Lecture October 18

Ordinary Least Squares (OLS)

- input data + output data

$$D = \{ [x_0, y_0], [x_1, y_1] \ldots [x_{n-1}, y_{n-1}] \}$$

target data

- Model

- Assessment of the quality of the model

Basic assumption in linear Regression

$$y = f(x) + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

$\mu = 0$ variance

$\sigma^2 = 1$

$f(x)$ = continuous function

non-stochastic variable

$f(x) \simeq \tilde{y}(x)$ our model

polynomial expansion

$$\tilde{y}(x_i) = \tilde{y}_i = \sum_{j=0}^{p-1} \beta_j x_i^j$$

$$= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots \beta_{p-1} x_i^{p-1}$$

$$\tilde{y}(x_0) = \tilde{y}_0 = \beta_0 + \beta_1 x_0 + \beta_2 x_0^2 + \dots + \beta_{p-1} x_0^{p-1}$$

$$= \beta_0 x_{00} + \beta_1 x_{01} + \beta_2 x_{02} + \dots + \beta_{p-1} x_{0\,p-1}$$

$1 =$

$\vdots$

$$\tilde{y}(x_{n-1}) = \tilde{y}_{n-1} = \beta_0 x_{n-1\,0} + \beta_1 x_{n-1\,1} + \dots + \beta_{p-1} x_{n-1\,p-1}$$

$$\tilde{y} = [\tilde{y}_0, \tilde{y}_1, \dots \tilde{y}_{n-1}]^T \in \mathbb{R}^n$$

$$\beta = [\beta_0, \beta_1, \dots \beta_{p-1}]^T \in \mathbb{R}^p$$

X = Design matrix

$$= \begin{bmatrix} x_{00} & x_{01} & \cdots & x_{0p-1} \\ x_{10} & x_{11} & \cdots & x_{1p-1} \\ & & \vdots & \\ x_{n-10} & x_{n-11} & \cdots & x_{n-1p-1} \end{bmatrix} \in \mathbb{R}^{n \times p}$$

every row represents an input x_i

every column represents a given feature (here a polynomial degree)

Binding energies

$$D = \{(x_0, y_0) \cdots (x_{n-1}, y_{n-1})\}$$

(x_0, y_0) could be ^4He $A = 4$ (x_0)

$y_0 = BE(^4He)$

(x_1, y_1)

$$\rightarrow \quad SL_i = x_i$$

$$y_i = BE(SL_i)$$

$$X = \begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^{p-1} \\ 1 & x_1 & x_1^2 & & \vdots \\ 1 & x_2 & x_2^2 & & \vdots \\ 1 & \vdots & \vdots & & \vdots \\ \vdots & & \vdots & & \\ 1 & x_{n-1} & x_{n-1}^2 & & x_{n-1}^{p-1} \end{bmatrix}$$

intercept

$$= \begin{bmatrix} \vdots & | & | & & | \\ x_0 & x_1 & x_2 & \cdots & x_{p-1} \\ | & | & | & & | \end{bmatrix}$$

$$x_0 = [x_{00} \ x_{10} \ x_{20} \cdots x_{n-1\,0}]^T$$

our model $\tilde{y} = X\beta$

(Design matrix
feature - i -)

Model is linear in β

- How to assess the model

$$MSE = \frac{1}{n}\sum_{i=0}^{n-1}(y_i - \tilde{y}_i)^2$$

cost/loss
risk/error
function

$$= \mathbb{E}[(y-\tilde{y})^2]$$

$$= \frac{1}{n}\sum_{i=0}^{n-1}\left(y_i - \sum_{j=0}^{p-1} x_{ij}\beta_j\right)^2$$

optimal $\beta = \hat{\beta}$

$$= \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{n}\sum_{i=0}^{n-1}\left(y_i - \sum_{j=0}^{p-1} x_{ij}\beta_j\right)^2$$

