ML Assignment 1

Srijan Mallick

February 28, 2021

Theorem 1. *Under Gaussian assumption linear regression amounts to least square (ordinary least square)*

Proof: In a linear regression problem, our primary objective is to estimate the regression coefficients by minimising the error term. In this case, we consider our loss function as the square of the error terms, and the cost function J is the loss, averaged over the entire dataset. In this section, we will give a set of probabilistic assumptions, under which least-squares regression is derived.

The required linear model is given by:

$$y_i = \theta^T x_i + \varepsilon_i \text{ , where } \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

Here ε_i is the error term for each individual data point, which captures the error in prediction, or some random noise. The density of ε_i is given by:

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

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

The conventional way to right this probability is:

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

where the notation in RHS denotes the distribution of y_i given x_i, parametrised by θ. When viewed as a function of θ, we call it the **likelihood function**,

given by

$$\begin{aligned} L(\theta) &= L(\theta \mid x_i, \vec{y}) \\ &= p(\vec{y} \mid X; \theta) \\ &= \prod_{i=1}^{m} p(y_i|x_i;\theta) \\ &= \prod_{i=1}^{m} \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{(y_i - \theta^T x_i)^2}{2\sigma^2}\right] \end{aligned}$$

Instead of maximizing L(θ) , we can also maximize any strictly increasing function of L(θ). So instead we maximize the **log likelihood function**.
Let the data be given by $\mathcal{D} = (x_i, y_i)_{i=1}^n$
With Bayes theorem we compute θ from data $\mathcal{D}$

$$\begin{aligned} p(\theta|\mathcal{D}) &= \frac{p(\mathcal{D}|\theta).p(\theta)}{p(\mathcal{D})} \\ &= \frac{\mathbf{L}(\theta|\mathcal{D}).p(\theta)}{p(\mathcal{D})} \end{aligned}$$

$p(\mathcal{D}|\theta)$ is a function of θ given $\mathcal{D}$ as we want to choose that particular θ which will maximize the probability i.e. the **Maximum Likelihood Estimator**.

$$\begin{aligned} \theta^* &= \underset{\theta}{argmax} \;\; L(\theta|\mathcal{D}) \\ &= \underset{\theta}{argmax} \;\; p(\mathcal{D}|\theta) \\ &= \underset{\theta}{argmax} \;\; p(y_1, x_1, y_2, x_2, y_3, x_3, \ldots, y_m, x_m \; ; \; \theta) \\ &= \underset{\theta}{argmax} \;\; \prod_{i=1}^{m} p(y_i, x_i \; ; \theta) \\ &= \underset{\theta}{argmax} \;\; \prod_{i=1}^{m} p(y_i|x_i \; ; \theta) p(x_i \; ; \theta) \\ &= \underset{\theta}{argmax} \;\; \prod_{i=1}^{m} p(y_i|x_i \; ; \theta) p(x_i) \\ &= \underset{\theta}{argmax} \;\; \prod_{i=1}^{m} p(y_i|x_i \; ; \theta) \\ &= \underset{\theta}{argmax} \;\; \sum_{i=1}^{m} log \; [p(y_i|x_i \; ; \theta)] \\ &= \underset{\theta}{argmax} \;\; \sum_{i=1}^{m} \log \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{(y_i - \theta^T x_i)^2}{2\sigma^2}\right] \end{aligned}$$

$$
\begin{aligned}
&= \underset{\theta}{argmax} \quad \frac{-1}{\sigma^2} \cdot \frac{1}{2} \sum_{i=1}^{m} (y_i - \theta^T x_i)^2 \\
&= \underset{\theta}{argmin} \; \frac{1}{m} \sum_{i=1}^{m} (y_i - \theta^T x_i)^2
\end{aligned}
$$

which we recognize to be $\mathbf{J}(\theta)$, our original least-squares cost function.