Lecture 3: Regression

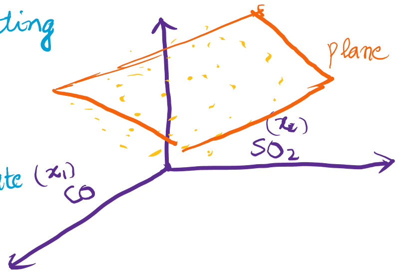
17 January 2024 10:59

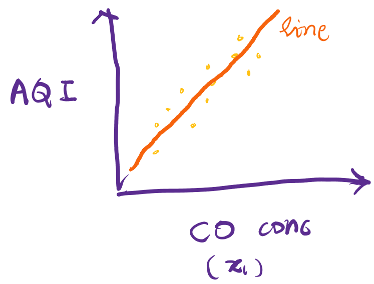
Lets begin with an exampe: AQI (Air Quality Index)

$AQI = \max\{f_1(x), \ldots f_n(x_n)\}$ where $f_i(x_i)$ is a function associated with i^{th} pollutant.

The device for measuring AQI, may not be measuring all pollutants, but some of them and use the data to estimate AQI.

This is done by fitting the perfect AQI data with the limited observation and estimate AQI.





Linear Regression

- "Simple but powerful"
- Works on transformation of raw data
- Intrepretable

? How will we best fit the given data?

Measure the goodness of the fit using an error fn.

Error/ Loss/cost/ Energy : $E(f, D)$ where $D = \{(\underline{x}_1, y_1), (\underline{x}_2, y_2) \ldots (\underline{x}_n, y_n)\}$

f is the estimated function

Some possible E-functions:

1. $\sum_{i=1}^{n} (f(\underline{x}_i) - y_i)$ (signed, not a good candidate for E-f^n)
2. $\sum_{i=1}^{n} |f(\underline{x}_i) - y_i|$ (unsigned, a good candidate)
3. $\sum_{i=1}^{n} (f(\underline{x}_i) - y_i)^2$ (Squared error - a good candidate)
4. $\sum_{i=1}^{n} (f(\underline{x}_i) - y_i)^3$ (signed, not a good candidate)

Squared error function

$$\sum_{\{i : x_i, y_i \in D\}} (f(\underline{x}_i) - y_i)^2$$

This function is useful as it is:

- Continuous, differentiable
- Geometrically, it is similar to reducing euclidean distance b/w estimated f^n and data set. Hence visualisable in Euclidean space
- Mathematical analysis becomes easier

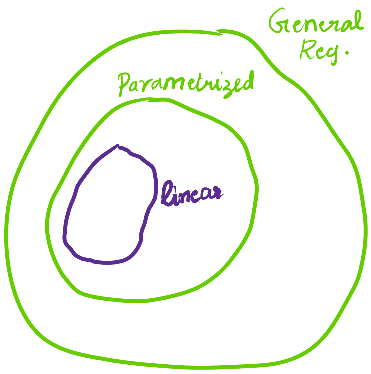
Let $x_i \in \mathbb{R}^d$, $y \in \mathbb{R}$:- $x_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{id} \end{bmatrix}$

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \qquad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

General regression problem

Find a f^n f^* such that $f^*(x)$ is the best predictor of y w.r.t D

$$f^* \in \underset{f \in \mathcal{F}}{\arg\min}\ E(f, D)$$



Parametrized Regression

Here we will find a f^n $f \equiv f(x, w)$

Eg: i. $f(x, (\alpha, \lambda)) = \alpha e^{-\lambda^T x}$

ii. $f(x, w) = w_0 + w_1 x + \ldots\ w_k x^k$

Here, $f^* \in \underset{w}{\arg\min}\ E(f(x,w), D)$

Linear Regression

Here we've $f(x, w) = w^T x + w_0 = \bar{w}^T \bar{x}$, where $w = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix} \in \mathbb{R}^d$

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

NB: Let $f(x, w) = w_0 + w_1 x + w_2 x^2 + \ldots\ w_n x^n$. by suitably replacing 'x' we can solve this problem as linear regression

Least square Optimization for Linear Regression

$$\sum^{n} \left(\sum^{d} \right)^2 \quad \text{where } D = \{(x_1, y_1) \ldots (x_n, y_n)\}$$

$$w^* \in \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n} \left(\sum_{j=1}^{d} w_j x_{ij} - y_i \right)^2$$ where $D = \{(x_1, y_1), \cdots (x_n, y_n)\}$

Prediction: $\hat{y}_i = \sum_{j=1}^{d} w_j^* x_{ij}$

d = 1 LR:

$$E(w, D) = \sum_{i=1}^{n} (y_i - w_0 - w_1 x_i)^2$$

We need to find $w = (w_0, w_1)$ such that $\frac{\partial E}{\partial w_0} = \frac{\partial E}{\partial w_1} = 0$

$$\frac{\partial E}{\partial w_0} = -2 \sum_{i=1}^{n} (y_i - w_0 - w_1 x_i) = 0$$

$$\Rightarrow \quad w_0 = \frac{\sum y_i - w_1 \sum x_i}{n} = \bar{y} - w_1 \bar{x}$$

where $\bar{y} = \sum y_i / n$ and $\bar{x} = \sum x_i / n$

Similarly, $\frac{\partial E}{\partial w_1} = 0 \Rightarrow w_1 = \frac{\sum x_i y_i - w_0 \sum x_i}{\sum x_i^2}$

Let $\alpha = \frac{\sum x_i y_i}{\sum x_i^2}$, $\beta = \frac{\sum x_i^2}{n}$.

we've: $w_1 = \frac{\alpha \beta - \bar{x}\bar{y}}{\beta - \bar{x}^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

d-D Data LR

$$x_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{id} \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}^{j-x}$$

$$w^* \in \arg\min_w \sum_{i=1}^{n} (y_i - w^T x_i)^2 = \sum_{i=1}^{n} z_i^2$$

where $z = \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix} = \begin{bmatrix} y_1 - x_1^T w \\ \vdots \\ y_n - x_n^T w \end{bmatrix} = y - Xw$

Hence $\arg\min_w \; (Xw - y)^T (Xw - y)$

$$E(w, D) = w^T X^T X w - 2y^T X w + y^T y$$

$$\nabla_w E = 0 \Rightarrow 2X^T X w - 2X^T y = 0$$

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

Also, $\hat{y} = w^{*T} x$