Under Gaussian noise assumption linear regression amounts to least square

Alimpan Barik

February 2021

1 Introduction

Linear regression attempts to model the relationship between two variables by fitting a line to the observed data. One variable is considered to be the independent variable and the other is considered to be the dependent variable

2 Probabilistic model of Linear Regression

2.1 Linear Model:

Suppose we are given a dataset D= { x_i, y_i}
x_i is known as the feature and y_i's are known as label/target
We want to fit a line to the given data.
Suppose we fit a line

$$y_i \simeq \theta^T x_i$$

or, $y_i = \theta^T x_i + \epsilon_i$ (1)

where ϵ_i's are random noise to model unknown effects
ϵ_i's are i.i.ds and by our assumption they follow the Gaussian distribution i.e.
$\epsilon_i \sim N(0, \sigma^2)$

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

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

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

However the conventional way to write the probability is

$$p(y_i|x_i;\theta) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(\frac{-(y_i-\theta^T x_i)^2}{2\sigma^2}\right) \tag{3}$$

Now from Bayes' Theorem

$P(\theta|D) = P(D|\theta)\frac{P(\theta)}{P(D)} \Rightarrow P(\theta|D) = \frac{P(\theta,D)}{P(D)} = \frac{P(D|\theta)P(\theta)}{\sum_\theta P(D|\theta)P(\theta)}$

2.2 Maximum Likelihood Estimation:

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

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

$= \text{argmax}_\theta P(y_1, x_1, ..., y_m, x_m; \theta)$

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

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

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

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

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

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

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

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

3 Conclusion:

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