ML Assignment 1

Nilabjanayan Bera

February 28, 2021

Theorem 1. ***under Gaussian assumption linear regression amounts to least square***

Proof: In probabilistic modelling we consider a linear model -

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

Considering ϵ_i as the random noise to model unknown effects -

$$y_i = \theta^T x_i + \epsilon_i \quad , \quad \text{where } \epsilon_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$$

The density of ϵ_i is given by -

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

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

However the conventional way to write the probablity is

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

The notation $p(y_i | x_i \ ; \theta)$ indicates that this is the distribution of y_i given x_i and parameterized by θ (we can not condition on θ since θ is not a random variable). We can also write the distribution of y_i as $(y_i | x_i \ ; \theta) \sim \mathcal{N}(\theta^T x_i, \sigma^2)$

Given X (the design matrix, which contains all the x_i's) and θ, what is the distribution of the y_i's? The probability of the data is given by $p(\vec{y} | X; \theta)$. This quantity is typically viewed a function of $\vec{y}$ (and perhaps X), for a fixed value of θ. When we wish to explicitly view this as a function of θ, we will instead call it the **likelihood** function

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

Instead of maximizing $\mathbf{L}(\theta)$, we can also maximize any strictly increasing function of $\mathbf{L}(\theta)$. In particular, the derivations will be a bit simpler if we instead maximize the **log likelihood**

$$\begin{aligned}
\ell(\theta) = \log \mathbf{L}(\theta) &= \log\left[\prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \theta^T x^{(i)})^2}{2\sigma^2}\right)\right] \\
&= \sum_{i=1}^{m} \log\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_i - \theta_i^x)^2}{2\sigma^2}\right]\right) \\
&= m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2} \cdot \frac{1}{2} \sum_{i=1}^{m} (y_i - \theta^T x_i)^2
\end{aligned}$$

Hence, maximizing gives the same answer as minimizing $\ell(\theta)$

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

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

In an alternative way let the data is $\mathcal{D} = (x_i, y_i)_{i=1}^{n}$
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**.

$$\begin{aligned}
\theta^* &= \underset{\theta}{argmax}\ \ \mathbf{L}(\theta|\mathcal{D}) \\
&= \underset{\theta}{argmax}\ \ p(\mathcal{D}|\theta) \\
&= \underset{\theta}{argmax}\ \ p(y_1, x_1, y_2, x_2, y_3, x_3, \ldots, y_m, x_m\ ;\ \theta) \\
&= \underset{\theta}{argmax} \prod_{i=1}^{m} p(y_i, x_i\ ; \theta) && \text{[as } (y_i, x_i)\text{'s are independent]} \\
&= \underset{\theta}{argmax} \prod_{i=1}^{m} p(y_i|x_i\ ; \theta).p(x_i\ ; \theta) \\
&= \underset{\theta}{argmax} \prod_{i=1}^{m} p(y_i|x_i\ ; \theta).p(x_i) && \text{[as } x_i\text{'s are independent of } \theta)] \\
&= \underset{\theta}{argmax} \prod_{i=1}^{m} p(y_i|x_i\ ; \theta) \\
&= \underset{\theta}{argmax} \sum_{i=1}^{m} log\ [p(y_i|x_i\ ; \theta)]
\end{aligned}$$

$$
\begin{aligned}
&= \underset{\theta}{argmax} \sum_{i=1}^{m} \log\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(y_i-\theta^T x_i)^2}{2\sigma^2}\right]\right) \qquad [\text{ from } * \text{ we have this }]\\
&= \underset{\theta}{argmax}\ m\log\frac{1}{\sqrt{2\pi}\sigma}-\frac{1}{\sigma^2}\ \cdot\ \frac{1}{2}\sum_{i=1}^{m}(y_i-\theta^T x_i)^2\\
&= \underset{\theta}{argmin}\ \frac{1}{m}\sum_{i=1}^{m}(y_i-\theta^T x_i)^2
\end{aligned}
$$

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