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

$$X^T X \in \mathbb{R}^{p\times p}$$

$$X^T \in \mathbb{R}^{p \times n} \quad X \in \mathbb{R}^{n \times p}$$

$$y \in \mathbb{R}^n$$

$$\beta \in \mathbb{R}^p \qquad n >> p$$

$$MSE = C(\beta) = \frac{1}{n}\sum_{i=0}^{n-1}\left(y_i - \sum_{j=0}^{p-1} x_{ij}\beta_j\right)^2$$

$$\frac{\partial C}{\partial \beta_j} = 0 = -\frac{2}{n}\sum_{i=0}^{n-1} x_{ij}\left(y_i - \sum_{j=0}^{p-1} x_{ij}\beta_j\right)$$

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

$$\frac{\partial C}{\partial \beta} = 0 = X^T(y - X\beta)$$

(multiplied away $-2/n$

$$X^T y = X^T X \beta \Rightarrow$$

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

$X \in \mathbb{R}^{n \times p}$ $\quad X^T \in \mathbb{R}^{p \times n}$

$X^T X \in \mathbb{R}^{p \times p}$

in supervised learning a common situation is

$n \gg p$

$(X^T X)^{-1}$ can have singular inverses. standard procedure is to implement SVD + pseudoinverse (next week).

statistics

$$\mathbb{E}[x] = \mu_x = \int_D x\, p(x)\, dx$$

$$\left(\sum_{i \in D} x_i\, p(x_i) \right)$$

$$\mathrm{var}[x] = \sigma^2 = \int_D (x - \mu_x)^2 p(x)\, dx$$

$$\mathrm{cov}[x_i, x_j] = \int_D (x_i - \mu_{x_i})(x_j - \mu_{x_j}) \times p(x_i)\, p(x_j)\, dx_i\, dx_j$$

in ML, we normally use a frequentist approach.
We don't know $p(x)$!

Sample expectation values

$$\mathbb{E}[x] = \frac{1}{n} \sum_{i=0}^{n-1} x_i \neq \mu_x = \sum x_i\, p(x_i)$$

$$= \bar{\mu}$$

$$p(x_i) \simeq \frac{1}{n}$$

$$\sigma^2 = \frac{1}{n} \sum_{i=0}^{n-1} (x_i - \bar{\mu})^2$$

$$\mathrm{cov}[x_i x_j] = \frac{1}{n} \sum_{\ell=0}^{n-1} (x_\ell^{(i)} - \mu_{x_i})(x_\ell^{(j)} - \mu_{x_j})$$

if we have iid
(= independent and identically distributed)

$$\mathrm{cov}[x_i, x_j] = 0$$

$$y = f(x) + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

$$f(x) \simeq X\beta$$

$$\mathbb{E}[y] = \mathbb{E}[X\beta] + \underbrace{\mathbb{E}[\varepsilon]}_{=0}$$

$$= X\beta = \mu_y$$

$$\mathrm{var}[y] = \mathbb{E}[(y - \mu_y)^2]$$

$$= \sigma^2, \text{ same as } \varepsilon \Rightarrow$$

$$y \sim N(X\beta, \sigma^2)$$

Exercise

$$\mathbb{E}[\hat{\beta}] = \hat{\beta} \quad \text{unbiased}$$

$$\mathrm{var}[\hat{\beta}] = \sigma^2 (X^T X)^{-1}$$

$$\mathrm{var}[\hat{\beta}_j] = \sigma^2 (X^T X)^{-1}_{jj}$$

Regularization

Ridge regression

$$C(\beta) = \frac{1}{n}(y - X\beta)^T(y - X\beta) + \lambda \underbrace{\sum_{j=0}^{p-1} \beta_j^2}_{\|\beta\|_2^2}$$

$$\lambda > 0$$

$$\sum_{j=0}^{p-1} \beta_j^2 < t$$

$$\frac{\partial C}{\partial \beta} = 0 = -\frac{2}{n} X^T(y - X\beta) - 2\lambda \sum_{j=0}^{p-1} \beta$$

$$\Rightarrow \hat{\beta} = (X^T X + \lambda I_p)^{-1} X^T y$$

cheap cheat

$(X^T X)^{-1}$ which is singular

Trick to avoid this is to add a small number λ to the diagonal

Lasso

$$C(\beta) = \frac{1}{n}(y - X\beta)^T(y - X\beta) + \lambda \sum_{j=0}^{p-1} |\beta_j|$$

$$\frac{d|x|}{dx} = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$$

no analytical expressions for $\hat{\beta}$