DDA3020 Tutorial 4
Linear Regression

Rongxiao Qu
School of Data Science
Email: rongxiaoqu@link.cuhk.edu.cn
Office hour: Tue 10:30 - 11:30, by appointment
Date: 2022.10.11

Contents

"Linear" Regression

- A linear combination of the input features
- $f(\boldsymbol{x}) = w_0 + \boldsymbol{x}^T \boldsymbol{w}$ $\quad (f_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{X}\,\boldsymbol{w})$
- $f(\boldsymbol{x}) = w_0 + \sum_{j=1}^{p} w_j x_j$

$x_1 \quad x_2$

- Have advantage when the data size is small: avoid overfitting
- But it imposes significant limitations on the model

Feature Transformation with Basis Functions

- $f_{\{\boldsymbol{w},b\}}(\boldsymbol{x}) = \sum_{j=1}^{p} w_j \phi(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x})$
- $(\boldsymbol{w} = (w_1, \dots, w_n)^T \ \boldsymbol{\phi} = \left(\phi_1, \dots, \phi_p\right))$
- Polynomial Regressions

$\phi(x) = x_1^2 \quad x_1^3 \quad x_1 x_2$

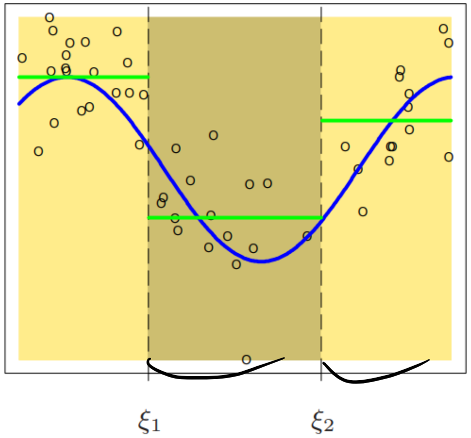
- *Gaussian Basis Function*: $\phi_j(x) = \exp\left\{-\frac{\left(x-\mu_j\right)^2}{2s^2}\right\}$
- *Sigmoid Basis Function*: $\phi_j(x) = \sigma(\frac{x-\mu_j}{s})$
  (*logistic sigmoid function*: $\sigma(a) = \frac{1}{1+\exp(-a)}$)
- *Splines* (*piecewise polynomials*)
- $f(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 (x - \xi_1)_+^3 + w_5 (x - \xi_2)_+^3$

- Splines:

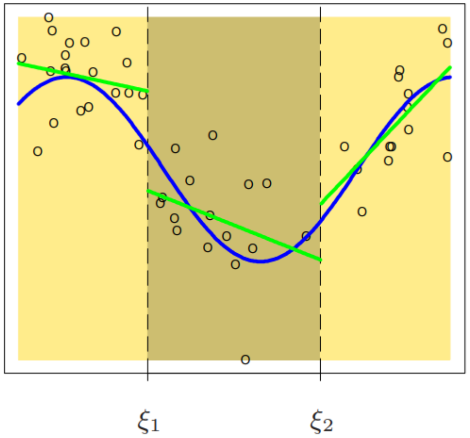
$$(x-\xi_1)_+ = \begin{cases} x-\xi_1 & x > \xi_1 \\ 0 & x \le \xi_1 \end{cases}$$

$$f_w(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 (x-\xi_1)_+^3 + w_5 (x-\xi_2)_+^3$$

