Ridge/Lasso Regression, Model selection

Xuezhi Wang

Computer Science Department
Carnegie Mellon University

10701-recitation, Apr 22

Outline

Outline

Linear Regression

Data X: $N \times P$ matrix, Target y: $N \times 1$ vector

- N samples, each sample has P features

Want to find θ so that y and $X\theta$ are as close as possible

- Pick θ that minimizes the cost function

$$L = \frac{1}{2}\sum_i (y_i - X_i\theta)^2 = \frac{1}{2}||y - X\theta||^2$$

- use gradient descent

$$\theta_j^{t+1} = \theta_j^t - step * \frac{\partial L}{\partial \theta_j} = \theta_j^t - step * \sum_i (y_i - X_i\theta)(-X_{ij})$$

Linear Regression

Matrix form:

$$L = \frac{1}{2}\sum_i (y_i - X_i\theta)^2 = \frac{1}{2}||y - X\theta||^2$$
$$= \frac{1}{2}(y - X\theta)^\top (y - X\theta)$$
$$= \frac{1}{2}(y^\top y - y^\top X\theta - \theta^\top X^\top y + \theta^\top X^\top X\theta)$$

Take derivative w.r.t. θ

$$\frac{\partial L}{\partial \theta} = \frac{1}{2}(-2X^\top y + 2X^\top X\theta) = 0$$

Hence we get

$$\theta = (X^\top X)^{-1} X^\top y$$

Linear Regression

Comparison of iterative methods and matrix methods:

- matrix methods achieve solution in a single step, but can be infeasible for real-time data, or large amount of data.
- iterative methods can be used in large practical problems, but need to decide learning rate

Any problems?

- Data X is an $N \times P$ matrix
- Usually $N > P$, i.e., number of data points larger than feature dimensions. And usually X is of full column rank.
- Under this case $X^\top X$ have rank P, i.e., invertible
- What if X has less than full column rank?

Outline

Regularization: ℓ_2 norm

Ridge Regression:

$$\min_{\theta} \frac{1}{2} \sum_i (y_i - X_i\theta)^2 + \lambda ||\theta||_2^2$$

Solution is given by:

$$\theta = (X^\top X + \lambda I)^{-1} X^\top y$$

- Results in a solution with small θ
- Solves the problem that $X^\top X$ is not invertible

Regularization: ℓ_1 norm

Lasso Regression:

$$\min_{\theta} \frac{1}{2} \sum_i (y_i - X_i\theta)^2 + \lambda ||\theta||_1$$

Solution is given by taking subgradient:

$$\sum_i (y_i - X_i\theta)(-X_{ij}) + \lambda t_j$$

where t_j is the subgradient of ℓ_1 norm,

$$t_j = sign(\theta_j) \text{ if } \theta_j \neq 0, t_j \in [-1, 1] \text{ otherwise}$$

- Sparse solution, i.e., θ will be a vector with more zero coordinates.
- Good for high-dimensional problems

Solving Lasso regression

Efron et al. proposed LARS (least angle regression) which computes the LASSO path efficiently

