ML Assignment 1

March 1, 2021

Theorem: Prove that under Gaussian noise assumption linear regression amounts to least square.

Proof: Let us assume that the target variables and the inputs are related via the equation

$$y_i = \theta^T x_i \; + \; \epsilon_i \quad \text{for } i = 1\,(1)\; m$$

where ϵ_i is an error term.

Let us further assume that the ϵ_i are distributed IID according to a Gaussian distribution with mean zero and some variance σ i.e.,

$$\epsilon_i \; \sim \; \mathcal{N}(0, \sigma^2)$$

i.e., the density of ϵ_i is given by

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

This implies that

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

To check the distribution of y_i's given x_i's for a fixed value θ we have the likelihood function

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

In order to make the data as high probability as possible we have to maximize $\mathbf{L(\theta)}$ over θ.

Let $\ell(\theta)$=log $\mathbf{L}(\theta)$

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

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

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

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