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

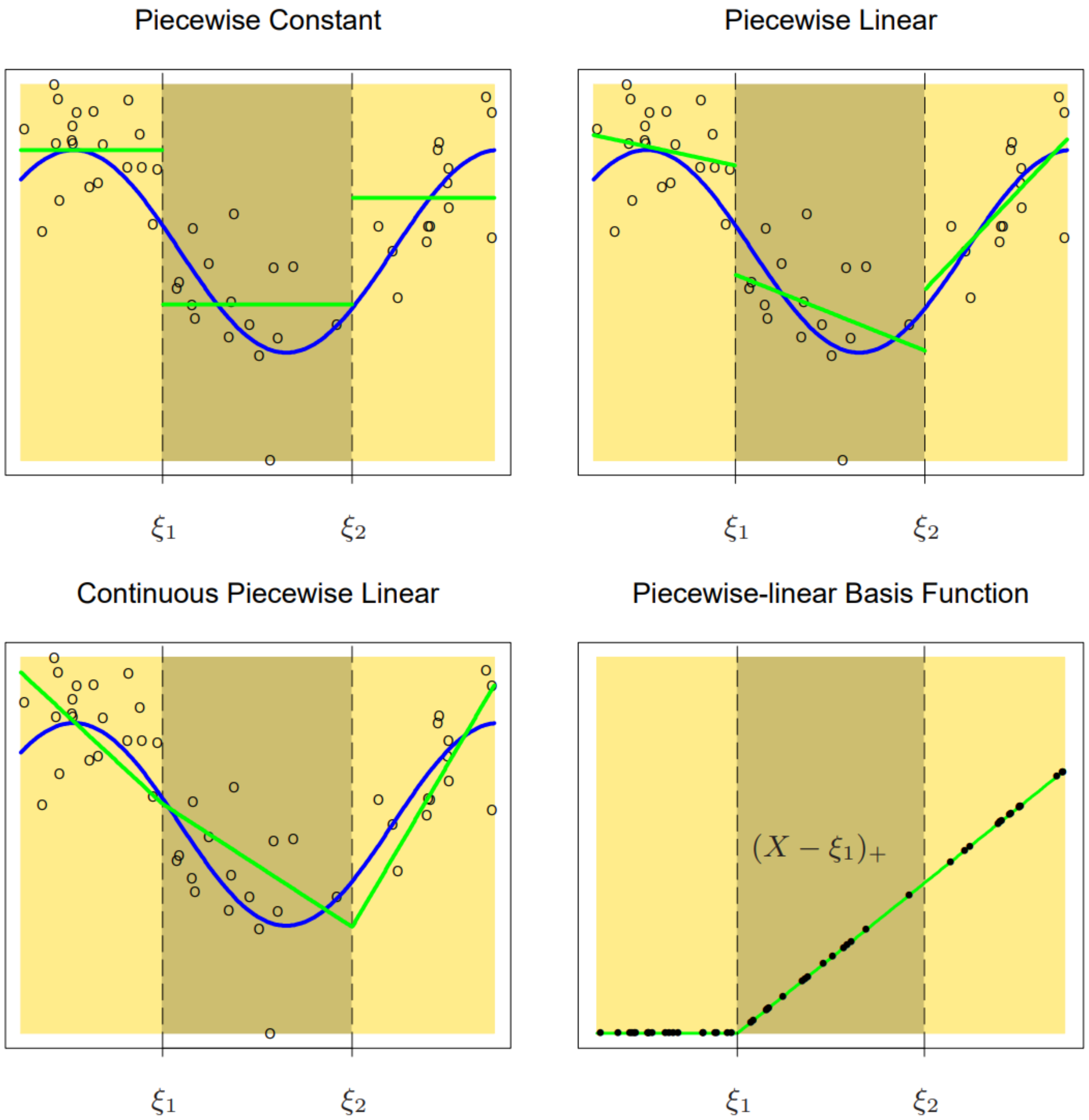
- A linear combination of the input features
- $f(\boldsymbol{x}) = w_0 + \boldsymbol{x}^T\,\boldsymbol{w}$ $\qquad (f_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{X}\,\boldsymbol{w})$
- $f(\boldsymbol{x}) = w_0 + \sum_{j=1}^{p} w_j x_j$