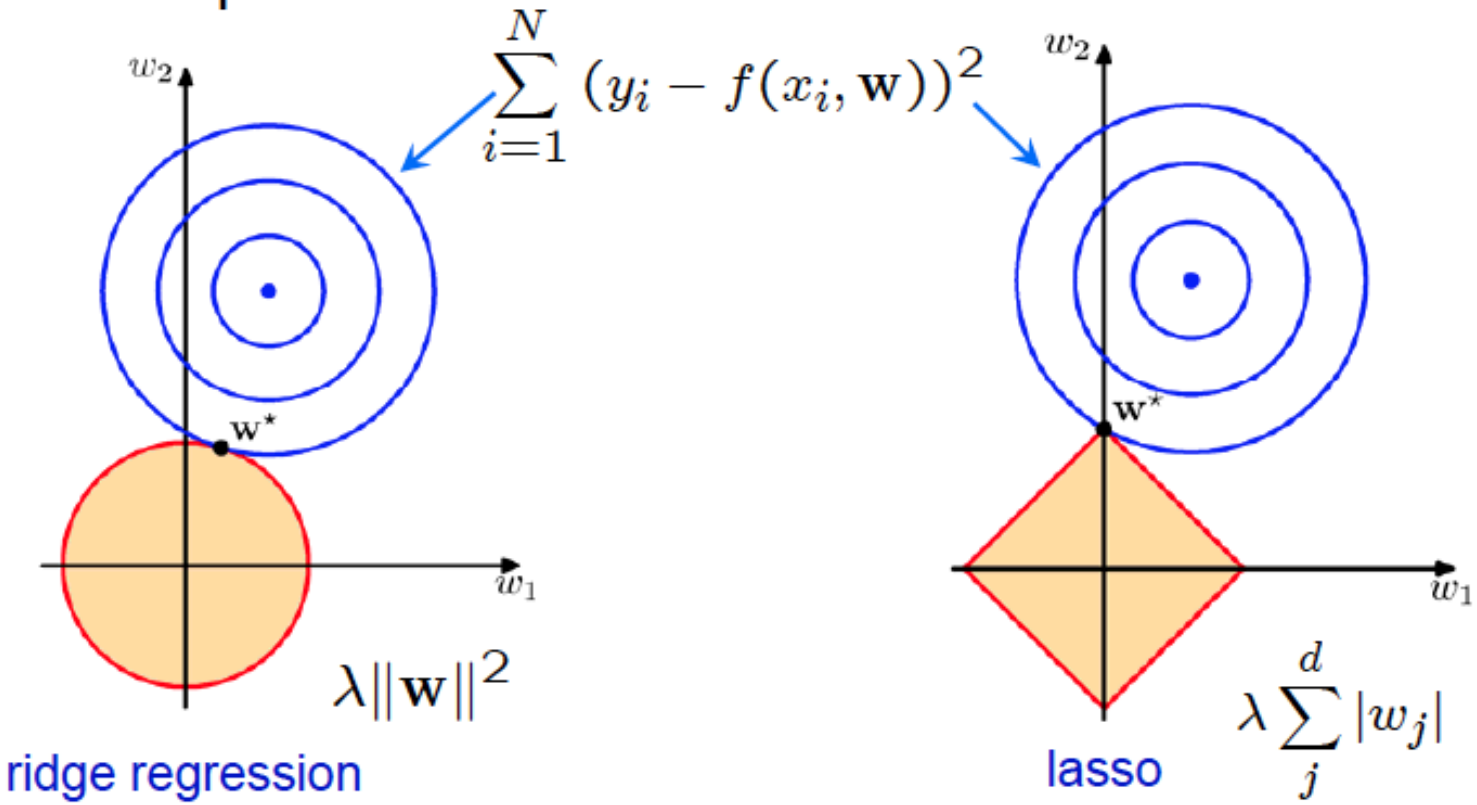
Forward stagewise algorithm

- Assume X is standardized and y is centered
- choose small ϵ
  - Start with initial residual $r = y$, and $\theta_1 = ... = \theta_P = 0$
  - Find the predictor Z_j (jth column of X) most correlated with r
  - Update $\theta_j \leftarrow \theta_j + \delta_j$, where $\delta_j = \epsilon \cdot \text{sign}(Z_j^\top r)$
  - Set $r \leftarrow r - \delta_j Z_j$, repeat steps 2 and 3

Comparison of Ridge and Lasso regression:

