1

14-11-2016

Statistical Learning III

- Ridge regression computation
- Comparison Ridge, Lasso, Best-subset
- Probability Theory remarks
- Bayesian Statistics
  a) Intro
  b) Laplace's rule of succession
  c) MAP interpretation of Ridge, Lasso

1. Ridge regression computation

$$\hat{\beta}_{Ridge} = \underset{\beta}{\arg\min}\ RSS(\beta) + \lambda \sum_{j=2}^{p} \beta_j^2 \qquad T = \begin{pmatrix} y_1 \\ x_1 \end{pmatrix}, \dots, \begin{pmatrix} y_N \\ x_N \end{pmatrix}$$

Can interpret as least squares with extra fake training data

$$\begin{pmatrix} y_{N+1} \\ x_{N+1} \end{pmatrix}, \dots, \begin{pmatrix} y_{N+p-1} \\ x_{N+p-1} \end{pmatrix}$$

$$y_{N+j} = 0 \qquad x_{N+j} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \sqrt{\lambda} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow (j+1)\text{-th position for } j = 1, \dots, p-1$$

Then $(y_{N+j} - x_{N+j}^T \beta)^2 = \lambda \beta_{j+1}^2$, so

$$\hat{\beta}_{Ridge} = \underset{\beta}{\arg\min} \sum_{i=1}^{N+p-1} (y_i - x_i^T \beta)^2 = \text{least squares with extra fake data.}$$

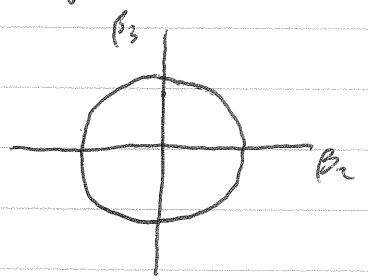
5. Comparison of Ridge, Lasso, Best-subset Selection

a) $\underbrace{\text{Ridge, Lasso}}$ vs Best-subset

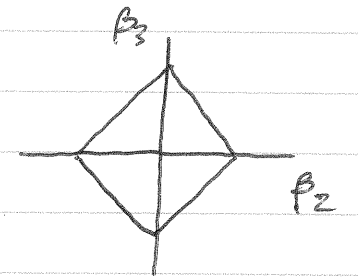
small changes in data cause
small changes in estimate, so less variance than best-subset.

b) Ridge vs Lasso

Ridge
$\sum_{j=2}^{p} \beta_j^2 \leq t$

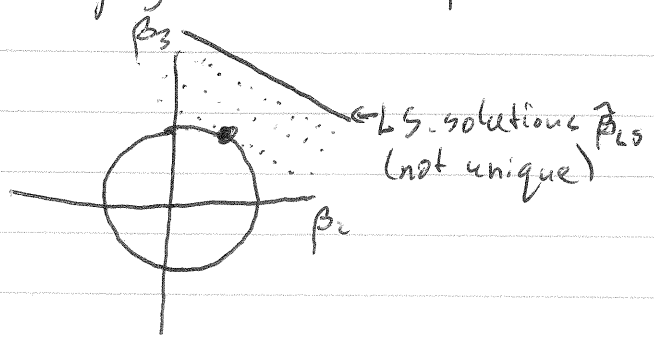


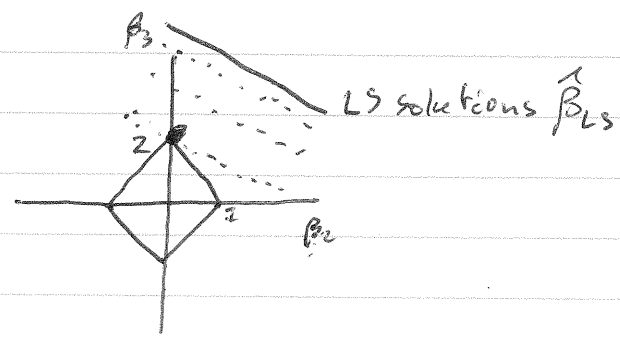
Lasso
$\sum_{j=1}^{p} |\beta_j| \leq t$



* Figure 3.11

* For highly correlated features:





β_2 and β_3 get about equal weight

minor noise determines which of corners 1 and 2 is chosen: one variable gets all weight, the other 0.