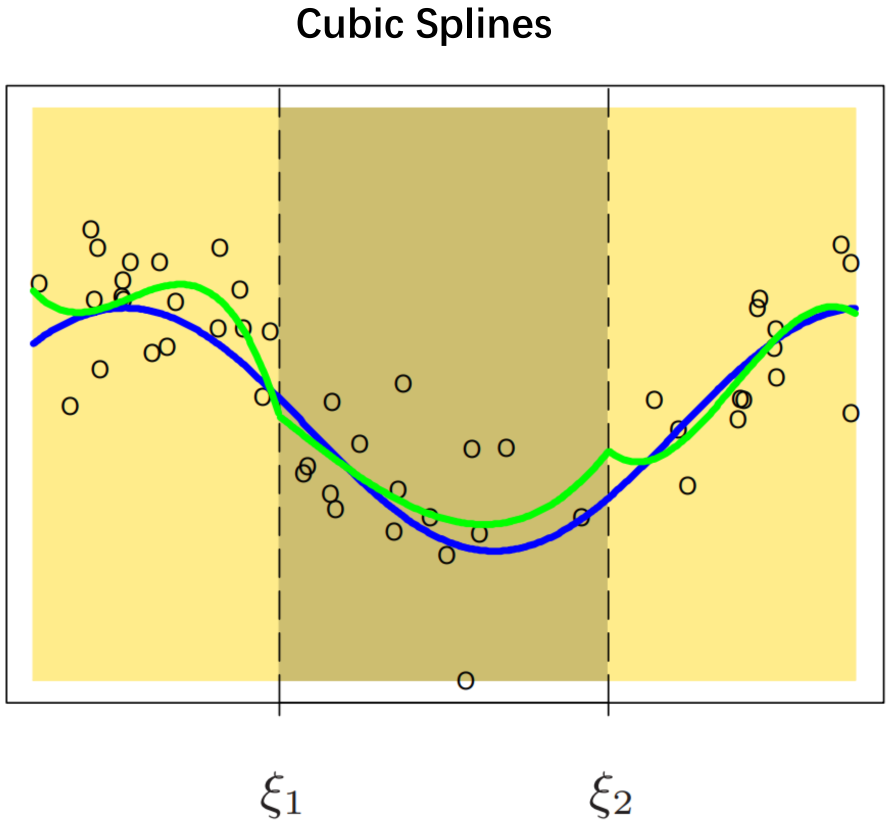
- Have advantage when the data size is small: avoid overfitting
- But it imposes significant limitations on the model

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

- Splines:


Piecewise Constant
Piecewise Linear
Continuous Piecewise Linear
Piecewise-linear Basis Function
$(X - \xi_1)_+$
ξ_1
ξ_2


Cubic Splines
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 $\boldsymbol{y} = \boldsymbol{f}_{\boldsymbol{w}}(\boldsymbol{x}) + \boldsymbol{\epsilon} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}$

- Assumptions:
  - $E(\boldsymbol{\epsilon}) = 0$ (Mean of the errors are zeros)
  - $Cov(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}$ (errors are uncorrelated with equal variance)
    (usually hard to satisfy)

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

Ridge Regression

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

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

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

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

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

  (subject to $\left| |\boldsymbol{\beta}| \right|^{\boldsymbol{2}} \leq \boldsymbol{t}$)

Lasso

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

$$\left(= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \{ (\boldsymbol{X\beta} - \boldsymbol{y})^{\boldsymbol{T}} (\boldsymbol{X\beta} - \boldsymbol{y}) + \lambda |\boldsymbol{\beta}|_{\mathbf{1}} \} \right)$$

