Notes On Linear Regression

CSE 446: Machine Learning
Autumn 2019

1 The Maximum Likelihood Estimator

In the linear regression lectures, we presented our supervised learning problem in terms of a loss function that scores our predictions relative to the ground truth as determined by some training data. However, an alternative view is to think about a probability procedure that might have given rise to the data. This probability procedure would have some parameters, and our job is to figure out which parameters would assign the highest probability to the data that was observed. This is the principle of maximum likelihood estimation.

To help us understand this concept more concretely, let's say that we have N training samples, $x_1, x_2, ..., x_N$, and N labels, $y_1, y_2, ..., y_N$, where y_i is the label of x_i. Furthermore, let's say that each label y_i is independently and identically generated using the equation $y_i = w^T x_i + \epsilon_i$, where $\epsilon_i = N(w^T x_i, \sigma^2)$. Mathematically, the goal of MLE is to find a $\hat{w}$ that maximizes the probability of the labels, $y_1, y_2, ..., y_N$, given the inputs, $x_1, x_2, ..., x_N$.

Recall that the probability density function of the normal distribution whose mean is μ and variance is σ^2 is

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Bigg(- \frac{(x-\mu)^2}{2\sigma^2} \Bigg)$$

Let's denote the likelihood of seeing a sequence of labels D. Since the inputs are i.i.d., we know that

$$D(y_1, y_2, ..., y_N | x_1, x_2, ..., x_N, w, \epsilon_1, \epsilon_2, ..., \epsilon_N) = \prod_{i=1}^{N} D(y_i | x_i, w, \epsilon_i)$$

Then,

$$\begin{aligned} \hat{w}_{MLE} &= \text{argmax}_w \prod_{i=1}^{N} P_w(y_i | x_i, w, \epsilon_i) \\ &= \text{argmax}_w log\Bigg(\prod_{i=1}^{N} P_w(y_i | x_i, w, \epsilon_i) \Bigg) \\ &= \text{argmax}_w \sum_{n=1}^{N} log\Big(P_w(y_i | x_i, w, \epsilon_i) \Big) \\ &= \text{argmin}_w \sum_{n=1}^{N} -log\Big(P_w(y_i | x_i, w, \epsilon_i) \Big) \end{aligned} \tag{1}$$

According to our normal distribution model, we know that

$$P_w(y_i | x_i, w, \epsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Bigg(- \frac{(w^T x_i - y_i)^2}{2\sigma^2} \Bigg) \tag{2}$$

Plugging (2) into (1), we get

$$\hat{w}_{MLE} = \text{argmin}_w \frac{1}{N} \sum_{n=1}^{N} (y_i - w^T x_i)^2$$

2 Closed-Form Solution

If we think of

- the training data as a large matrix X of size $N \times D$, where $X_{n,d}$ is the value of the dth feature on the nth sample
- the labels as a column vector Y of dimension N
- the weights as a column vector w of dimension D

we could rewrite w_{MLE} as

$$\hat{w}_{MLE} = \text{argmin}_w \frac{1}{N} \left\| Xw - Y \right\|^2$$

For notational convenience, let's define

$$L(w) = \sum_{n=1}^{N} (y_i - w \cdot x_i)^2 = \left\| Xw - Y \right\|^2$$

Therefore, now, in order to find $\hat{w}_{MLE}$, we need to minimize $L(w)$. How do we minimize a function? We take its derivative and set it to 0!

Recall that, for an arbitrary vector x, its l2-norm can be written as

$$\begin{aligned} f(x) &= \|x\|_2^2 \\ &= \left(\left(\sum_{k=1}^{N} {x_k}^2 \right)^{\frac{1}{2}} \right)^2 \\ &= \sum_{k=1}^{N} {x_k}^2 \end{aligned}$$

Then, it follows that

$$\begin{aligned} \frac{\partial}{\partial x_j} f(x) &= \frac{\partial}{\partial x_j} \sum_{k=1}^{N} {x_k}^2 \\ &= \sum_{k=1}^{N} \frac{\partial}{\partial x_j} {x_k}^2 \\ &= 2x_j \end{aligned}$$

Then, it follows that

$$\nabla f(x) = 2x$$

Since $Xw - Y$ is a vector, we see that

$$\begin{aligned} \nabla_w \big(L(w)\big) &= 2X^T(Xw - Y) \\ &= 2X^T Xw - 2X^T Y \end{aligned} \tag{3}$$

Equating (3) to 0, we get

$$\begin{aligned} 2X^T Xw - 2X^T Y &= 0 \\ X^T Xw - X^T Y &= 0 \\ X^T Xw &= X^T Y \\ \hat{w}_{MLE} &= (X^T X)^{-1} X^T Y \end{aligned}$$

This is the closed-form solution to our linear regression problem.

3 Error Analysis

Now, let's think about how to evaluate how well our model is doing.

Suppose that $y_i = {x_i}^T w + \epsilon_i$, where each ϵ_i is drawn i.i.d from $N(0, \sigma^2)$. Then, it follows that

$$\begin{aligned} \hat{w}_{MLE} &= (X^T X)^{-1} X^T \hat{Y} \\ &= (X^T X)^{-1} X^T (Xw + \epsilon) \\ &= w + (X^T X)^{-1} X^T \epsilon \end{aligned}$$

Recall that $MSE_{train} = \left\| \hat{Y} - Y \right\|_2^2$. Therefore,

$$\begin{aligned} MSE_{train} &= E\left[\left\| \hat{Y} - Y \right\|_2^2 \right] \\ &= E\left[\left\| X(X^T X)^{-1} X^T \epsilon \right\|_2^2 \right] \end{aligned}$$

For notational convenience, let's define

$$A = X(X^T X)^{-1} X^T$$

Then,

$$\begin{aligned} MSE_{train} &= E\Big[\, \|A\epsilon\|^2 \Big] \\ &= E\Big[\epsilon^T A^T A\epsilon\Big] \\ &= E\Bigg[\sum_i \Big(\sum_j \epsilon_j A_{i,j}\Big)^2\Bigg] \\ &= E\Bigg[\sum_i \sum_{j_1} \sum_{j_2} \epsilon_{i,j_1} A_{i,j_1} \epsilon_{j_2} A_{i,j_2}\Bigg] \\ &= E\Bigg[\Big((A\epsilon)^T A\epsilon\Big)_{i,j}\Bigg] \\ &= E\Big[\text{Trace}(A\epsilon\epsilon^T A^T)\Big] \end{aligned}$$

Notice that $E[\epsilon\epsilon^T] = \sigma^2\mathbb{I}$. Therefore,

$$MSE_{train} = \sigma^2 \ \text{Trace}(AA^T)$$

Now, what if each error ϵ_i is drawn i.i.d. from a Laplace distribution? Recall that the probability density function for Laplace distribution is

$$\frac{1}{2b} \exp\Big(- \frac{\|x-\mu\|_{L1}}{b}\Big)$$

To make the math prettier, let's assume that $\mu = 0$, in which case the probability density function of the Laplace distribution becomes

$$\frac{1}{2b} \exp\Big(- \frac{\|x\|_{L1}}{b}\Big)$$

[Exercise]

a. What is $\hat{w}_{MLE}$ under the Laplace distribution?

b. Does $\hat{w}_{MLE}$ have an analytical (closed-form) solution?