3. Probability Theory Remarks

Bayes' rule:

$$Pr(A|B) = \frac{Pr(B|A) \cdot Pr(A)}{Pr(B)}$$

Often used in Bayesian ~~stu~~ and standard frequentist statistics

Densities are slippery:

* Location of maximum depends on choice of parametrization
* Uniform density in one parametrization is not uniform in another parametrization

(see slides)

2. Bayesian Statistics

a] Intro

Probability model: $\mathcal{P} = \{ P_\theta(x, y) \mid \theta \in \Theta \}$ or $\mathcal{P} = \{ P_\theta(y) \mid \theta \in \Theta \}$

Frequentist statistics (standard):
- optimal parameter θ^* is fixed, but unknown
- diversity of methods
- need proofs/experiments to justify methods

Bayesian statistics:
- pretend that true parameter θ^* is a random variable distributed according to prior distribution $\pi(\theta)$ that we know.

$T = Y_1, \ldots, Y_N$ (no features for simplicity)

$$\Pr(T, \theta) = P_\theta(T) \cdot \pi(\theta)$$ is joint distribution of data and parameters

Since we know the full joint distribution, can simply use probability theory to compute any probability we are interested in. ← single method

↳ if we estimate different probabilities this way, they are beautifully consistent.

Bayesian premise is too strong:
- prior π is chosen for computational or information theoretic properties in practice, so cannot blindly assume θ^* is random sample from π.
- need proofs/experiments to justify Bayesian methods

Modern frequentist motivation:

- If we choose π right, then often works really well (both in theory and in practice).
- Learn faster if true θ^* has high prior probability, slower if true θ^* has small prior probability, so can use π to express prior knowledge about our data.

Posterior Distribution:

$$\pi(\theta|T) := Pr(\theta|T) = \frac{Pr(\theta,T)}{Pr(T)} = \frac{P_\theta(T)\cdot\pi(\theta)}{Pr(T)}$$

- Often puts its probability mass closer and closer to θ^* as $N \to \infty$. (vd Vaart et al.)
- Expresses uncertainty about θ

Predictive Distribution:

$$Pr(Y|T) = \int_\Theta P_\theta(Y)\cdot\pi(\theta|T)\,d\theta = \frac{Pr(Y,T)}{Pr(T)}$$

new sample outside of training set

often better predictions than frequentist $P_{\hat\theta}(Y)$ because $\pi(\theta|T)$ keeps track of uncertainty better than fixed single choice $\hat\theta$.

b] Example: Laplace Rule of Succession

Bernoulli model: $P_\theta(Y) = \begin{cases}\theta & \text{for } Y=1\\ 1-\theta & \text{for } Y=0\end{cases}$ $\quad \theta\in[0,1]$

Maximum likelihood: $\hat\theta = \frac{n_1}{N} \to P_{\hat\theta}(Y=1) = \hat\theta = \frac{n_1}{N}$ ← dangerous for prediction if $n_1=0$

Suppose $\pi(\theta)=1$ is uniform prior density, ⇐ not the same as "no prior knowledge" because depends on parametrisation

Then $Pr(Y=1|T) = \frac{n_1+1}{N+2}$

(Laplace, 1814)

$Pr(Y=0|T) = \frac{n_0+1}{N+2}$

beta function

Proof: $$\frac{Pr(Y,T)}{Pr(T)} = \frac{\int P_\theta(Y,T)\cdot\pi(\theta)\,d\theta}{\int P_\theta(T)\cdot\pi(\theta)\,d\theta} = \frac{\int \theta^{n_1+Y}(1-\theta)^{n_0+1-Y}\,d\theta}{\int \theta^{n_1}(1-\theta)^{n_0}\,d\theta} = \frac{n_Y+1}{N+2} \quad \square$$

c] MAP interpretation of Ridge and Lasso

Maximum a Posteriori (MAP) parameters:

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\ \pi(\theta|T) = \underset{\theta}{\operatorname{argmax}}\ \frac{P_\theta(T)\cdot\pi(\theta)}{Pr(T)}$$

$$= \underset{\theta}{\operatorname{argmax}}\ P_\theta(T)\cdot\pi(\theta)$$

- maximizes posterior <u>density</u>, so for continuous parameters depends on parametrisation!
- "real" Bayesians prefer prediction with predictive distribution.

