ISyE 6740 Computational Data Analysis

Supervised learning

input → algorithm → target#

new input → ? (Classification / Regression)

→ if the new input does not meet some specific criteria, algorithm does not know what to do with the data

vs

Unsupervised learning

subset input → algorithm → target

new input → algorithm → target

Algorithm learns to make decisions after training/based on some criteria, it gets to target (E-M algorithm, PCA, NNMF, etc.)

↪ latent variables algorithms

Linear Regression

$\vec{x}_1, \dots, \vec{x}_n \in \mathbb{R}^d$ (each x_i is a vector, see below to understand)

$y_1, \dots, y_n \in \mathbb{R}$ ($y_1, \dots, y_n$ is a single vector)

$$y_i = x_i^T \theta^* + \varepsilon_i \quad \text{for } i = 1, \dots, n$$

is equivalent to:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix}_{n\times d} \begin{bmatrix} \theta_1^* \\ \theta_2^* \\ \vdots \\ \theta_d^* \end{bmatrix}_{d\times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}$$

$\vec{x}_1^T \in \mathbb{R}^d$, $\vec{x}_n^T \in \mathbb{R}^d$

this matrix can be also written as follows.

$$\begin{bmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & & & \\ x_{1d} & x_{2d} & \cdots & x_{nd} \end{bmatrix} \quad (T) \rightarrow \text{transpose the original}$$

$\vec{x}_1 \ \vec{x}_2 \ \cdots \ \vec{x}_n \in \mathbb{R}^d$

$$\Leftrightarrow [\vec{x}_1 \ \vec{x}_2 \cdots \vec{x}_n]^T \Leftrightarrow \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}_{n\times d}$$

hence

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1}}_{n\times 1} = \underbrace{\begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}_{n\times d} \begin{bmatrix} \theta_1^* \\ \theta_2^* \\ \vdots \\ \theta_d^* \end{bmatrix}_{d\times 1}}_{= n\times 1} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}}_{n\times 1}$$

✓ dimensions ok!

or equivalently: $y_i = x_i^T \theta^* + \varepsilon_i \quad \text{for } i = 1, \dots, n$

Ordinary Least Square

$$\hat{\theta}^{OLS} = \underset{\theta}{\operatorname{argmin}} \frac{1}{2n} \|Y - \hat{Y}\|_2^2$$

(sum of squared errors)

↳ this term here is actually optional, some literature includes, some not!

but

$$\|Y - \hat{Y}\|_2^2 = \|Y - X\hat{\theta}\|_2^2 = \sum_{i=1}^{n} (y_i - x_i^T \theta)^2$$

(equivalent)

(recall from our Linear Algebra Lesson.
the ℓ_2-norm $\|x\|_2 = \sqrt{\langle x, x \rangle}$
Hence $\|x\|_2^2 = \langle x, x \rangle$ (inner product)
$\|x\|_2^2 = \sum_{i=1}^{n} x_i^2$)

Let's do both ways

(1) Using summation

$$\hat{\theta}^{OLS} = \underset{\theta}{\operatorname{argmin}} \frac{1}{2n} \left(\sum_{i=1}^{n} (y_i - x_i^T \theta)^2 \right) = \underset{\theta}{\operatorname{argmin}} \frac{1}{2n} \left(\sum_{i=1}^{n} \left(y_i^2 - 2 y_i x_i^T \theta + (x_i^T \theta)^2 \right) \right)$$

$$= \underset{\theta}{\operatorname{argmin}} \frac{1}{2n} \left(\sum_{i=1}^{n} y_i^2 - 2 \sum_{i=1}^{n} \underbrace{y_i}_{1\times 1} \underbrace{(x_i^T \theta)}_{1\times 1} + \sum_{i=1}^{n} \underbrace{(x_i^T \theta)^T}_{1\times 1} \underbrace{(x_i^T \theta)}_{1\times 1} \right) \qquad \vec{x} \in \mathbb{R}^{d\times 1}, \ \vec{\theta} \in \mathbb{R}^{d\times 1}$$

($(x_i^T\theta)^T (x_i^T\theta) = \theta^T x_i x_i^T \theta$)

F.O.C.