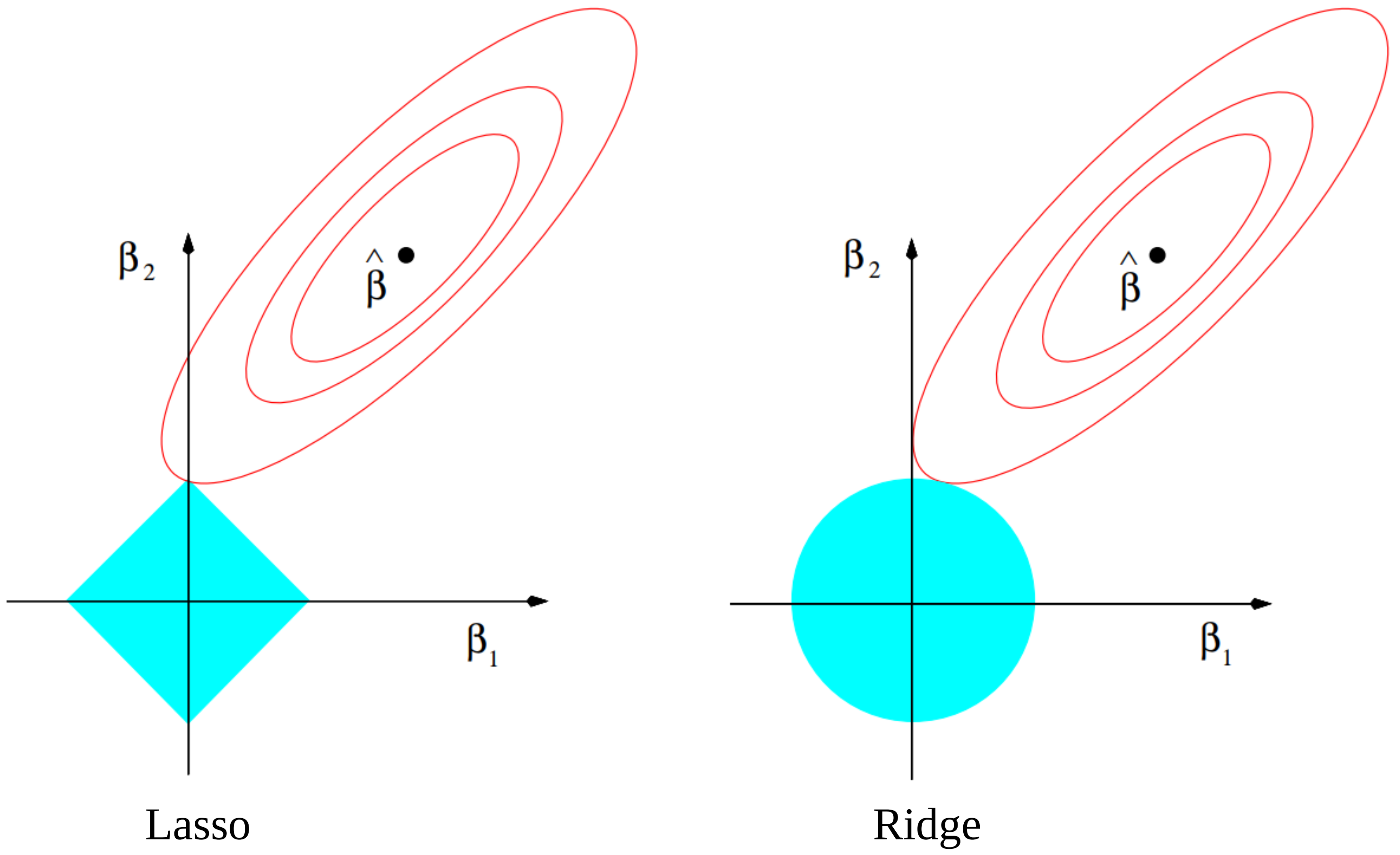
- Equivalently:
- $\hat{\beta}^{ridge} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} y_i \; - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right\}$

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

subject to $|\boldsymbol{\beta}|_{\mathbf{1}} \leq t$

(subject to $|\boldsymbol{\beta}|_{\mathbf{1}} \leq \boldsymbol{t}$)

Geometry of Ridge and Lasso regression



Code Demo

Data Processing

Ridge Regression and Lasso