SDS 387
Linear Models

Fall 2025

Lecture 18 - Tue, Oct 30, 2025

Instructor: Prof. Ale Rinaldo

- Recall that if $R(\beta) = \mathbb{E}\left[\left(Y - X^T\beta\right)^2\right]$ is the prediction risk associated to $\beta \in \mathbb{R}^d$ then

$$R(\beta) = \underbrace{\|\beta - \beta^*\|_\Sigma^2}_{\text{estimation}} + \underbrace{\mathbb{E}\left[\left(Y - \mathbb{E}[Y|X]\right)^2\right]}_{\sigma^2 \text{ variance}} + \underbrace{\mathbb{E}\left[\left(\mathbb{E}[Y|X] - X^T\beta^*\right)^2\right]}_{\eta^2 \text{ non linearity}}$$

where $\beta^* = \mathbb{E}[XX]^{-1}\,\mathbb{E}[Y \cdot X]$

$\downarrow$ projection parameter

- If $\beta = \beta^*$ then $R(\beta^*) = \sigma^2 + \eta^2$

$$= \inf_{\beta \in \mathbb{R}^d} R(\beta)$$

because $\|\beta - \beta^*\|_\Sigma^2 = 0$ iff $\beta = \beta^*$

and $\inf_{\beta \in \mathbb{R}^d} \mathbb{E}\left[\left(\mathbb{E}[Y|X] - X^T\beta\right)^2\right] = \eta^2$

$= \mathbb{E}\left[\left(\mathbb{E}[Y|X] - X^T\beta^*\right)^2\right]$

$\Downarrow$

$0 \leq R(\beta) - R(\beta^*)$ Excess risk of β

- Of course if linearity holds, i.e.

$$Y = X^T\beta^* + \varepsilon \qquad \mathbb{E}[\varepsilon] = 0, \quad \varepsilon \perp\!\!\!\perp X$$

then $\eta = 0$ and

$$R(\beta^*) = \sigma^2 = \mathbb{E}\left[\left(Y - X^T\beta^*\right)^2\right]$$

$$= \mathbb{E}[\varepsilon^2] = \mathrm{Var}[\varepsilon]$$

- Back to last lecture: $(X_i, Y_i)_{i=1,\dots,n} \overset{iid}{\sim} P_{X,Y}$

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \in \mathbb{R}^n \qquad \underset{n \times d}{\Phi} = \begin{bmatrix} \Phi_1^T \\ \vdots \\ \Phi_n^T \end{bmatrix}$$

$\Downarrow$

design matrix or feature matrix

$\Phi_i = \varphi(X_i) \in \mathbb{R}^d$ ith feature

The plug-in estimator of β^* (either the projection parameter or the linear parameter if the model is linear)

is $\hat{\beta} = \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \; \hat{R}(\beta)$ where

empirical risk $\leftarrow$ $\hat{R}(\beta) = \hat{\mathbb{E}}_n\left[\left(Y_i - \Phi_i^T \beta\right)^2\right]$

$$= \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \Phi_i^T \beta\right)^2$$

$$= \frac{1}{n} \| \underset{n\times 1}{Y} - \underset{n\times d}{\Phi} \underset{d\times 1}{\beta} \|^2$$

- Then $\boxed{\hat{\beta} = \left(\Phi^T \Phi\right)^{-1} \Phi^T Y}$

OLS estimator

provided that Φ has full column rank $(= d)$

$\hookrightarrow$ $\underset{d\times d}{\Phi^T \Phi}$ invertible

This requires $n \geq d$ (check)!

Pf/ The function $\beta \in \mathbb{R}^d \to \hat{R}(\beta) \in \mathbb{R}$ is strictly convex because its Hessian ($d \times d$ matrix of mixed 2^{nd} derivatives) is

$$\hat{\Sigma} = \frac{1}{n} \Phi^T \Phi \succ 0$$

by assumption for all $\beta \in \mathbb{R}^d$

So $\hat{\beta}$ is the only vector s.t.

$$\nabla \hat{R}(\hat{\beta}) = 0$$

$$\Updownarrow$$

$$-\frac{2}{n} \Phi^T (Y - \Phi \hat{\beta}) = 0$$

$$\Updownarrow$$

$$\Phi^T \Phi \hat{\beta} = \Phi^T Y \qquad \text{Normal equations}$$

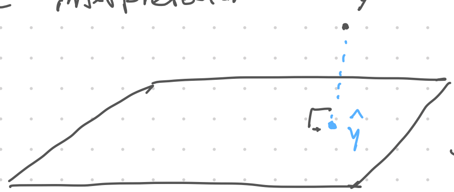
Using inverting of $\Phi^T \Phi$ we get

$$\hat{\beta} = (\Phi^T \Phi)^{-1} \Phi^T Y$$

$\blacksquare$

- Geometric interpretation

in $\mathbb{R}^d$



$\to$ linear subspace in $\mathbb{R}^n$ spanned by columns of Φ, i.e. $C(\Phi)$

$\dim(\cdot) = d \leq n$

4

where $\underset{\substack{\downarrow \\ \text{fitted values}}}{\hat{Y}} = \Phi \hat{\beta} = \underbrace{\Phi (\Phi^T \Phi)^{-1} \Phi^T}_{H} Y$

$= H Y$

where H, the hat matrix, is the orthogonal projection onto $C(\Phi)$. (H is symmetric and $H^2 = H \cdot H = H$).