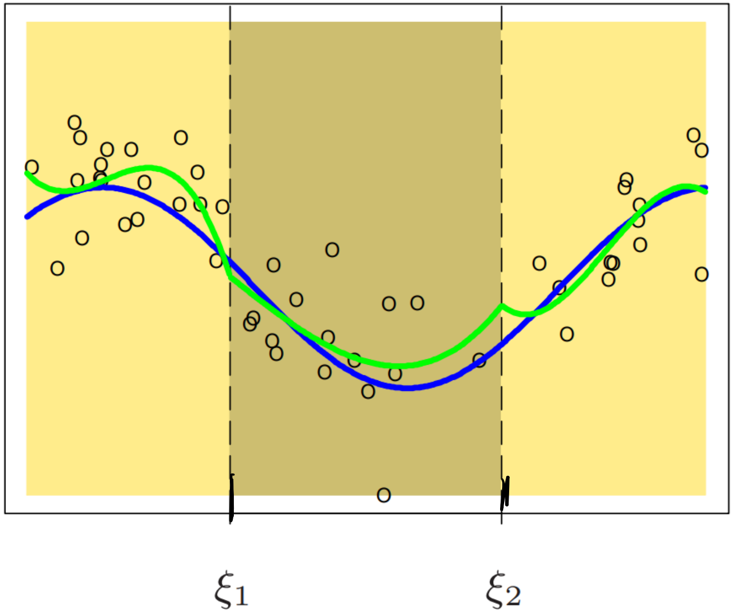
Piecewise Constant



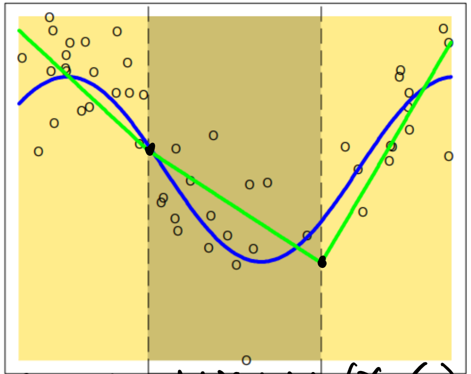
Piecewise Linear



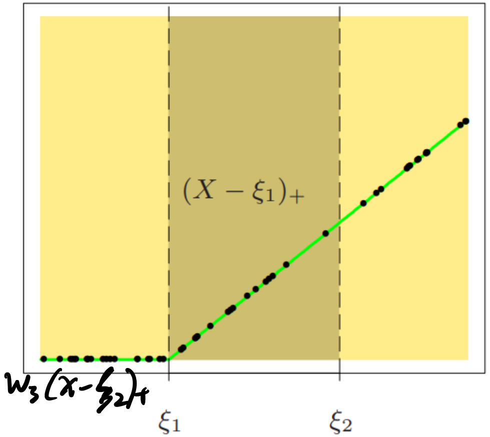
Cubic Splines



Continuous Piecewise Linear



Piecewise-linear Basis Function



$$f_w(x) = w_0 + w_1 x + w_2 (x-\xi_1)_+ + w_3 (x-\xi_2)_+$$

ξ_1 ξ_2

Least Squares Regression

- Minimizing the squared error:
- $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{w}}$
- $\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmin}}\, RSS = \underset{\boldsymbol{w}}{\operatorname{argmin}} \sum_{i=1}^{n}(f_{\boldsymbol{w}}(\boldsymbol{x}) - y)^2$
  $= \underset{\boldsymbol{w}}{\operatorname{argmin}}(\boldsymbol{Xw} - \boldsymbol{y})^{\boldsymbol{T}}(\boldsymbol{Xw} - \boldsymbol{y})$

- Take derivative w.r.t **w** and set the derivative to be 0 →
- $\hat{\boldsymbol{w}} = \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-\mathbf{1}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$
- $\hat{\boldsymbol{y}} = \boldsymbol{X}_{\boldsymbol{new}}\hat{\boldsymbol{w}} = \boldsymbol{X}_{\boldsymbol{new}}\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-\mathbf{1}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$

Properties of the LS estimator: $\hat{\boldsymbol{w}}$

For $\boldsymbol{y} = \boldsymbol{f_w}(\boldsymbol{x}) + \boldsymbol{\epsilon} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}$

- Assumptions: ① $f_w(x) \to$ true ② $\varepsilon \sim$ not correlated with y / x
  - $E(\boldsymbol{\epsilon}) = 0$ (Mean of the errors are zeros)
  - $Cov(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}$ (errors are uncorrelated with equal variance) (usually hard to satisfy)

- Conclusions:
  - $E(\hat{\boldsymbol{w}}) = \boldsymbol{w}$
  - $Cov(\hat{\boldsymbol{w}}) = \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$ $\to$ conduct tests of significance for $\mathrm{w}_\mathrm{i}'\mathrm{s}$
  - ($\hat{\boldsymbol{w}}$ is the best linear unbiased estimator (BLUE) of $\boldsymbol{w}$)

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

$$= E((X^T X)^{-1} X^T (Xw + e))$$

$$= E((X^T X)^{-1} X^T X w + (X^T X)^{-1} X^T e)$$

$$= E(w) + (X^T X)^{-1} X^T \underbrace{E(e)}_{=0}$$

$$= w$$

$$Cov(\hat{w}) = E((\hat{w} - E(\hat{w}))(\hat{w} - E(\hat{w}))^T)$$

$$= E(((X^T X)^{-1} X^T (Xw + e) - w)((X^T X)^{-1} X^T (Xw + e) - w)^T)$$

$$= E(((X^T X)^{-1} X^T e)((X^T X)^{-1} X^T e)^T)$$

$$= E((X^T X)^{-1} X^T e e^T X (X^T X)^{-1})$$

$$= (X^T X)^{-1} X^T \underbrace{E(ee^T)}_{= Cov(e)} X (X^T X)^{-1}$$

$$= 6^2 (X^T X)^{-1} X^T X (X^T X)^{-1}$$

$$= 6^2 (X^T X)^{-1}$$

Ridge Regression

- $\hat{\beta}^{ridge} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$

  $= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ (\boldsymbol{X\beta} - \boldsymbol{y})^T (\boldsymbol{X\beta} - \boldsymbol{y}) + \lambda ||\boldsymbol{\beta}||^2 \right\})$