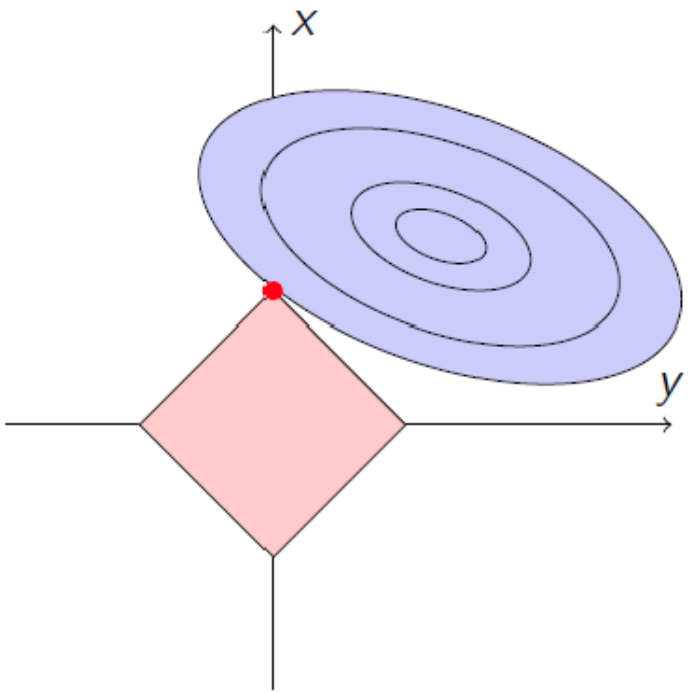
Two-dimensional case:

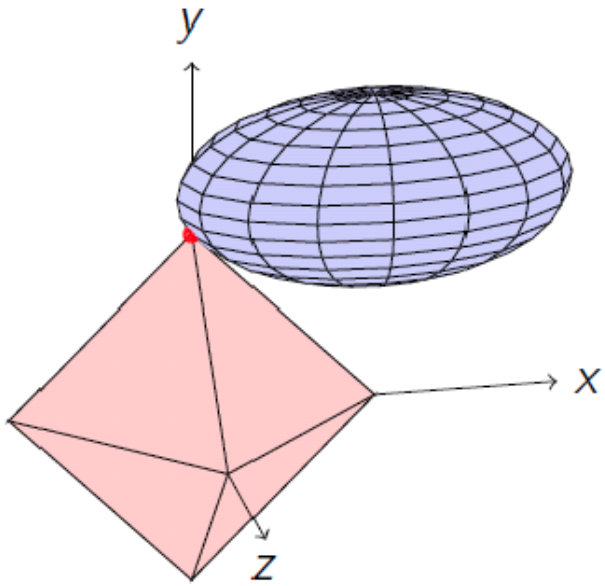
- contour plots for d = 2



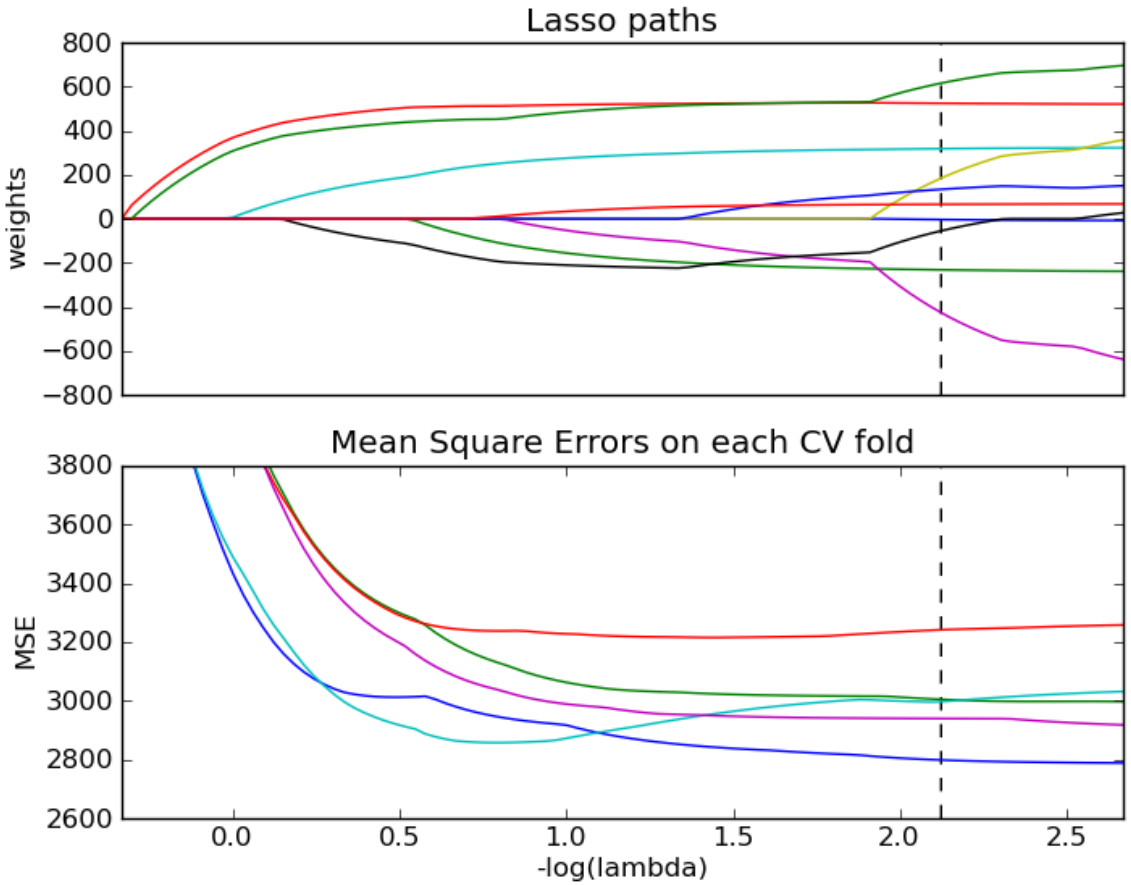
Comparison of Ridge and Lasso regression:

