Lecture october 11

Basic elements in machine learning and Linear regession

- Our data: inputs and outputs (target)

  $D = \{ (x_0, y_0), (x_1, y_1) \dots (x_{n-1}, y_{n-1}) \}$

  in total n-entries

- Model

- Cost/error/loss function.

Example:

Important assumption:

$$y = [y_0 \; y_1 \dots y_{n-1}]^T$$

$$y = \underbrace{f(x)}_{\text{there exists a continuous function}} + \varepsilon$$

random

noise

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

$(\sigma^2 = 1)$

$$f(x) \simeq P_{p-1}(x)$$

$$P_{p-1}(x_i) = \sum_{j=0}^{p-1} \beta_j x_i^j$$ Model

$$y_i \simeq P_{p-1}(x_i) = \tilde{y}_i$$

$$\tilde{y}_0 = \beta_0 \underbrace{x_0^0}_{=1} + \beta_1 x_0^1 + \beta_2 x_0^2 + \dots \beta_{p-1} x_0^{p-1}$$

$$\tilde{y}_1 = \beta_0 + \beta_1 x_1^1 + \beta_2 x_1^2 + \dots + \beta_{p-1} x_1^{p-1}$$

$$\vdots = \qquad \vdots$$

$$\tilde{y}_{n-1} = \beta_0 + \beta_1 x_{n-1}^1 + \beta_2 x_{n-1}^2 + \dots \beta_{p-1} x_{n-1}^{p-1}$$

Def

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

$$
= \begin{bmatrix} x_{00} & x_{01} & \cdots & x_{0\,p-1} \\ x_{10} & x_{11} & \cdots & x_{1\,p-1} \\ x_{20} & & & \vdots \\ \vdots & & & \vdots \\ x_{n-1\,0} & & & x_{n-1\,p-1} \end{bmatrix}
$$

$$
\in \mathbb{R}^{n \times p}
$$

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

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

$$
\boxed{\tilde{y} = \underbrace{X\beta} \in \mathbb{R}^n}
$$

X contains input data assuming that x_{ij} are not stochastic variables

X = Design / feature matrix

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

$$x_j = \begin{bmatrix} x_{0j} & x_{1j} & \cdots & x_{n-1\,j} \end{bmatrix}^T$$

each represents a feature.

β are unknown parameters. To be determined.

Our Model

$$\boxed{\tilde{y} = X\beta}$$

linear in the unknown parameter β

$$\begin{bmatrix} x_{00} & x_{01} & \cdots & x_{0p-1} \end{bmatrix}$$

$$X = \begin{bmatrix} x_{10} & & & \\ x_{20} & & & \\ \vdots & & & \\ x_{n-1\,0} & - & - & x_{n-1\,p-1} \end{bmatrix}$$

rows are the data entries

columns are the features.

— Assess if this is a good model.

Cost function $C(D;\beta)$

$$= C(\beta)$$

$$\tilde{y}_i = \sum_{j=0}^{p-1} x_{ij}\beta_j$$

$\tilde{y}_i$ should reproduce $\underline{\underline{y_i}}$

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

$=$ MSE $=$ mean square

error

Absolute error

$$|y - \tilde{y}|$$

or relative error

$$\left|\frac{y - \tilde{y}}{y}\right|$$

$$\nabla_\beta C(\beta) = 0$$

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

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

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

$$= \frac{1}{n}\|(y - X\beta)\|_2^2$$

$$\|z\|_2^2 = \sum_{i=1}^{m} z_i^2$$

$$\frac{\partial C}{\partial \beta} = -\frac{2}{n} X^T (y - X\beta) = 0$$

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

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

$$(X^T X)^{-1} X^T y \Rightarrow$$

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

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min}\; C(\beta)$$