Machine Learning
Assignment 1

Shivam Verma
Email address: shivam.59910103@gmail.com

Contents

1 Gaussian assumption ⇒ Linear regression amounts to least square

1.1 Step 1: Parameter Identification

If we are given a data set, we identify one quantity in a column of dataset as *s.t.* the feature variable (x^i), and another as the target variable (y^i), alongwith the offset/error/unmodelled effects/residuals (ε^i) which tells us the error caused by a given value of θ we choose for the best fit of the curve with the dataset given.

$$y^i = \theta^T x^i + \varepsilon^i \tag{1}$$

1.2 Step 2: Assumption

We assume that the residuals follow a gaussian distribution.

$$i.e. \quad P(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{\theta^2}{2\sigma^2}\right] \tag{2}$$

where, σ is the variance of the dataset. Now, from Eq[1] and Eq[2], we can write,

$$i.e. \quad P(y^i|x^i;\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{\left(y^i - \theta^T x^i\right)^2}{2\sigma^2}\right] \tag{3}$$

The Eq[3] tells us the probablity of the target value y^i for a given feature x^i and residual θ.

1.3 Step 3: Proof

We can define a new function that tells us about the "likelihood" of the Gaussian Distribution we obtained in Eq[3]. we can write it as,

$$L(\theta) = \prod_{i=1}^{n} P(y^i|x^i;\theta) \quad = \quad \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{\left(y^i - \theta^T x^i\right)^2}{2\sigma^2}\right] \tag{4}$$

Taking Log of Eq[4] we get,

$$l(\theta) = \ln\left[L(\theta)\right] \quad = \quad \underbrace{n \ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right)}_{\text{constant}} - \frac{1}{\sigma^2} \sum_{i=1}^{n} \underbrace{\frac{1}{2}\left(y^i - \theta^T x^i\right)^2}_{ls(\theta)} \tag{5}$$

Our aim is to maximum the function $l(\theta)$ also referred to as "log likelihood"[1].

$$ls(\theta) = \frac{1}{2}\left(y^i - \theta^T x^i\right)^2 \tag{6}$$

We can clearly see in the Eq[5] that to maximize the log likelihood we need to minimize $ls(\theta)$, and to minimize $ls(\theta)$ we have to differentiate it and find it's minima. which is nothing but least square criterion.

2 Conclusion

We started with a dataset where we first identified the elemts of it into a more familiar nomenclature to understand it's role. Upon the assumption of Gaussian distribution of the residuals, we arrive at the notion of likelihood which we intend to maximize in order to get the curve we intend to fit our data with is as close as possible. Doing so, we end up minimizing the function $ls(\theta)$ which we end up realizing is least squaring criterion of the dataset.

References

[1] Andrew Ng. CS229 Lecture Notes. http://cs229.stanford.edu/notes2020spring/cs229-notes1.pdf, 2020.