Higher dimensional case:





Choosing λ



- Standard practice now is to use cross-validation

Outline

Probabilistic Intepretation of Linear regression

Assume $y_i = X_i\theta + \epsilon_i$, where ϵ is the random noise.
Assume $\epsilon \sim \mathcal{N}(0, \sigma^2)$

$$p(y_i|X_i;\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{(y_i - X_i\theta)^2}{2\sigma^2}\}$$

Since data points are i.i.d, we have the data likelihood

$$L(\theta) = \prod_{i=1}^{N} p(y_i|X_i;\theta) \propto \exp\{-\frac{\sum_{i=1}^{N}(y_i - X_i\theta)^2}{2\sigma^2}\}$$

The log likelihood is:

$$\ell(\theta) = -\frac{\sum_{i=1}^{N}(y_i - X_i\theta)^2}{2\sigma^2} + \text{const}$$

Maximizing the log-likelihood is equivalent to minimize $\sum_{i=1}^{N}(y_i - X_i\theta)^2$, i.e., the loss function in LR!

Probabilistic Intepretation of Ridge regression

Assume a Gaussian prior on $\theta \sim \mathcal{N}(0, \tau^2 I)$, i.e.,

$$p(\theta) \propto \exp\{-\theta^\top \theta / 2\tau^2\}$$

Now get the MAP estimate of θ

$$p(\theta|X, y) \propto p(y|X; \theta)p(\theta) = \exp\{-\frac{\sum_{i=1}^{N}(y_i - X_i\theta)^2}{2\sigma^2}\} \exp\{-\theta^\top \theta / 2\tau^2\}$$

The log likelihood is:

$$\ell(\theta|X, y) = -\frac{\sum_{i=1}^{N}(y_i - X_i\theta)^2}{2\sigma^2} - \theta^\top \theta / 2\tau^2 + \text{const}$$

which matches $\min_\theta \frac{1}{2}\sum_i (y_i - X_i\theta)^2 + \lambda||\theta||_2^2$, where λ is a constant associated with σ^2, τ^2.

Probabilistic Intepretation of Lasso regression

Assume a Laplace prior on $\theta_i \overset{iid}{\sim}$ Laplace$(0, t)$, i.e.,

$$p(\theta_i) \propto \exp\{-|\theta_i|/t\}$$

Now get the MAP estimate of θ

$$p(\theta|X, y) \propto p(y|X; \theta)p(\theta) = \exp\{-\frac{\sum_{i=1}^{N}(y_i - X_i\theta)^2}{2\sigma^2}\} \exp\{-\sum_i |\theta_i|/t\}$$

The log likelihood is:

$$\ell(\theta|X, y) = -\frac{\sum_{i=1}^{N}(y_i - X_i\theta)^2}{2\sigma^2} - \sum_i |\theta_i|/t + \text{const}$$

which matches $\min_\theta \frac{1}{2}\sum_i(y_i - X_i\theta)^2 + \lambda||\theta||_1$, where λ is a constant associated with σ^2, t.

