07-280: Maximum Likelihood Estimation

Nihar B. Shah

Motivation

Maximum Likelihood Estimation (MLE) is a powerful technique to derive ML methods. As we will see below, MLE can be used to derive logistic regression and linear regression from scratch. It is not just confined to linear/logistic regression, but is a much more general method. The recipe involves four steps:

1. Assume a probabilistic model on how data is generated.
2. Collect training data.
3. Estimate parameters of the assumed model using the data.
4. Use this model with the estimated parameters to make predictions.

Let us now go into specific detail. There are many applications where:

- You have data comprising pairs $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$.
- You have a probabilistic model (i.e., assumptions) on how any $y^{(i)}$ is generated given the $x^{(i)}$.
- The probabilistic model is associated with some parameters $\theta^* \in \Theta$ which are not known to you.
- Your goal is to estimate θ^*.

Simple Example

You have a potentially biased coin. You toss it n times and observe the outcomes. Your goal is to estimate the bias of the coin.

In terms of our framework above: Let us capture the outcomes of the tosses in $y^{(1)}, \ldots, y^{(n)}$ with $y^{(i)} = 1$ if the i^{th} outcome was heads and 0 if tails. There are no "inputs" here so simply set $x^{(1)} = \cdots = x^{(n)} = \text{null}$. Let

$$\mathbb{P}(y = 1) = \theta^*, \qquad \mathbb{P}(y = 0) = 1 - \theta^*.$$

This is our probabilistic model. Finally $\Theta = [0, 1]$, i.e., $\theta^* \in [0, 1]$. Assume all tosses are independent.

Now suppose you observe that in $n = 100$ tosses, there were 60 heads. What would be your estimate of θ^*?

A reasonable guess would be the fraction of heads, i.e., 0.6.

Is there a principled way to determine this? Can we use the same principles to solve problems that are much more general and complex?

Maximum Likelihood Estimation (MLE)

Informally: What value of the parameters would make the observed data (y's) most likely? Formally:

$$\arg\max_{\theta\in\Theta} \mathbb{P}(y^{(1)}, y^{(2)}, \ldots, y^{(n)} \; ; \; \theta, \, x^{(1)}, x^{(2)}, \ldots, x^{(n)})$$

The term $\mathbb{P}()$ above captures the probability of observing $y^{(1)}, \ldots, y^{(n)}$ under the assumed probabilistic model when the underlying parameter is θ and inputs are $x^{(1)}, \ldots, x^{(n)}$.

If we assume that all n samples are independent, this becomes

$$\arg\max_{\theta\in\Theta} \prod_{i=1}^{n} \mathbb{P}(y^{(i)} \; ; \; \theta, x^{(i)}).$$

Products are messy to deal with. Sums are nicer. Let's do something about it. Suppose you want to do $\arg\max_{\theta\in\Theta} g(\theta)$ for some function $g : \Theta \to \mathbb{R}$. Then for any function $f : \mathbb{R} \to \mathbb{R}$ that is strictly increasing,

$$\arg\max_{\theta\in\Theta} g(\theta) = \arg\max_{\theta\in\Theta} f(g(\theta)).$$

Using log (any base) as f, we get an equivalent MLE:

$$\arg\max_{\theta\in\Theta} \log\left(\prod_{i=1}^{n} \mathbb{P}(y^{(i)} \; ; \; \theta, x^{(i)})\right)$$

$$= \boxed{\arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \log \mathbb{P}(y^{(i)} \; ; \; \theta, x^{(i)})}$$

Coin Example (continued)

Coming back to our simple coin setting:

$$\mathbb{P}(y^{(i)}; \, \theta, \, x^{(i)}) = \begin{cases} \theta & \text{if } y^{(i)} = 1 \\ 1-\theta & \text{if } y^{(i)} = 0 \end{cases} = \theta^{\,y^{(i)}}(1-\theta)^{1-y^{(i)}}.$$

MLE:

$$\arg\max_{\theta\in[0,1]} \sum_{i=1}^{n} \Big[(1-y^{(i)})\log(1-\theta) + y^{(i)}\log\theta\Big]$$

$$= \arg\max_{\theta\in[0,1]} \left(\sum_{i=1}^{n} y^{(i)}\right)\log\theta + \left(n - \sum_{i=1}^{n} y^{(i)}\right)\log(1-\theta)$$

Taking derivatives with respect to θ and setting to 0, we get

$$\theta = \frac{\sum_{i=1}^{n} y^{(i)}}{n}.$$

A More Complex Setting

Now suppose you are testing the durability of some material. You want to figure out its breakability with respect to the ambient temperature and pressure. You apply various temperatures (first coordinate of $x^{(i)}$) and pressures (second coordinate of $x^{(i)}$) and observe whether it breaks ($y^{(i)} = 1$) or not ($y^{(i)} = 0$).

Model: There are some unknown parameters $\theta_1, \theta_2 \in \mathbb{R}$. *Breakability* is determined by $\theta_1 x_1^{(i)} + \theta_2 x_2^{(i)}$. Actual breaking is stochastic: breaks with probability = sigmoid(*breakability*). In our framework we thus have

$$\mathbb{P}(y^{(i)} \,;\, \theta,\, x^{(i)}) = \begin{cases} \text{sigmoid}(x^{(i)\top}\theta) & \text{if } y^{(i)} = 1 \\ 1 - \text{sigmoid}(x^{(i)\top}\theta) & \text{if } y^{(i)} = 0 \end{cases}$$

$$\log \mathbb{P}(y^{(i)};\, \theta,\, x^{(i)}) = \begin{cases} \log(\text{sigmoid}(x^{(i)\top}\theta)) & \text{if } y^{(i)} = 1 \\ \log(1 - \text{sigmoid}(x^{(i)\top}\theta)) & \text{if } y^{(i)} = 0 \end{cases}$$

$$= y^{(i)} \log(\text{sigmoid}(x^{(i)\top}\theta)) + (1 - y^{(i)}) \log(1 - \text{sigmoid}(x^{(i)\top}\theta))$$

MLE:

$$\arg\max_\theta \sum_{i=1}^{n} \left[y^{(i)} \log(\text{sigmoid}(x^{(i)\top}\theta)) + (1 - y^{(i)}) \log(1 - \text{sigmoid}(x^{(i)\top}\theta)) \right]$$

$$= \arg\min_\theta -\sum_{i=1}^{n} \left[y^{(i)} \log(\text{sigmoid}(x^{(i)\top}\theta)) + (1 - y^{(i)}) \log(1 - \text{sigmoid}(x^{(i)\top}\theta)) \right]$$

This is **Logistic Regression!**

Real-Valued y's

Now instead of observing a binary variable of whether the material broke or not, suppose you can make a measurement of the breakability $\theta^\top x^{(i)}$ directly. However, any measuring device has some noise, and so you actually observe

$$y^{(i)} = \theta^\top x^{(i)} + z^{(i)}$$

where $z^{(i)}$ is a random variable distributed Normally with mean 0 and variance σ^2. A normal (also called Gaussian) distribution with mean μ and variance σ^2 is

$$\mathbb{P}(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2} \right).$$

For MLE:

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

MLE:

$$\arg\max_\theta \sum_{i=1}^{n} -(y^{(i)} - \theta^\top x^{(i)})^2 = \arg\min_\theta \sum_{i=1}^{n} (y^{(i)} - \theta^\top x^{(i)})^2$$

This is **Linear Regression!**

We see that we can derive logistic and linear regression via the MLE principle. *Finally, note that this MLE approach is far more general and widely applicable.*