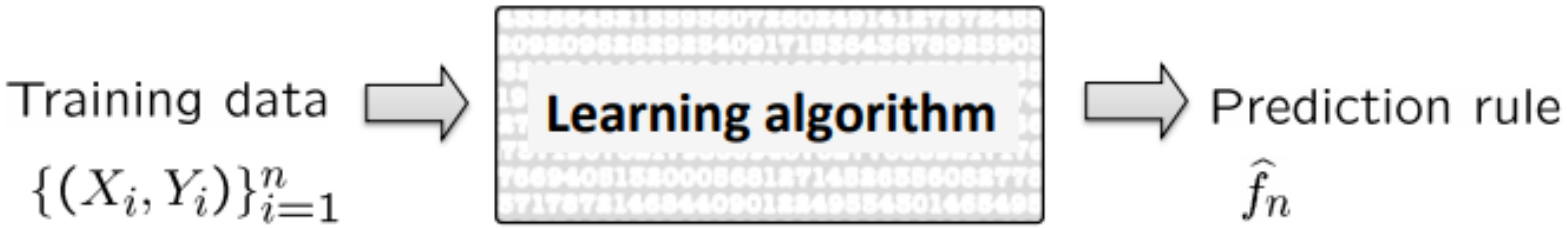
Linear Regression

Xiangli Chen

Regression algorithm



- Linear regression
- Regularized linear regression – ridge, lasso regression
- Polynomial regression
- Kernel regression
- Regression trees, splines, wavelet estimators, …

Restrict class of predictors

Optimal predictor: $f^* = \arg\min_f \mathbb{E}[(f(X) - Y)^2]$

Empirical Minimizer: $\hat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2$

Class of predictors

Law of Large Numbers:

$$\frac{1}{n} \sum_{i=1}^{n} [\text{loss}(Y_i, f(X_i))] \xrightarrow{n \to \infty} \mathbb{E}_{XY} [\text{loss}(Y, f(X))]$$

Restrict class to avoid overfitting

- Class of linear functions
- Class of polynomial functions
- Class of nonlinear functions

Linear regression

least Square Estimator

$$\hat{f}_{LIN} = \underset{f \in \mathcal{F}_L}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2$$

$\mathcal{F}_L$ — class of linear functions

Uni-variate case:

$$f(x) = \beta_0 + \beta_1 x$$

β_0: intercept β_1: slope

Multi-variate case:

$$f(X) = \beta^T x \quad \text{where} \quad x = [1, x_1, \ldots, x_p], \ \beta = [\beta_0, \ldots, \beta_p]^T$$

Least squares estimator

$$\widehat{f}_n^L = \arg\min_{f \in \mathcal{F}_L} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2 \qquad f(X_i) = X_i\beta$$

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (X_i\beta - Y_i)^2 \qquad \widehat{f}_n^L(X) = X\widehat{\beta}$$

$$= \arg\min_{\beta} \frac{1}{n} (X\beta - y)^T (X\beta - y)$$

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} = \begin{bmatrix} 1, x_{11}, \ldots, x_{1p} \\ \vdots \\ 1, x_{n1}, \ldots, x_{np} \end{bmatrix} \qquad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

Least Square Estimator

Let $J(\beta) = (X\beta - y)^T (X\beta - y)$

So $\hat{\beta} = \arg\min_{\beta} J(\beta)$

$$\left.\frac{\partial J(\beta)}{\partial \beta}\right|_{\hat{\beta}=0} \Rightarrow (X^T X)\hat{\beta} = X^T y$$

If $X^T X$ is invertible $\hat{\beta} = X(X^T X)X^T y$

- $X^T X$ is invertible, when $X^T X$ has full rank
- If $X^T X$ is not invertible, try regularization, gradient descent learning method