Outline

Variable Selection

- Consider "best" subsets, order $O(2^P)$ (combinatorial explosion)
- Stepwise selection
  - A new variable may be added into the model even with a small improvement in LMS
  - When applying stepwise to a perturbation of the data, probably have different set of variables enter into the model at each stage
- LASSO produces sparse solutions, which takes care of model selection
  - we can even see when variables jump into the model by looking at the LASSO path

Outline

Example

Suppose you have data $Y_1, ..., Y_n$ and you want to model the distribution of Y. Some popular models are:

- the Exponential distribution: $f(y; \theta) = \theta e^{-\theta y}$
- the Gaussian distribution: $f(y; u, \sigma^2) \sim \mathcal{N}(u, \sigma^2)$
- ...

How do you know which model is better?

AIC

Suppose we have models $M_1, ..., M_k$ where each model is a set of densities:

$$M_j = \{p(y; \theta_j) : \theta_j \in \Theta_j\}$$

We have data $Y_1, ..., Y_n$ drawn from some density f (not necessarily drawn from these models). Define

$$\text{AIC}(j) = \ell_j(\hat{\theta}_j) - 2d_j$$

where $\ell_j(\theta_j)$ is the log-likelihood, and $\hat{\theta}_j$ is the parameter that maximizes the log-likelihood. d_j is the dimension of Θ_j.

BIC

Bayesian Information Criterion We choose j to maximize

$$\text{BIC}_j = \ell_j(\hat{\theta}_j) - \frac{d_j}{2} \log n$$

which is similar to AIC but the penalty is harsher, hence BIC tends to choose simpler models.

Simple example

Let

$$Y_1, ..., Y_n \sim \mathcal{N}(\mu, 1)$$

we want to compare two model:

$$M_0 : \mathcal{N}(0, 1) \text{ and } M_1 : \mathcal{N}(u, 1)$$

Simple example: AIC

The log-likelihood is

$$\ell = \log \prod_i e^{-(Y_i-u)^2/2} = -\sum_i (Y_i - u)^2/2$$

$$\text{AIC}_0 = -\sum_i Y_i^2/2 - 0$$

$$\text{AIC}_1 = -\sum_i (Y_i - \bar{Y})^2/2 - 2 = -\sum_i Y_i^2/2 + \frac{n}{2}\bar{Y}^2 - 2$$

we choose model 1 if $\text{AIC}_1 > \text{AIC}_0$ i.e.,

$$-\sum_i Y_i^2/2 + \frac{n}{2}\bar{Y}^2 - 2 > -\sum_i Y_i^2/2$$

or $\bar{Y} > \sqrt{\frac{4}{n}}$.

Simple example: BIC

$$\text{BIC}_0 = -\sum_i Y_i^2/2 - \frac{0}{2}\log n = -\sum_i Y_i^2/2$$

$$\text{AIC}_1 = -\sum_i (Y_i - \bar{Y})^2/2 - \frac{1}{2}\log n = -\sum_i Y_i^2/2 + n/2\,\bar{Y}^2 - \frac{1}{2}\log n$$

we choose model 1 if $\text{BIC}_1 > \text{BIC}_0$ i.e.,

$$-\sum_i Y_i^2/2 + n/2\,\bar{Y}^2 - \frac{1}{2}\log n > -\sum_i Y_i^2/2$$

or $\bar{Y} > \sqrt{\frac{\log n}{n}}$.

Comparison

Generally speaking,

- AIC/CV finds the most predictive model
- BIC find the true model with high probability, i.e., BIC assumes that one of the models is true and that you are trying to find the model most likely to be true in the Bayesian sense.