First order condition
= derivative with respect to θ and set to 0

$$\frac{\partial f(\theta)}{\partial \theta} = -\frac{2}{2n} \sum_{i=1}^{n} \underbrace{(y_i x_i^T)^T}_{1\times d} + \frac{2}{2n} \sum_{i=1}^{n} \underbrace{(x_i \underbrace{x_i^T}_{d\times 1} \underbrace{\theta}_{1\times d})}_{d\times 1}$$

$$= -\frac{1}{n} \left(\sum_{i=1}^{n} \underbrace{x_i y_i}_{(d\times 1)} - \sum_{i=1}^{n} (x_i x_i^T \theta) \right) \overset{\text{F.O.C}}{=} 0$$

$$\Leftrightarrow -\frac{1}{n} \sum x_i y_i = \frac{1}{n} \sum x_i x_i^T \theta \quad \Rightarrow \quad \hat{\theta}^{OLS} = (\Sigma x_i x_i^T)^{-1} \Sigma x_i y_i$$

where:

$$\sum_{i=1}^{n} x_i x_i^T = \sum_{i=1}^{n} \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{pmatrix}_{d\times 1} (x_{i1} \ x_{i2} \ \dots \ x_{id})_{1\times d} = \begin{pmatrix} \Sigma x_{i1}^2 & \Sigma x_{i2} x_{i1} & \Sigma x_{id} x_{i1} \\ \vdots & \Sigma x_{i2}^2 & \\ & & \\ \Sigma x_{i1} x_{id} & \dots & \Sigma x_{id}^2 \end{pmatrix}_{d\times d}$$

↳ must be invertible

And

$$\sum_{i=1}^{n} x_i y_i = \begin{pmatrix} \Sigma x_{i1} y_i \\ \Sigma x_{i2} y_i \\ \\ \Sigma x_{id} y_i \end{pmatrix}_{d\times 1}$$

using matrix notation. $Y \in \mathbb{R}^{n\times 1}$ $\theta \in \mathbb{R}^{d\times 1}$ $X \in \mathbb{R}^{n\times d}$

$\times \frac{1}{2n}$ $\|Y - \hat{Y}\|_2^2 = \|Y - X\hat{\theta}\|_2^2 = (Y - X\hat{\theta})^T (Y - X\hat{\theta}) = Y^TY - \underbrace{Y^TX\theta}_{1\times 1} - \underbrace{\theta^TX^TY}_{1\times 1} + \underbrace{\theta^TX^TX\theta}_{1\times n \; n\times n \; n\times 1}$

$f(\theta) = Y^TY - 2\theta^TX^TY + \theta^TX^TX\theta$

F.O.C. $\frac{\partial f(\theta)}{\partial \theta} = \frac{1}{2N}\left(-2X^TY + 2X^TX\theta\right) = 0$

$\Leftrightarrow \frac{1}{N}X^TX\theta = \frac{1}{N}X^TY \Rightarrow \hat{\theta}^{OLS} = (X^TX)^{-1}X^TY$

Equivalently

$\hat{\theta}^{OLS} = \arg\min_\theta \underbrace{\frac{1}{2n}\|y - X\theta\|_2^2}_{\mathcal{L}(\theta)}$

→ Gradient

$\nabla \mathcal{L}(\hat{\theta}) = \frac{1}{2n}(2)(-1)X^T(y - X\theta)$

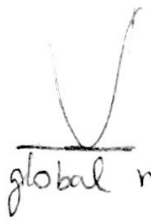
$\Leftrightarrow \frac{1}{n}X^T(X\theta - y) = 0 \quad \Leftrightarrow \quad X^TX\theta = X^Ty$

$\hat{\theta} = (X^TX)^{-1}X^Ty$

→ Hessian

$\nabla^2 \mathcal{L}(\hat{\theta}) = \frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta^2} = \frac{1}{n}X^TX \; (\succ 0)$ → positive definite

(strictly convex)



Likelihood function (since $\varepsilon \sim N(0, \sigma^2 I_n)$ → $\hat{\theta} \sim MN(\vec{\theta}, \sigma^2(X^TX)^{-1})$ | $\vec{\theta}$ is a vector

$\mathcal{L}(\theta) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{\frac{-1}{2\sigma^2}\underbrace{(y - X\theta)^T(y - X\theta)}_{\|y - X\theta\|_2^2}}$

take log: $\log \mathcal{L}(\theta) = +\log 1 - \frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|y - X\theta\|_2^2$ ($\log 1 \to 0$)

1ST derivative $\frac{\partial \log \mathcal{L}(\theta)}{\partial \theta}$ ⇒ for some σ^2 constant, with $\varepsilon \sim N(0, \sigma^2 I_n)$

OLS provides same solution as MLE.