Geometric Interpretation of LS

$$\hat{y} = \hat{f}_{LR}(X) = X\hat{\beta} = X(X^T X)^{-1} X^T y$$

$\hat{y}$ is a linear combination of columns of X

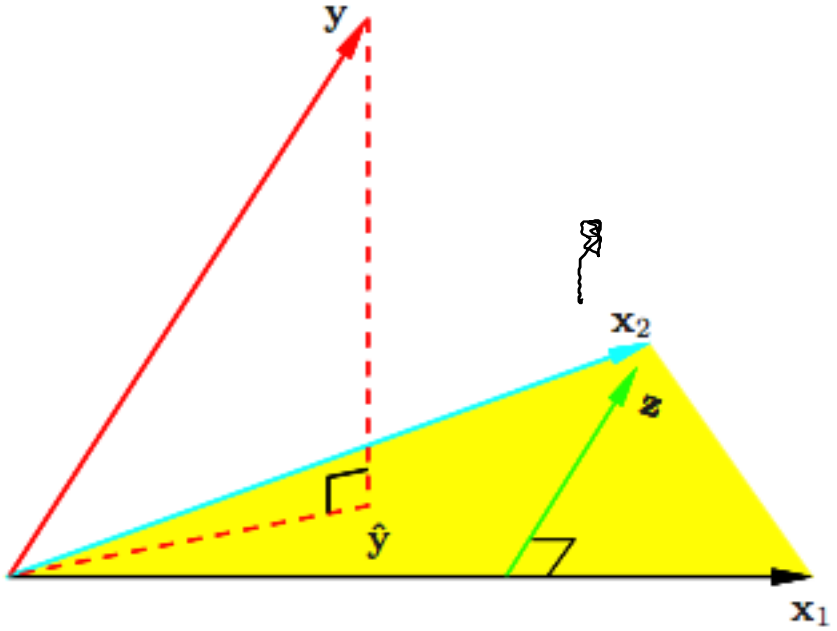
on the space spanned by those columns

Note that

$$X^T(y - \hat{y}) = X^T y - X^T X (X^T X)^{-1} X^T y = 0$$

So $y - \hat{y}$ is orthogonal to the space

$\hat{y}$ is the orthogonal project of y into the space



Gradient Descent

When $X^T X$ is invertible or computationally expensive if X is huge

$$\hat{\beta} = \arg\min_{\beta} J(\beta)$$

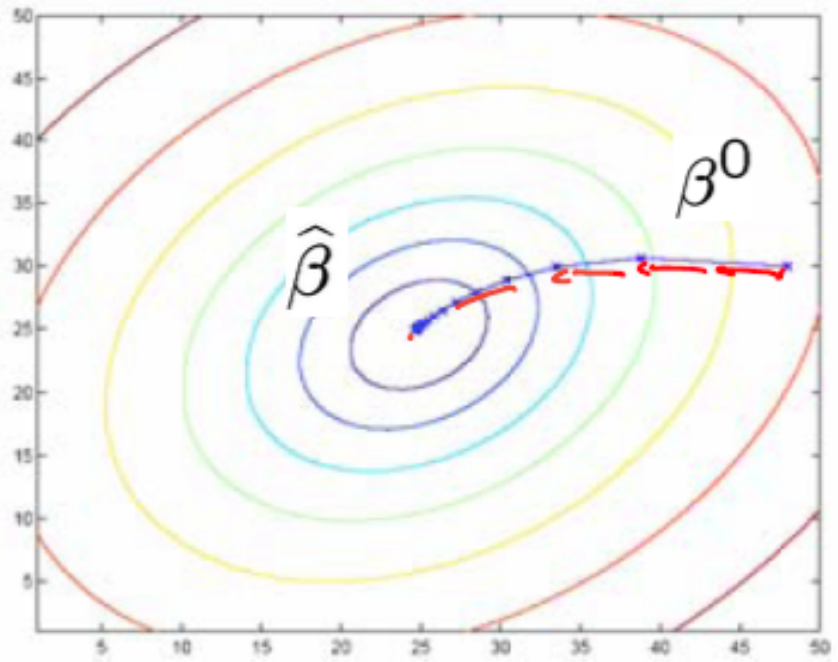
$J(\beta)$ is convex with respect to β

Gradient descent step

$$\beta^{t+1} = \beta^t - \alpha \frac{\partial J(\beta)}{\partial \beta}$$

$$= \beta^t - 2\alpha X^T (X\beta^t - y)$$

α is the step size



Gaussian Assumption and Likelihood

$$y = \beta^T x + \varepsilon \quad \varepsilon \sim N(0, \sigma^2) \quad Y \mid X = x \sim N(\beta^T x, \sigma^2)$$

Given random sample (i.i.d) $\{x_i, y_i\}_{i=1}^N$, conditional likelihood

$$f(y_{i=1}^N \mid x_{i=1}^N, \beta, \sigma^2) = \prod_{i=1}^N f(y_i \mid x_i, \beta, \sigma^2)$$

