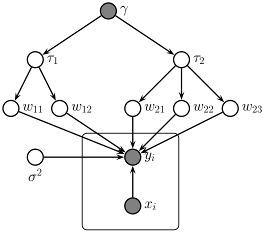


Figure: Graphical model for group lasso with 2 groups, the first has size $G_1 = 2$, the second has size $G_2 = 3$.

- The group lasso does not yield sparsity within a group. That is, if a group of parameters is non-zero, they will all be non-zero.
- Consider **sparse group lasso** criterion:

$$J(\boldsymbol{w}) = \mathsf{NLL}(\boldsymbol{w}) + \lambda_1 \sum_{g=1}^{G} \|\boldsymbol{w}_g\|_2 + \lambda_2 \|\boldsymbol{w}_g\|_1$$

Fused lasso

- **Fused lasso**: we want neighboring coefficients to be similar to each other, in addition to being sparse, by using a prior

$$p(\boldsymbol{w}|\sigma^2) \propto \exp(-\frac{\lambda_1}{\sigma}\sum_{j=1}^{D}|w_j| - \frac{\lambda_2}{\sigma}\sum_{j=1}^{D-1}|w_{j+1}| - w_j)$$

Outline

Non-convex regularizers

Potential problems of Laplace prior:

- It does not put enough probability mass near 0, so it does not sufficiently suppress noise.
- It does not put enough probability mass on large values, so it causes shrinkage of relevant coefficients, corresponding to "signal".

Solution:

- Use more flexible kinds of priors which have a larger spike at 0 and heavier tails.

Generalized Norms: Bridge Regression

Bridge regression has the form

$$\hat{\boldsymbol{w}} = \mathsf{NLL}(\boldsymbol{w}) + \lambda \sum_j |w_j|^b$$

for $b \geq 0$. This corresponds to MAP estimation using a **exponential power distribution** given by

$$\mathsf{ExpPower}(\boldsymbol{w}|\mu, a, b) = \frac{b}{2a\Gamma(1 + 1/b)} \exp(-\frac{|\boldsymbol{w} - \mu|^b}{a})$$

- Convex objective function (true norm): $b \geq 1$.
- Encourages sparse solutions (cusp at zero): $b \leq 1$.
- Lasso/Laplacian (convex & sparsity): $b = 1$.
- Ridge/Gaussian (classical, closed form solutions): $b = 2$.
- Sparsity via discrete counts (greedy search): $b \to 0$.

Hierarchical adaptive lasso

- Recall: lasso may use a large value of λ to "squash" the irrelevant parameters, but this then over-penalizes the relevant parameters.
- Bayesian can associate a different penalty parameter with each parameter.
- How? Let τ_j^2 have its own private tuning parameter, γ_j, which coming from the conjugate prior

$$\begin{aligned}
\gamma_j &\sim \mathsf{IG}(a, b) \\
\tau_j^2 | \gamma_j &\sim \mathsf{Ga}(1, \gamma_j^2/2) \\
w_j | \tau_j^2 &\sim N(0, \tau_j^2)
\end{aligned}$$