$\boxed{\text{6 Oct 2019}}$

Lecture 2

$\in \mathbb{N}$

[counts in discrete space]

[intervals:]

Data = [EDA]

I. { categorical } — classes

{ quantitative } — numeral $x, y \in \mathbb{R}$ $\quad x \in \mathbb{N}$

$\vec{x} \in \mathbb{R}^d$

$X \in \mathbb{R}^{m \times n}$

[+ continuous] — $\mathbb{R}$

[+ [discrete / ordinals]] — $\mathbb{N}$

II. [univariate $\updownarrow$ multivariate] $\mathbb{R}^{d}$

III. **Predictor** $\Rightarrow$ **Outcome**

dependency: [independent] $\Rightarrow$ [dependent]

[control / input] $\Rightarrow$ [observable / output]

Both can have randomness but

$$x + \varepsilon_{in} \rightsquigarrow f(x + \varepsilon_{in}) + \delta_{obs}.$$

Learning : from input $\longrightarrow$ output

Brief summary for supervised learning

given data $\longrightarrow$ EDA (clean-up)

$\downarrow$

decide on variables

+

task

$\rightarrow$ Declare predictor space X ; outcome space Y

$\rightarrow$ [N noisy Samples to be used for learning] (x_i, y_i) $i=1,\cdots,N$

$x_i \in X$ $y_i \in Y$

In Supervised learning ~~means that~~ this is the training set.

Keep some samples for validation

For a given task with training set $\{(x_i, y_i)\}_{i=1}^{N}$

$\Rightarrow$ Find function

$$f : X \rightarrow Y$$

and define

loss function $L(f(x), y)$

such that

(i) in-sample loss (error) is minimised

$$\mathbb{E}\left[L\left(f(\{x_i\}), \{y_i\} \right)\right] \ll$$

<u>and</u> $\{(x_i, y_i)\}_{i \in \text{training}}$

(ii) expected out-of-sample loss is also small

$$\mathbb{E}\left[L\left(f(\{x_k\}), \{y_k\} \right)\right] \ll$$

$\{(x_k, y_k)\}_{k \in \text{test}}$ (validation)

$\Downarrow$

we expect then that

$f(x_{\text{unknown}})$ will be a good predictor for <u>y_{unknown}</u>

Linear regression

Data: Inputs (predictors) X | Output (outcome) y

Quantitative

Samples:

$i = 1, \ldots, N$ samples (observations)

$$
\begin{array}{cccc|c}
x_1^{(1)} & x_2^{(1)} & \cdots & x_p^{(1)} & y^{(1)} \\
x_1^{(2)} & x_2^{(2)} & \cdots & x_p^{(2)} & y^{(2)} \\
\vdots & & & & \\
x_1^{(N)} & x_2^{(N)} & \cdots & x_p^{(N)} & y^{(N)}
\end{array}
$$

Inputs: $\{\vec{x}^{(i)}\}_{i=1}^{N}$ output $\{y^{(i)}\}_{i=1}^{N}$

Assume a linear relationship: $f(\vec{x}) = \underline{\beta_0} 1 + \underline{\beta_1} x_1 + \underline{\beta_2} x_2 + \ldots + \underline{\beta_p} x_p$

$\left.\begin{array}{l}\vec{x} \in \mathbb{R}^p \\ y \in \mathbb{R}\end{array}\right\}$ $\hat{y} = f_{LR}(\vec{x}\,; \vec{\theta})$

$\vec{\theta} = (\beta_0, \ldots \beta_p)$

(Hyper) Parameters

$\boxed{\text{Need to find } \vec{\theta}} \leftarrow \boxed{\text{Defines the model}}$

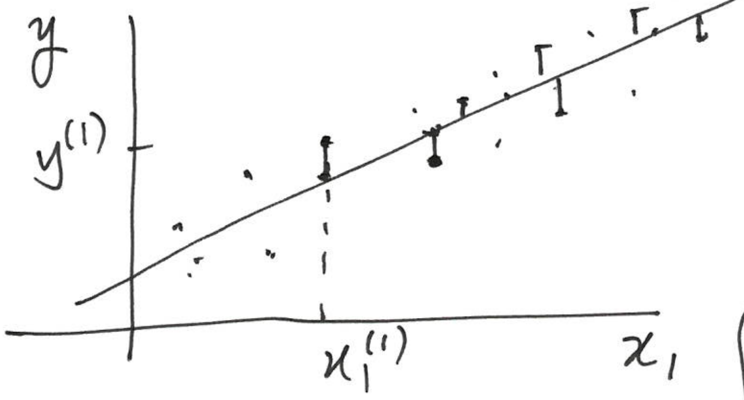
$f: X \to Y$

$f: \mathbb{R}^p \to \mathbb{R}$

Simplest case (so that we can draw):