Maximize conditional log likelihood (minimize empirical conditional log loss)

$$\hat{\beta}_{MLE} = \arg\max_\beta \sum_{i=1}^N \log f(y_i \mid x_i, \beta, \sigma^2) \quad \text{(discriminative approach)}$$

$$= \arg\min_\beta \sum_{i=1}^N (\beta^T x_i - y_i)^2 = \hat{\beta}$$

Least square estimation is same as maximize conditional log likelihood estimate under a Gaussian model

Shrinkage Methods

Ridge regression – regularized by L_2 penalty

$$\hat{\beta}_{ridge} = \underset{\beta}{\arg\min} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \qquad \lambda \geq 0$$

Note that β_0 has been left out of the penalty term

- Adding c to $y_i's$, β_0 increasing by c as well and also predicted values $\hat{y}_i's$. It is not true if β_0 is penalized
- β_0 affects the shift, not the complexity

Same Problem

Re-express the ridge regression

$$\sum_{i=1}^{N}\left(y_i-\beta_0^c-\sum_{j=1}^{p}(x_{ij}-\bar{x}_j)\beta_j^c\right)^2+\lambda\sum_{j=1}^{p}\beta_j^{c\,2}$$

where $\beta_0^c=\beta_0+\sum_{j=1}^{p}\bar{x}_j\beta_j \qquad \beta_j^c=\beta_j \quad j=1,2,\cdots,p$

Note that $\hat{\beta}_0^c=\bar{y} \qquad \hat{\beta}_j^c=\hat{\beta}_j$

Let $y_i=y_i-\bar{y}$ and $x_{ij}=x_{ij}-\bar{x}_j$ gives ridge regression without intercept

$$\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2 \Longleftrightarrow (y-X\beta)^T(y-X\beta)+\lambda\beta^T\beta$$

$\hat{\beta}_{ridge}=(X^TX+\lambda I)^{-1}X^Ty$ $\quad (X^TX+\lambda I)$ always has full rank

SVD Decomposition

MAP Interpretation

Maximize a posterior (MAP)

$$f(\beta \mid y_{i=1}^N, x_{i=1}^N, \sigma^2) \propto f(y_{i=1}^N \mid \beta, x_{i=1}^N, \sigma^2) f(\beta \mid x_{i=1}^N, \sigma^2)$$

Gaussian Prior

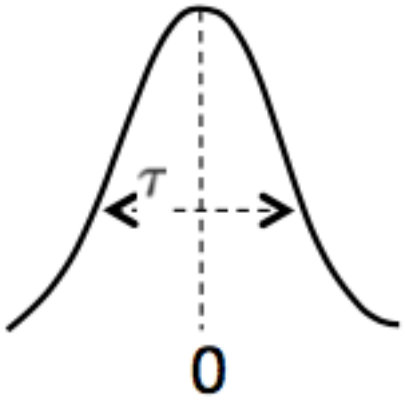
$$\beta = N(0, \tau^2 I) \qquad f(\beta) \propto e^{-\beta^T \beta / 2\tau^2}$$

$$\hat{\beta}_{MAP} = \arg\max_{\beta} \log f(y_{i=1}^N \mid \beta, x_{i=1}^N, \sigma^2) f(\beta)$$

$$= \arg\min_{\beta} \sum_{i=1}^N (y_i - \beta^T x_i)^2 + \left(\frac{\sigma}{\tau}\right)^2 \|\beta\|_2^2$$

Let $\lambda = \left(\frac{\sigma}{\tau}\right)^2$, the expression is the same to ridge regression

$$\hat{\beta}_{MAP} = \hat{\beta}_{ridge}$$



Ridge Regression

Lasso

Laplace prior

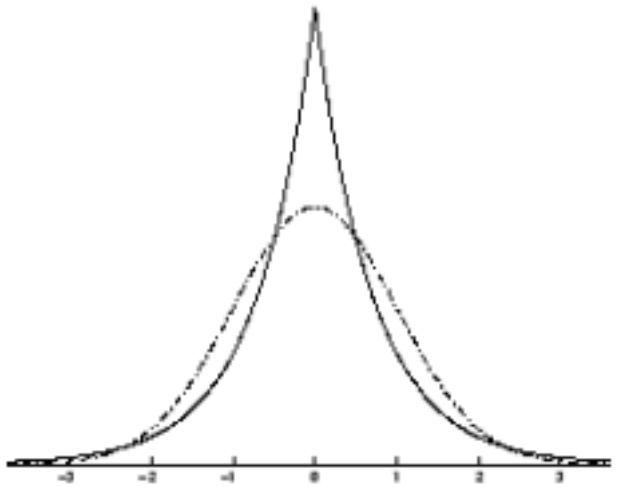
$\beta \sim \text{Laplace}(0, t)$ $\quad f(\beta) \propto e^{-|\beta|/t}$