- Equivalently:

- $\hat{\beta}^{ridge} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right\}$

  $(= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \{ (\boldsymbol{X\beta} - \boldsymbol{y})^T (\boldsymbol{X\beta} - \boldsymbol{y}) \}$

  subject to $\sum_{j=1}^{p} \beta_j^2 \le t$

  (subject to $||\boldsymbol{\beta}||^2 \le \boldsymbol{t}$)

$$RSS = (Xw - y)^T (Xw - y) + \lambda w^T w$$

$$= w^T X^T X w - w^T X^T y - y^T X w + y^T y + \lambda w^T w$$

$$\frac{\partial RSS}{\partial w} = 2X^T X w - X^T y - X^T y + 2\lambda w = 0$$

$$X^T X w + \lambda w = X^T y$$

$$(X^T X + \lambda I) w = X^T y$$

$$\hat{w}_{ridge} = (X^T X + \lambda I)^{-1} X^T y$$

Lasso

- $\hat{\beta}^{ridge} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$

$$\left(= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \{ (\boldsymbol{X\beta} - \boldsymbol{y})^T (\boldsymbol{X\beta} - \boldsymbol{y}) + \lambda |\boldsymbol{\beta}|_1 \} \right)$$

- Equivalently:
- $\hat{\beta}^{ridge} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right\}$

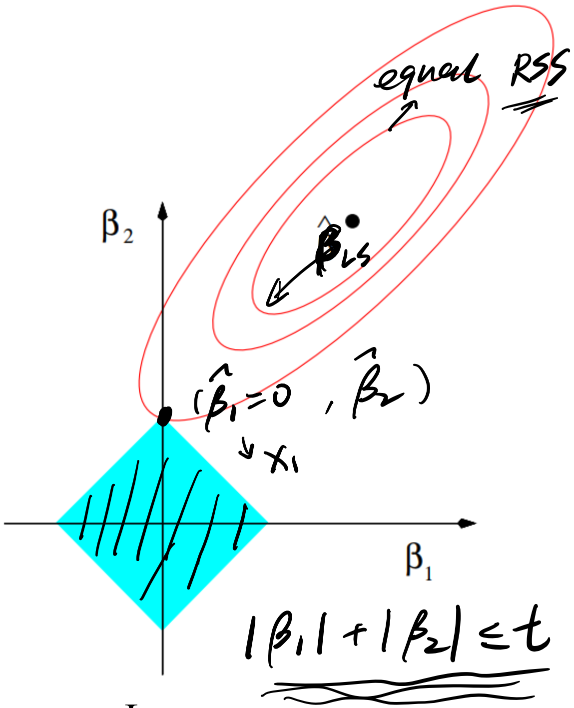
$$\left(= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \{ (\boldsymbol{X\beta} - \boldsymbol{y})^T (\boldsymbol{X\beta} - \boldsymbol{y}) \} \right)$$

subject to $|\boldsymbol{\beta}|_1 \leq t$

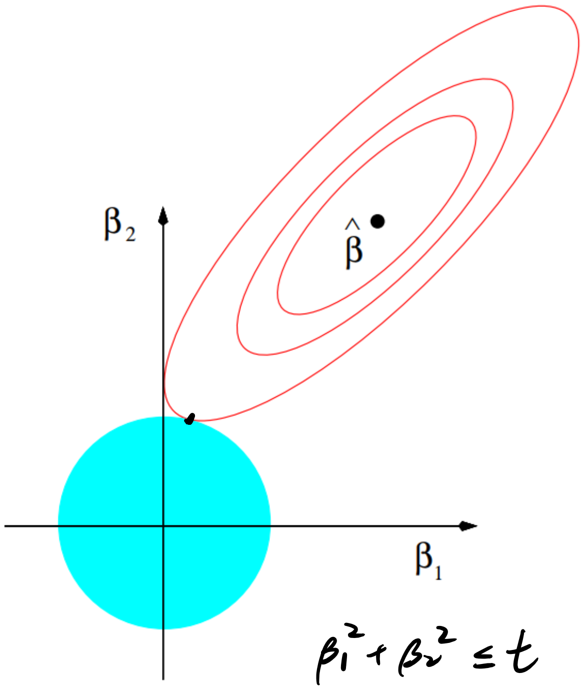
(subject to $|\boldsymbol{\beta}|_1 \leq \boldsymbol{t}$)

Geometry of Ridge and Lasso regression

$\beta = (\beta_1, \beta_2)$



Lasso



Ridge

Code Demo

Data Processing

Ridge Regression and Lasso