Under Gaussian noise assumption linear regression amounts to least square.

Sourav Bhattacharjee

February 2021

1 Probabilistic Modeling of Linear Regression

1.1 Linear Model:

Suppose we are given a dataset D= { $\mathrm{x}_i, y_i\}_{i=1}^m$
x_i's are known as the feature and y_i's are known as target
We want to fit a line to the given data.
Suppose we fit a line

$$y_i \approx \theta^T x_i$$

$$or,\ y_i = \theta^T x_i + \epsilon_i$$

where ϵ_i's are random noise to model unknown effects and follow Gaussian distribution.

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

$$\Rightarrow p(\epsilon_i) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{\epsilon_i^2}{2\sigma^2}\right)$$

$$\Rightarrow p(y_i - \theta^T x_i) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{(y_i - \theta^T x_i)^2}{2\sigma^2}\right]$$

However the conventional way to write the probability is

$$p(y_i|x_i;\theta) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{(y_i - \theta^T x_i)^2}{2\sigma^2}\right]$$

1.2 Parameter Estimation:

Bayes Theorem:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} = \frac{p(\theta, D)}{p(D)}$$

1.3 Maximum Likelihood Estimation(MLE):

$\theta^* = argmax_\theta L(\theta|\mathcal{D})$

$= argmax_\theta P(\mathcal{D} \mid \ \theta)$

$= \text{argmax}_\theta P(y1, x_1, ..., y_m, x_m; \theta)$

$= \text{argmax}_\theta \prod_{i=1}^m P(y_i, x_i; \theta)$

$= \text{argmax}_\theta \prod_{i=1}^m [P(y_i|x_i; \theta).P(x_i; \theta]$

$= \text{argmax}_\theta \prod_{i=1}^m [P(y_i|x_i; \theta)].P(x_i)$

$= \text{argmax}_\theta \prod_{i=1}^m [P(y_i|x_i; \theta)]$

$= \text{argmax}_\theta \quad \sum_{i=1}^m logP(y_i|x_i; \theta)$

$= \text{argmax}_\theta \quad \sum_{i=1}^m \left[log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) + log\left(exp\left(-\frac{(\theta^T x_i - y_i)^2}{2\sigma^2}\right)\right)\right]$

$= \text{argmax}_\theta - \frac{1}{2\sigma^2} \quad \sum_{i=1}^m (\theta^T x_i - y_i)^2$

$= \text{argmin}\theta \frac{1}{m} \quad \sum_{i=1}^m (\theta^T x_i - y_i)^2$

2 Conclusion:

Hence under the Gaussian assumption (i.e. when the error/noise terms follow Normal distribution) , the linear regression amounts to least square