ML Assignment 1

Oishik Dasgupta

25 February 2021

Theorem: Prove that under gaussian noise assumption linear regression amounts to least squares.

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

$$y_i = \theta^T x_i + \epsilon_i$$

where ϵ_i is an error term that captures either unmodelled effects or random noise. Let us further assume that the ϵ_i are distributed IID according to a Gaussian distribution with mean zero and some variance σ^2. We can write this assumption as

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

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

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

The notation "$p\left(y_i | x_i; \theta\right)$" indicates that this is the distribution of y_i given x_i and parameterized by θ. Note that we should not condition on $\theta\left(p\left(y_i | x_i; \theta\right)\right)$, since θ is not a random variable.We can also write the distribution of y_i as $y_i | x_i; \theta \sim N\left(\theta^T x_i, \sigma^2\right)$.
Given X(the design matrix, which contains all the x_i's) and θ, what is the distribution of the y_i's? 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:

$$L(\theta) = L(\theta; X, \vec{y}) = p(\vec{y}|X; \theta)$$

Note that by the independence assumption on the ϵ_i's (and hence also the y_i's given the x_i's), this can also be written

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

Now given this probabilistic model relating the y_i's and the x_i's, what is a reasonable way of choosing our best guess of the parameters θ? The principle of maximum likelihood says that we should choose θ so as to make the data as high probability as possible. i.e.,we should choose θ to maximize L(θ). Instead of maximizing L(θ),we can maximize any strictly increasing function of L(θ). In particular, the derivations will be a bit simpler if we instead maximize the log likelihood l(θ):

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

Hence, maximizing l(θ) gives the same answer as minimizing

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

which we recognize to be J(θ),our original least-squares cost function.
To summarize: Under the previous probabilistic assumptions on the data, least-squares regression corresponds to finding the maximum likelihood estimate of θ.This is thus one set of assumptions under which least-squares regression can be justified as a very natural method that's just doing MLE.