Linear Regression

Let's take a look at a probabilistic interpretation of linear regression. Why do we use least squares? What's the purpose of using squared error?
Let's assume that

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

where $\epsilon^{(i)}$ is the error term that includes unmodeled effects and random noise. We also assume that

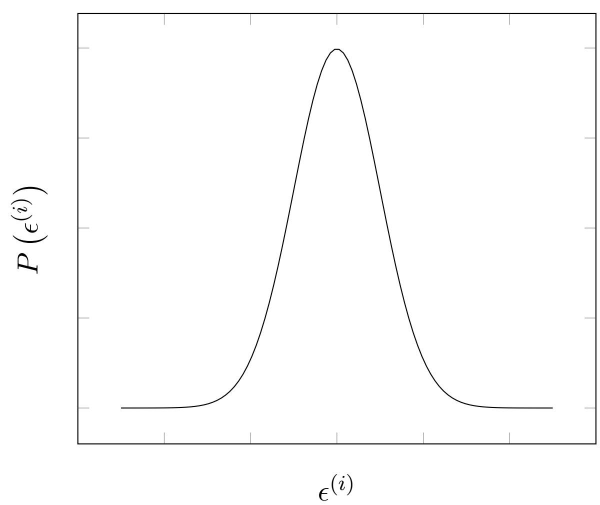
$$\epsilon^{(i)} \sim \mathcal{N}\left(0, \sigma^2\right)$$

What this means is the probability density of the error $P\left(\epsilon^{(i)}\right)$ is equivalent to the Gaussian density equation with a mean of 0 and standard deviation σ (σ^2 is the variance)

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(\epsilon^{(i)}\right)^2}{2\sigma^2}\right)$$

which is a function that integrates to 1. We can see this distribution represented as a Gaussian as such

Gaussian Distribution



Under this set of assumptions, it is implied that

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

since $y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$. Another way of representing this is with the distribution relation

$$y^{(i)} \mid x^{(i)}; \theta \sim \mathcal{N}\left(\theta^T x^{(i)}, \sigma^2\right)$$

which tells us that our random variable is $y^{(i)}$ and our mean is $\theta^T x^{(i)}$ with a variance of σ^2. Therefore, we've concluded that, based on the previous assumptions, the probability density of $y^{(i)}$ given $x^{(i)}$ parameterized by θ follows a Gaussian distribution.

Under the assumptions we just made, the likelihood of the parameters $\mathcal{L}(\theta)$ is defined as the probability of the data $P(\vec{y} \mid x; \theta)$ which is equivalent to

$$\prod_{i=1}^{m} P\left(y^{(i)} \mid x^{(i)}; \theta\right)$$

Substituting for the probability function in the product for what we determined previously,

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

where, as before, m is the number of training samples, $\vec{y}$ is our target vector, x is our feature matrix, and θ is our set of parameters.

We then define the log likelihood of the parameters $\ell(\theta)$ to be $\log \mathcal{L}(\theta)$. Following the behavior of logarithm, we know that

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

is the same as

$$\sum_{i=1}^{m} \log\left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(y^{(i)} - \theta^T x^{(i)}\right)^2}{2\sigma^2}\right)\right]$$

because the log of a product is the same as the sum of the log of every term in the product. This can also be applied further to the expression, resulting in

$$\sum_{i=1}^{m} \log \frac{1}{\sigma\sqrt{2\pi}} + \sum_{i=1}^{m} \log\left[\exp\left(-\frac{\left(y^{(i)} - \theta^T x^{(i)}\right)^2}{2\sigma^2}\right)\right]$$

We see that the sum term on the left does not include i, so it is simply just added to itself m times which we know as the definition of multiplication. Therefore this is equivalent to

$$m \log \frac{1}{\sigma\sqrt{2\pi}} + \sum_{i=1}^{m} \log\left[\exp\left(-\frac{\left(y^{(i)} - \theta^T x^{(i)}\right)^2}{2\sigma^2}\right)\right]$$

which simplifies to

$$m \log \frac{1}{\sigma\sqrt{2\pi}} - \sum_{i=1}^{m} \frac{\left(y^{(i)} - \theta^T x^{(i)}\right)^2}{2\sigma^2}$$

When we want to get the maximum likelihood estimation, preferably we should use the log likelihood because it is a strictly monotonically increasing function and should also result in maximizing the likelihood. When we look back at our log likelihood equation, $m \log \frac{1}{\sigma\sqrt{2\pi}}$ is a constant and our second term is negative but dependent on θ. We can ignore σ because it is a constant. To maximize the likelihood, we want to minimize

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

which we know is our cost function $J(\theta)$, which circles back to why we want to use squared error.