Gaussian Error Assumption Over Linear Regression

Dristanta Das

February 2021

1 Introduction

Linear regression attempts to model the relationship between two variables by fitting a line to the observed data. One variable is considered to be the independent variable and the other is considered to be the dependent variable.

2 Probabilistic Modelling

2.1 Linear Model

$$y_i \simeq \theta^T x_i$$

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

where,

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

Here ϵ_i's are i.i.d random variables.

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

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

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

However the conventional way is,

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

2.2 Parameter Estimation

Suppose, we are given with a dataset,

$$\mathcal{D} = \{x_i, yi\}_{i=1}^{m}$$

Then, Bayes' Theorem States that,

$$\begin{aligned} P(\theta|\mathcal{D}) &= P(\mathcal{D}|\theta).P(\theta) \\ &= P(\theta, \mathcal{D})\frac{1}{P(\mathcal{D})} \end{aligned} \tag{1}$$

2.3 Maximum Likelihood Estimation (MLE)

The idea behind maximum likelihood estimation (MLE) is to define a function of the parameters that enables us to find a model that fits the data well. The estimation problem is focused on the likelihood function, or more precisely its negative logarithm. For data represented by a random variable x and for a family of probability densities $p(\, x \mid \, \theta \,)$ parametrized by θ , the negative log-likelihood is given by

$$\mathcal{L}_x(\theta) = -logp(\, x \mid \, \theta \,)$$

The notation $\mathcal{L}_x(\theta)$ emphasizes the fact that the parameter θ is varying and the data x is fixed. We very often drop the reference to x when writing the negative log-likelihood, as it is really a function of θ , and write it as $\mathcal{L}(\theta)$ when the random variable representing the uncertainty in the data is clear from the context. Let us interpret what the probability density $p(\, x \mid \, \theta \,)$ is modeling for a fixed value of θ . It is a distribution that models the uncertainty of the data. In other words, once we have chosen the type of function we want as a predictor, the likelihood provides the probability of observing data x . In a complementary view, if we consider the data to be fixed (because it has been observed), and we vary the parameters θ , what does $\mathcal{L}(\theta)$ tell us? It tells us how likely a particular setting of θ is for the observations x.Based on this second view, the maximum likelihood estimator gives us the most likely parameter θ for the set of data. We consider the supervised learning setting, where we obtain pairs (x_1, y_1), . . . , (x_N, y_N) with and labels $y_N \in R^n$. We are interested in constructing a predictor that takes a feature vector x_n as input and produces a prediction y n (or something close to it), i.e., given a vector x_n we want the probability distribution of the label y_n . In other words, we specify the conditional probability distribution of the labels given the examples for the particular parameter setting θ .

We assume that the set of examples (x_1, y_1), . . . , (x_N, y_N) are independent and identically distributed (i.i.d.). The word "independent" implies that the likelihood of the whole dataset $\mathrm{Y} = y_1, ..., y_N$ and $\mathrm{X} = x_1, ..., x_N$ factorizes into a product of the likelihoods of each individual example,

$$p(Y|X, \theta) = \prod_{n=1}^{N} p(y_n|x_n, \theta)$$

where $p(y_n|x_n, \theta)$ is a particular distribution (which was Gaussian). The expression "identically distributed" means that each term in the product, is of the same distribution, and all of them share the same parameters. It is often easier from an optimization viewpoint to compute functions that can be decomposed into sums of simpler functions. Hence, in machine learning we often consider the negative log-likelihood,

$$\mathcal{L}(\theta) = -log\ p(Y|X, \theta) = -\sum_{n=1}^{N} log\ p(y_n|x_n, \theta)$$

While it is temping to interpret the fact that θ is on the right of the conditioning in $p(y_n|x_n, \theta)$, and hence should be interpreted as observed and fixed, this interpretation is incorrect. The negative log-likelihood $\mathcal{L}(\theta)$ is a function of θ . Therefore, to find a good parameter vector θ that explains the data (x_1, y_1), . . . , (x_N, y_N) well, minimize the negative log- likelihood $\mathcal{L}(\theta)$ with respect to θ .

Remark. The negative sign is a historical artifact that is due to the convention that we want to maximize likelihood, but numerical optimization literature tends to study minimization of functions.

We will make a Gaussian model assumption with $\theta \in R^n$

$$
\begin{aligned}
\theta^* &= \underset{\theta}{\text{argmax}}\ \mathcal{L}(\theta|\mathcal{D}) \\
&= \underset{\theta}{\text{argmax}}\ P(\mathcal{D} \mid \theta) \\
&= \underset{\theta}{\text{argmax}}\ P(y_1, x_1, ..., y_m, x_m; \theta) \\
&= \underset{\theta}{\text{argmax}} \prod_{i=1}^{m} P(y_i, x_i; \theta) \\
&= \underset{\theta}{\text{argmax}} \prod_{i=1}^{m} [P(y_i|x_i; \theta).P(x_i; \theta] \\
&= \underset{\theta}{\text{argmax}} \prod_{i=1}^{m} [P(y_i|x_i; \theta)].P(x_i) \\
&= \underset{\theta}{\text{argmax}} \prod_{i=1}^{m} [P(y_i|x_i; \theta)] \\
&= \underset{\theta}{\text{argmax}} \sum_{i=1}^{m} logP(y_i|x_i; \theta) \\
&= \underset{\theta}{\text{argmax}} \sum_{i=1}^{m} [log(\frac{1}{\sqrt{2\pi}\sigma}) + log(exp(-\frac{(\theta^T x_i - y_i)^2}{2\sigma^2}))] \\
&= \underset{\theta}{\text{argmax}}\ -\frac{1}{2\sigma^2} \sum_{i=1}^{m} (\theta^T x_i - y_i)^2 \\
&= \underset{\theta}{\text{argmin}}\ \frac{1}{n} \sum_{i=1}^{m} (\theta^T x_i - y_i)^2
\end{aligned}
\tag{2}
$$

3 Conclusion

Hence, Under Gaussian error assumption linear regression amounts to least square.