Ridge / Lasso:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\ RSS(\beta) + \lambda\, pen(\beta)$$

ridge: $pen(\beta) = \sum_{j=2}^{p} \beta_j^2$ lasso: $pen(\beta) = \sum_{j=2}^{p} |\beta_j|$

Suppose Gaussian noise:

$$Y = X^T\beta + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0,\sigma^2)$$

$$P_\beta(Y_1,\dots,Y_N|X_1,\dots,X_N) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(Y_i - X_i^T\beta)^2}{2\sigma^2}}$$

$$\hat{\beta}_{MAP} = \underset{\beta}{\operatorname{argmax}}\ P_\beta(Y_1,\dots,Y_N|X_1,\dots,X_N)\cdot\pi(\beta)$$

$$= \underset{\beta}{\operatorname{argmin}}\ -\log P_\beta(Y_1,\dots,Y_N|X_1,\dots,X_N) - \log\pi(\beta)$$

$$= \underset{\beta}{\operatorname{argmin}}\ N\cdot\left(-\log\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \frac{1}{2\sigma^2}RSS(\beta) - \log\pi(\beta)$$

$$= \underset{\beta}{\operatorname{argmin}}\ RSS(\beta) - 2\sigma^2\log\pi(\beta)$$

Suppose data have been pre-processed such that we can assume that the intercept $\hat{\beta}_1 = 0$.

Ridge: Choose π s.t.

$$\beta_1 = 0 \quad \text{with prob. } 1$$
$$(\beta_2, \dots, \beta_p) \sim \mathcal{N}(0, \tau^2 I)$$

$$-\log \pi(\beta) = \sum_{j=2}^{p} -\log\left(\frac{1}{\sqrt{2\pi\tau^2}} \cdot e^{-\frac{(\beta_j - 0)^2}{2\tau^2}}\right)$$

$$= (p-1)\left(-\log \frac{1}{\sqrt{2\pi\tau^2}}\right) + \sum_{j=2}^{p} \frac{\beta_j^2}{2\tau^2}$$

$$\hat{\beta}_{MAP} = \underset{\beta}{\operatorname{argmin}}\ RSS(\beta) + \frac{\sigma^2}{\tau^2} \sum_{j=2}^{p} \beta_j^2$$

is ridge with $\lambda = \frac{\sigma^2}{\tau^2}$

Is also posterior mean $\mathbb{E}_{\pi(\beta|T)}[\beta]$

Lasso: Choose π s.t.

$$\beta_1 = 0 \quad \text{with prob. } 1.$$
$$(\beta_2, \dots, \beta_p) \sim \prod_{j=2}^{p} \frac{1}{2\tau} e^{-\frac{|\beta_j|}{\tau}}$$

$$-\log \pi(\beta) = \sum_{j=2}^{p} -\log\left(\frac{1}{2\tau} \cdot e^{-\frac{|\beta_j|}{\tau}}\right)$$

$$= (p-1)\left(-\log \frac{1}{2\tau}\right) + \sum_{j=2}^{p} \frac{|\beta_j|}{\tau}$$

$$\hat{\beta}_{MAP} = \underset{\beta}{\operatorname{argmin}}\ RSS(\beta) + \frac{2\sigma^2}{\tau} \sum_{j=2}^{p} |\beta_j|$$

is Lasso with $\lambda = \frac{2\sigma^2}{\tau}$.

Remarks:

- "real" Bayesian prefers predicting with predictive distribution
- MAP + CV to determine λ is very unBayesian.