MACHINE LEARNING ASSIGNMENT 1

Shashwata Mondal

February 2021

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

Proof: When faced with a regression problem, why might linear regression, and specifically why might the least-squares cost function J, be a reasonable choice? In this section, we will give a set of probabilistic assumptions, under which least-squares regression is derived as a very natural algorithm. Let us assume that the target variables and the inputs are related via the equation

$$y^{(i)} = \theta^T x^{(i)} \ + \ \epsilon^{(i)}$$

where $\epsilon^{(i)}$ is an error term that captures either unmodeled effects (such as if there are some features very pertinent to predicting housing price, but that we'd left out of the regression), or random noise. Let us further assume that the $\epsilon^{(i)}$ are distributed IID (independently and identically distributed) according to a Gaussian distribution (also called a Normal distribution) with
mean zero and some varianceσ. We can write this assumption as

$$\epsilon^{(i)} \ \sim \ \mathcal{N}(0, \sigma^2)$$

I.e., the density of $\epsilon^{(i)}$ 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

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

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

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

$$= \prod_{i=1}^{m} p(y^i|x^i;\theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

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**

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

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

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

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