Assume $\underline{p=1}$ $\qquad$ $\underline{N \gg 2}$

$f(x_1)$ $\quad (x_1^{(i)}, y^{(i)}) \quad i=1,\dots,N$



$$f(x_1)=\beta_0+\beta_1 x_1$$

$$\begin{cases} y^{(1)} = \beta_0 + \beta_1 x_1^{(1)} \\ y^{(2)} = \beta_0 + \beta_1 x_1^{(2)} \\ \vdots \\ y^{(N)} = \beta_0 + \beta_1 x_1^{(N)} \end{cases}$$

Mean square error:

MSE

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left[y^{(i)} - f_{LR}(x_1^{(i)}; \beta_0, \beta_1)\right]^2 = \frac{1}{N}\sum_{i=1}^{N} e^{(i)\,2}$$

$$\| \quad L(y, f_{LR}(x_1))$$

Define: $\vec{y}^{T} = (y^{(1)}, \dots, y^{(N)}) \in \mathbb{R}^N$

$\vec{x}_1^{T} = (x_1^{(1)}, \dots, x_1^{(N)}) \in \mathbb{R}^N$

$\vec{1}^{T} = (1, \dots, 1) \in \mathbb{R}^N$

$$f_{LR}(\vec{x}) = \beta_0 \vec{1}_{N\times 1} + \beta_1 \vec{x}_{1\,N\times 1} = \underbrace{\begin{pmatrix} \vec{1} & \vec{x}_1 \end{pmatrix}_{N\times 2}}_{= X} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}_{2\times 1}$$

$$f_{LR}(\vec{x}) = X \cdot \vec{\beta}$$

Reminder $(\vec{\beta} = \vec{\theta})$

$$\vec{e} = \vec{y} - X\vec{\beta}$$

$$L \cdot [f_{LR}(\vec{x}_1), y] = \frac{1}{N} \vec{e}^T \cdot \vec{e} = \frac{\|\vec{e}\|^2}{N}$$

Minimize L in the space of parameter $\vec{\beta}$

$$L(\vec{\beta}) = L(\beta_0, \beta_1) = \frac{1}{N}\left[(\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta})\right]$$

(1) $$\left.\frac{dL}{d\vec{\beta}}\right|_{\vec{\beta}^*} = \nabla_{\vec{\beta}} L \Big|_{\vec{\beta}^*} = 0$$

(2) Check that $$H(\vec{\beta}^*)_{ij} = \left(\frac{\partial^2 L}{\partial\beta_i \partial\beta_j}\right)$$

$$H > 0$$

$$L(\vec{\beta}) = \frac{1}{N}\left[\vec{y}^T\cdot\vec{y} - \vec{y}^T X\vec{\beta} - \vec{\beta}^T X^T\vec{y} + \vec{\beta}^T X^T X\vec{\beta}\right]$$

$$\nabla_{\vec{\beta}} L \quad \left(\vec{y}^T X\vec{\beta}\right)$$

Aside

$$\nabla_{\vec{\beta}}\left(\vec{\alpha}^T\cdot\vec{\beta}\right) = \nabla_{\vec{\beta}}\left(\alpha_1\beta_0 + \alpha_2\beta_1\right) = \begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix} = \vec{\alpha}$$

$$\nabla_{\vec{\beta}}\left(\vec{\beta}^T\cdot\vec{\alpha}\right) = \vec{\alpha}$$

$$\nabla_{\vec{\beta}}\left(\vec{\beta}^T A\vec{\beta}\right) = A\vec{\beta} + A^T\vec{\beta} = \left(A + A^T\right)\vec{\beta}$$

$$\nabla_{\vec{\beta}} L(\vec{\beta}) = \frac{1}{N}\left[-X^T\vec{y} - X^T\vec{y} + \left(X^T X + (X^T X)^T\right)\vec{\beta}\right]$$

$$= \frac{-2}{N}\left[X^T\vec{y} - X^T X\vec{\beta}\right]$$

Normal equations.

$$\nabla_{\vec{\beta}} L \Big|_{\vec{\beta}^*} = 0 \qquad X^T \cdot \vec{y} = (X^T X)\vec{\beta}^*$$

$$X = \left[\vec{1} \,\vdots\, \vec{x}_1 \right]_{N \times 2} \text{ invertible}$$

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