67750: Advanced Practical Machine Learning 2018/9

Maximum Likelihood Estimation

TA: Daniel Gissin

1 Maximum Likelihood Estimation

Maximum likelihood estimation (MLE) is one of the more fundamental concepts in statistics. Here, instead of trying to learn a discriminative objective (classifying between labels), we are instead trying to learn the underlying distribution of the data.

This means that instead of trying to learn $\mathbb{P}(y|x)$, we ignore the y (or treat it as another feature) and directly try to learn $\mathbb{P}(x)$ (or the joint probability $\mathbb{P}(x, y)$). If we are able to do that, we can still perform classification tasks using the Bayes optimal classifier, but we can also do many other things which we will see during this course.

1.1 A Gentle Start - Estimating a Coin's Bias

Say we have a coin that when tossed has a chance of p to turn up heads. We want to estimate p.

So we toss the coin 100 times, and we end up getting 30 heads. If you had to guess p and you didn't have a reason to assume it should be anything in particular, you would probably guess $p = \frac{30}{100}$...

1.2 The Estimation Problem

Let's generalize the above problem. Say we have a sample S of N data points, and we want to learn the data distribution $\mathcal{D}$ such that $S \sim \mathcal{D}^N$. To do that, we need some prior knowledge:

We will assume that $\mathcal{D}$ is a member of a family of probability functions, parameterized by θ. We will then look for the $\mathcal{D}_\theta$ which best explains the given data. Intuitively, we would like to find the distribution which gives our sample the highest probability mass since if $\mathbb{P}(S \sim \mathcal{D}_{\theta_1}^N) > \mathbb{P}(S \sim \mathcal{D}_{\theta_2}^N)$, we should probably prefer θ_1...

This is precisely the maximum likelihood criterion. Given our sample S and our parameter θ, the Likelihood function returns the probability mass of drawing the sample S from $\mathcal{D}_\theta$:

$$\mathcal{L}(S, \theta) = \mathbb{P}(S \sim \mathcal{D}_\theta^N) = \prod_{i=1}^{N} p(x_i; \theta)$$

So maximizing the likelihood function with respect to θ should give us the best distribution D_θ.

1.3 The Log-Likelihood Trick

The likelihood function is multiplicative, which will make differentiating it quite annoying. Since the logarithm is a monotonically increasing function, it is usually easier to work with the Log-Likelihood function, which has the same argmax as the Likelihood function:

$$\ell(S,\theta) = log(\mathcal{L}(S,\theta)) = log(\prod_{i=1}^{N} p(x_i;\theta)) = \sum_{i=1}^{N} log(p(x_i;\theta))$$

$$\hat{\theta}_{MLE} = \underset{\theta}{\text{argmax}}(L(X,\theta)) = \underset{\theta}{\text{argmax}}(LL(X,\theta))$$

1.4 Returning to the Coin Toss

Let's solve our original problem using the MLE criterion. We have N coin tosses, and k turned out heads. Our family of distribution functions will be all Bernoulli distributions, $x_i \sim Ber(\theta)$, so our parameter is simply p. Our likelihood function will be:

$$\mathcal{L}(k,\theta) = \Big(\prod_{i=1}^{k} \theta\Big) \cdot \Big(\Pi_{i=1}^{N-k}(1-\theta)\Big) = \theta^k (1-\theta)^{N-k}$$

Our log likelihood function will be:

$$\ell(N,k,\theta) = k \cdot log(\theta) + (N-k) \cdot log(1-\theta)$$

Taking the derivative w.r.t. θ gives us our MLE:

$$0 = \frac{\partial LL}{\partial \theta} = \frac{k}{\theta} - \frac{N-k}{1-\theta} \rightarrow \hat{\theta}_{MLE} = \frac{k}{N}$$

Which is indeed what we would have guessed intuitively...

1.5 Caveats of the MLE

The MLE is a great tool for estimating the distribution that generated our data, but it has it's limitations. Specifically for our example, say we didn't have 100 coin tosses but only had 3. Even if the coin is completely fair, an eighth of the time we would say that it will always turn out heads...

This is a general problem with MLE, which is often bad for small samples.

1.6 MLE for a Gaussian Model

Let's look at another canonical example of maximum likelihood estimation - the 1-dimensional Gaussian:

$$\mathcal{N}(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Note that this parametric distribution has more than one parameter and so $\theta = (\mu, \sigma^2)$.

We are given N data points, the height of randomly selected people in Israel, and we assume that the heights follow a Gaussian distribution. We would like to find the most likely parameters for the mean and variance.

First, we will write down the Likelihood function for our sample:

$$\mathcal{L}(S, \mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2}$$

We want to maximize the Likelihood for both variables - μ and σ^2.

For μ we don't even have to look at the Log Likelihood, since maximizing the Likelihood w.r.t. μ simply means minimizing the exponent, which is simple:

$$\min_{\mu}(\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2) \rightarrow \min_{\mu}(\sum_{i=1}^{N}(x_i - \mu)^2)$$

$$\frac{\partial}{\partial\mu}(\sum_{i=1}^{N}(x_i - \mu)^2) = 0 = 2(\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \mu) \rightarrow \hat{\mu} = \frac{\sum_{i=1}^{N} x_i}{N} = \bar{x}$$

So as we might have expected, our mean turned out to be the empirical mean of our sample. Now we move on to the variance, so we'll use the Log Likelihood:

$$\ell(S, \mu, \sigma) = -\frac{N}{2}log(2\pi) - Nlog(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2$$

Simply differentiate w.r.t. σ to get our MLE:

$$\frac{\partial\ell}{\partial\sigma} = 0 = -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{N}(x_i - \mu)^2 \rightarrow \hat{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Also as we might have expected, we got the empirical variance as our MLE.

Finally, had we calculated the MLE of a multidimensional Gaussian, we would have gotten the following results (d is the dimension of our data):

$$\mathcal{N}(x; \overrightarrow{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^d det(\Sigma)}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

$$\hat{\mu} = \frac{\sum_{i=1}^{N} x_i}{N} = \bar{x}$$

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^T$$