MAP estimator

$$\hat{\beta}_{MAP} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \beta^T x_i)^2 + \frac{\sigma^2}{t} \|\beta\|_1$$

Let $\lambda = \frac{\sigma^2}{t}$, the expression is Lasso (regularized with L1 term)

No closed form solution — a quadratic programming problem



Lasso

Prime Problem

Prime problem

$$\hat{\beta} = \arg\min_{\beta} J(\beta)$$

subject to $pen(\beta) \le \eta$

$\Rightarrow$ Lagrangian form

$$\hat{\beta} = \arg\min_{\beta} J(\beta) + \lambda pen(\beta)$$

Ridge: $\|\beta\|_2^2 \le \eta$ Lasso: $\|\beta\|_1 \le \eta$

$\|\beta\|_p \le \eta$ gives convex set of β

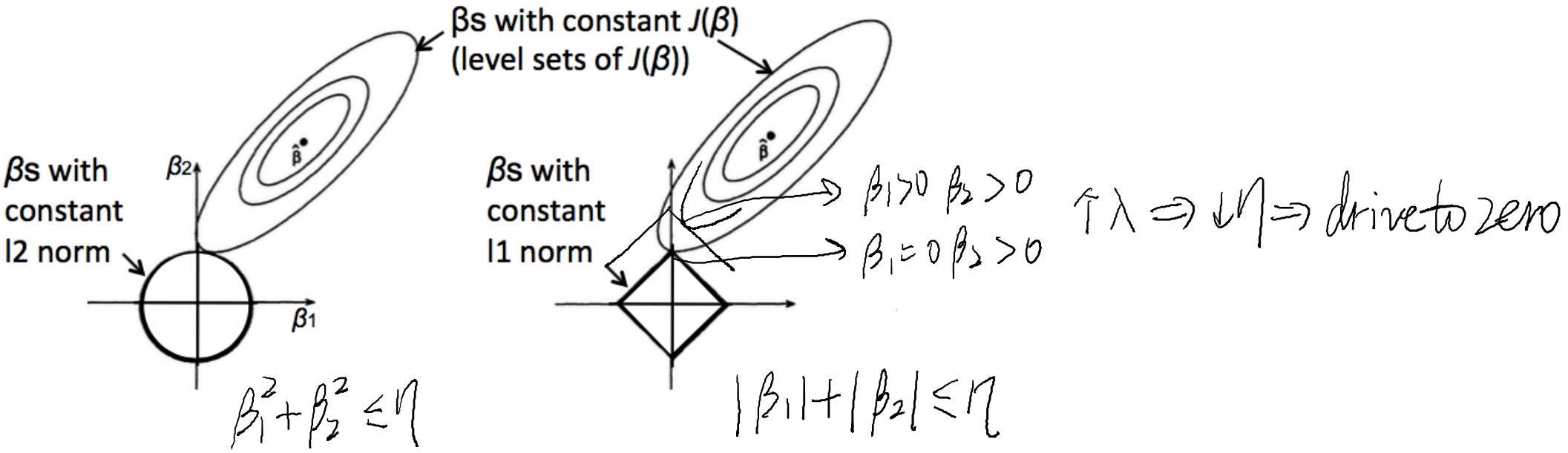
- There is a one to one correspondence between λ and η.
- When λ increases, η decreases, an increasing number of coordinates are driven to zero.

Ridge regression vs lasso

Ridge Regression:
$\text{pen}(\beta) = \|\beta\|_2^2$

Lasso:
$\text{pen}(\beta) = \|\beta\|_1$

HOT!



Lasso results in sparse solution – more zero coordinates, good for high dimension

Polynomial Regression

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_m x^m$$

$$X = \begin{bmatrix} 1 & x_1 & \cdots & x_1^m \\ \vdots & \vdots & & \\ 1 & x_n & \cdots & x_n^m \end{bmatrix} \qquad \hat{\beta} = (X^T X)^{-1} X^T y$$