Lecture 7: Maximum Likelihood Estimation

CSE 427: Machine Learning

Md Zahidul Hasan

Lecturer, Computer Science
BRAC University

Spring 2023

Maximum Likelihood Estimation

Let's say, our dataset contains the following instances, $x_1, x_2, \cdots, x_m$ picked from an unknown normal/Gaussian distribution. The parameters that generate a normal distribution is the mean μ and the standard deviation σ. We want to estimate the values of μ and σ that makes the observation of $x_1, x_2, \cdots, x_m$ most likely. In other words, we want to find that μ and σ that maximizes the following:

$$\begin{aligned}\arg\max_{\mu,\sigma} P(X) &= \arg\max_{\mu,\sigma} p(x_1) \times p(x_2) \times \cdots \times p(x_m) \\ &= \arg\max_{\mu,\sigma} \Pi_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\ &= \arg\max_{\mu,\sigma} (\frac{1}{\sqrt{2\pi\sigma}})^m e^{-\frac{\sum_{i=1}^{m}(x_i-\mu)^2}{2\sigma^2}}\end{aligned}$$

Let's call the right hand side, our objective function J. If we differentiate the objective function with respect to μ and σ, and solve the equations $\frac{\partial J}{\partial \mu} = 0$ and $\frac{\partial J}{\partial \sigma} = 0$, we find that, $\mu = \frac{\sum_{i=1}^{m} x_i}{m}$ and $\sigma = \sqrt{\frac{\sum_{i=1}^{m}(x_i-\mu)^2}{m}}$. These values maximize J because the second differentials are negative.

Conditional Log-Likelihood

In supervised learning, for each x, we are given some target y. So, it makes more sense to find the parameters w that maximize the following:

$\arg\max_w P(Y|X,w) = \arg\max_w p(y_1|x_1,w) \times \cdots \times p(y_m|x_m,w)$

Let's discuss linear regression our prediction is a linear function, $\bar{y}_i = w_1x_1 + w_2x_2 + \cdots + w_nx_n + w_{n+1} = \langle w, x\rangle = h_w(x_i)$.

And we know that the error $\epsilon_i = y_i - \bar{y}_i$ follows a Gaussian distribution with some $\mu = 0$ and σ. Therefore,

$$p(y_i|x_i,w) = p(y_i - h_w(x_i)|x_i,w) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y_i - h_w(x_i))^2}{2\sigma^2}}.$$

$$\text{So, } J = P(Y|X,w) = \Pi_{i=1}^{m}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y_i - h_w(x_i))^2}{2\sigma^2}}$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^m e^{-\frac{\sum_{i=1}^{m}(y_i - h_w(x_i))^2}{2\sigma^2}}$$

An Objective Function For Linear Regression

Now, if we take natural log on both sides, our objective function becomes $\log J = -m \log \sqrt{2\pi}\sigma - \frac{\sum_{i=1}^{m}(y_i - h_w(x_i))^2}{2\sigma^2}$
The above term $\log P(Y|X, w)$ is called the conditional log likelihood. In most cases of supervised learning, this serves as an objective function to maximize.
$\log P(Y|X, w) = -m \log \sqrt{2\pi}\sigma - \frac{\sum_{i=1}^{m}(y_i - h_w(x_i))^2}{2\sigma^2}$.
In the case of linear regression, notice that the right hand side can be maximized only when $\sum_{i=1}^{m}(y_i - h_w(x_i))^2$ is minimized. All of the other terms are constant. So, the objective function in linear regression is:

$$\arg\min_w \sum_{i=1}^{m}(y_i - h_w(x_i))^2$$

So, we have to minimize the sum of squared errors.

An Objective Function for Logistic Regression

In logistic regression, our hypothesis is: $h_w(x) = \frac{1}{1+e^{-\langle w,x \rangle}}$.
Therefore, $p(y=1|x,w) = h_w(x)$ and $p(y=0|x,w) = 1 - h_w(x)$.
Combining these two, we get the following Bernoulli distribution,

$$p(y|x,w) = h_w(x)^y (1 - h_w(x))^{1-y}$$

So, $p(Y|X,w) = \Pi_{i=1}^{m} p(y_i|x_i,w) = \Pi_{i=1}^{m} h_w(x_i)^{y_i}(1 - h_w(x_i))^{1-y_i}$.
If we take log on both sides,
$\log p(Y|X,w) = \sum_{i=1}^{m} y_i \log h_w(x_i) + (1 - y_i) \log (1 - h_w(x_i))$.
This is the negation of the cross entropy. Let's say,
$A = (a_1, a_2, \cdots, a_n)$ and $B = (b_1, b_2, \cdots, b_n)$ are two probability distributions. Then their cross entropy is defined as follows:

$$H(A,B) = -\sum_{i=1}^{n} a_i \log b_i.$$

This cross entropy is minimized when $a_i = b_i$ for all $i \in \{1, 2, \cdots, n\}$. So, in the case of logistic regression, our goal is to minimize the cross entropy.

An Objective Function for Softmax Regression

In softmax regression, $p(y = j|x, w) = \bar{y}_j = \frac{e^{h_{w_j}(x)}}{\sum_{i=1}^{k} e^{h_{w_i}(x)}}$. These $\bar{y}_j$ create a probability distribution. $\bar{y}_j$ is our estimated probability that the output class is j. For a particular instance i, the target vector y can be as follows, $y = (0, 1, 0, \cdots, 0)$ which is a one-hot vector. We can see that the target class is class 2. And our probability distribution across the classes can be as follows $\bar{y} = (0.1, 0.2, 0.0014, \cdots, 0.1)$. Using this we get the following Multinoulli distribution (also called categorical distribution):

$$p(y|x, w) = \bar{y}_1^{y_1}.\bar{y}_2^{y_2} \cdots \bar{y}_k^{y_k}$$

If we take log on both sides, we get, $\log p(y|x, w) = \sum_{i=1}^{k} y_i \log \bar{y}_i$. Which is the negation of the cross entropy. Now let's find out, $P(Y|X, w) = \Pi_{i=1}^{m} p(y_i|x_i, w)$.
Taking log on both sides,
$\log P(Y|X, w) = \sum_{i=1}^{m} \log p(y_i|x_i, w) = \sum_{i=1}^{m} \sum_{j=1}^{k} y_{i,j} \log \bar{y}_{i,j}$
$= \sum_{i=1}^{m} -H(y_i, \bar{y}_i)$.