- Next $\underset{\substack{\downarrow \\ \text{residuals}}}{e} = Y - \hat{Y} = \underbrace{(I_n - H)}_{\substack{\text{also a orthogonal projection} \\ \text{(symmetric + idempotent)} \\ \text{onto } C(\Phi)^{\perp}}} Y \in \mathbb{R}^n$

- Direct sum decomposition

$$Y = \hat{Y} + e \quad \text{where} \quad \langle \hat{Y}, e \rangle = 0$$

$$\frac{\|Y\|^2}{n} = \frac{\|\hat{Y}\|^2}{n} + \frac{\|e\|^2}{n}$$

$$\Updownarrow$$

$$\frac{\sum_i Y_i^2}{n} = \underset{\substack{\downarrow \\ \text{"proportion of energy" explained by the model}}}{\frac{\sum_i \hat{Y}_i^2}{n}} + \underset{\hookrightarrow \text{ least squares error}}{\frac{\sum_i e_i^2}{n}}$$

- Numerical considerations. How do you compute $\hat{\beta}$?

1) Requires matrix inversion of $\Phi^T \Phi$

order d^3 in fact order $n \cdot d^2$

2) Gradient descent iterative procedure that starting from $\beta_0 \in \mathbb{R}^d$ (i.e. $\beta_0 = 0$ or any point in the orthogonal complement of kernel(Φ))

apply the following recursion:

$t \geq 1$

$$\beta_t = \beta_{t-1} + \gamma \nabla \frac{\hat{R}}{2}(\beta_{t-1})$$

γ > 0 step size or learning rate

$$= \beta_{t-1} - \gamma \left(\Phi^T \left(\Phi \beta_{t-1} - Y \right) \right)$$

This has complexity $n \cdot d$ and as $t \to \infty$

$$\beta_t \to \hat{\beta}$$

See section 5.2 of Bach's book

- What if $\Phi^T \Phi$ is not invertible? Assume

$\text{rank}(\Phi) = n$ (for example, $d > n$).

Then the normal equations

$$\Phi^T \Phi \beta = \Phi^T Y$$

have infinitely many solutions, because if say $\hat{\beta}$ solve the normal equations, so does

$$\hat{\beta} + v$$

where $v \in \text{kernel}(\Phi)$

$\hookrightarrow \{ x \in \mathbb{R}^d : \Phi x = 0 \}$

because
$$\Phi^T \Phi (\hat{\beta} + v) = \Phi^T \Phi \hat{\beta} + \underbrace{\Phi^T \underbrace{\Phi v}_{=0}}_{=0}$$
$$= \Phi^T \Phi \hat{\beta}$$

- When $\text{rank}(\Phi) = n$ then any solution $\hat{\beta}$ to the normal equation is such that

$$\hat{Y} = \Phi \hat{\beta} = Y$$

i.e. $\hat{R}(\hat{\beta}) = 0$ and the model

<u>interpolate</u> the data

- Among the infinitely many solutions to the normal equations there is a canonical one:

the min-norm solution, the one with smallest Euclidean norm. This is defined as

$$\hat{\beta}_{MN} = (\Phi^T\Phi)^+ \Phi^T Y$$

where for a $\underset{m\times n}{A}$ its Moore-Penrose pseudo inverse A^+ is a $n\times m$ matrix s.t.

i) $AA^+A = A$ (AA^+ maps columns of A to themselves, so it is an identity on $C(A)$)

ii) $A^+AA^+ = A^+$

iii) AA^+ is symmetric
A^+A

- Extra properties $\text{kernel}(A^+) = \text{kernel}(A^T)$
$C(A^+) = C(A^T)$

AA^+ and A^+A are idempotent

$AA^+ \downarrow$ orthogonal projection onto $C(A)$

$A^+A \downarrow$ orthogonal projection onto $C(A^T)$ $\hookrightarrow$ row space of A

- If $\underset{m \times n}{A} = \underset{m \times k}{U} \; \underset{k \times k}{\Sigma} \; \underset{k \times n}{V^T}$ where Σ is diagonal with positive diagonal elements (the singular values) and $k = \text{rank}(A)$

Then

$$A^+ = V \Sigma^{-1} U^T$$

- Back to interpolation (i.e. $\text{rank}(\Phi) = n$)

$$\hat{\beta}_{MN} = \Phi^+ Y = \text{argmin} \{ \|\beta\| \text{ s.t. } \Phi\beta = Y \}$$

and gradient descent $\to \hat{\beta}_{MN}$