Statistical ML Notes

Neil Thistlethwaite

28 May 2020

Maximum Likelihood Estimation

Maximum likelihood estimation, or MLE, is a statistical method for estimating the parameters of some distribution. For instance, suppose we have several data points x_1, x_2, …, x_n drawn from some distribution. If we know the family of distributions that it belongs to (e.g. Gaussian, Uniform, Binomial, etc.), then we can use MLE to give us the "most likely" parameters of the specific distribution the data was drawn from.

For instance, suppose we flip a coin with an unknown bias 20 times, recording whether each flip resulted in "HEADS" or "TAILS". Then, we can view each coin flip as one sample from a Bernoulli distribution (a Bernoulli distribution is parameterized by p, the "chance of success", and takes value 1 with probability p and value 0 with probability $1-p$). Using Maximum Likelihood Estimation, we will find the most likely estimate for p for this coin.

Specifically, we want to find $\hat{p}_{\text{MLE}}$, the best estimate for p given the data.

$$\hat{p} = \hat{p}_{\text{MLE}} = \operatorname*{arg\,max}_{p\in[0,1]} \mathrm{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \mid X \sim \text{Bernoulli}(p))$$

This simply says that the most likely p is the one that maximizes the probability that we would obtain this sample of data, given that each X is sampled from the Bernoulli distribution with parameter p. If this seems slightly backwards, it's because it is: in the next section, we will more rigorously explain where this notion comes from. But for now, continuing with this, we can next use the fact that the trials are independent:

$$\hat{p} = \operatorname*{arg\,max}_{p\in[0,1]} \mathrm{P}(X_1 = x_1 \mid p)\mathrm{P}(X_2 = x_2 \mid p)\ldots\mathrm{P}(X_n = x_n \mid p)$$

$$\hat{p} = \operatorname*{arg\,max}_{p\in[0,1]} \prod_{i=1}^{n} \mathrm{P}(X_i = x_i \mid p)$$

However, maximizing this is often difficult. In practice, it's usually easier to work with *log likelihood*, which simply means taking the logarithm of the expression inside the argmax. This is valid because $\log(x)$ is a monotonically increasing function, meaning the maximum of $\log(f(x))$ will occur at the same place as the maximum of $f(x)$.

$$\hat{p} = \underset{p \in [0,1]}{\arg\max} \; \log \prod_{i=1}^{n} \mathrm{P}(X_i = x_i \mid p)$$

$$\hat{p} = \underset{p \in [0,1]}{\arg\max} \sum_{i=1}^{n} \log \mathrm{P}(X_i = x_i \mid p)$$

Next, using the pdf for a Bernoulli distribution with parameter p, we can actually solve for $\hat{p}$. Noting that $\mathrm{P}(X_i = x_i) = p$ if a "HEADS" was flipped ($x_i =$ HEADS), and $\mathrm{P}(X_i = x_i) = 1 - p$ if a "TAILS" was flipped, we get:

$$\hat{p} = \underset{p \in [0,1]}{\arg\max}\, k \log p + (n-k)\log(1-p)$$

where k is the number of "HEADS" that were flipped. Next, to maximize this expression, we can set the derivative of the objective function with respect to p equal to 0, in order to find critical points.

$$\frac{d}{dp}\left(k \log p + (n-k)\log\left(1-p\right)\right) = 0$$

$$\frac{k}{p} - \frac{n-k}{1-p} = 0$$

$$\frac{1-p}{p} = \frac{n-k}{k}$$

$$\frac{1}{p} - 1 = \frac{n}{k} - 1$$

$$p = \frac{k}{n}$$

Some simple checking can be done to verify that $\frac{k}{n}$ is indeed a maximum and not a minimum, which gives us our final answer:

$$\hat{p} = \frac{k}{n}$$

Logically, this makes sense: if we flip 7 heads and 13 tails, the best guess for the bias of the coin is that it lands heads $\frac{7}{7+13}$ of the time. However, there's a slight problem here: we made the implicit assumption that all biases are equally likely. What if we want to model the fact that fair coins are much more common than biased coins? That's where maximum a posteriori (MAP) comes in.

Maximum a Posteriori

Maximum a Posteriori (MAP) is closely related to MLE, but with the addition of a *prior* over the distribution's parameters. That is, MAP allows us to take advantage of the knowledge that some parameters might be inherently more likely than others. For example, when we're given a coin, and we're trying to estimate the bias, it's much more likely that the coin is fair (or near fair) than it is that it has a significant bias. This is captured in the prior, so-called because it represents our "prior beliefs" before having looked at any data (in this case, the coin flips).

More rigorously, the difference between MAP and MLE is in the expression we're maximizing: in MAP we maximize the posterior probability, whereas in MLE we maximize the likelihood function. The difference is subtle; letting θ represent our distribution's parameters and X represent the sampled data, we have

$$\begin{aligned} \text{Posterior probability: } & \mathrm{P}(\theta \mid X) \\ \text{Likelihood function: } & \mathrm{P}(X \mid \theta) \end{aligned}$$

In fact, using Bayes' Theorem, we can show relatively easily that maximizing the posterior probability (MAP) is equivalent to the maximum likelihood estimate when the prior is *uniform*. That is, when all possible θ are equally likely.

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \mathrm{P}(\theta \mid X)$$

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \frac{\mathrm{P}(X \mid \theta)\ \mathrm{P}(\theta)}{\mathrm{P}(X)}$$

Since $\mathrm{P}(X)$ is constant (no θ dependence), we can simply remove it from the argmax:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \mathrm{P}(X \mid \theta)\ \mathrm{P}(\theta)$$

Likewise, if $\mathrm{P}(\theta)$ is uniform, then it also does not depend on θ:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \mathrm{P}(X \mid \theta)$$

$$\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{MLE}}$$

Thus, we have shown that MLE is simply a special case of MAP, when all parameters are equally likely. Tying this back into our coin bias example, the implicit assumption was that any bias of the coin was equally likely: $\mathrm{P}(p = 0.99) = \mathrm{P}(p = 0.5) = \mathrm{P}(p = x) \forall x \in [0, 1]$. A more realistic prior might be a Gaussian centered around $p = 0.5$. We leave calculating the MAP estimate for this case as an exercise for the reader, as we now focus on a more interesting application of MAP and MLE.

Mean Squared Error as MLE

Now, we will show that the use of a mean squared error (MSE) as a loss function in linear regression corresponds directly to the maximum likelihood estimate under Gaussian noise. That is, we make the assumption that our observations were generated from a linear model with normally-distributed noise:

$$y_i = \langle \theta, x_i \rangle + \epsilon,\ \epsilon \sim N(0, \sigma^2)$$

If we have reason to believe a linear model[1] could fit our data well, then this is a relatively natural assumption to make, since noise in real-world phenomena is often normally distributed with zero-mean. Our goal is then to estimate the parameter vector θ, given data $X = \{x_1, x_2, \ldots, x_n\}$ with observations $Y = \{y_1, y_2, \ldots, y_n\}$, where $\theta, x_i \in \mathbb{R}^m$, and $y_i \in \mathbb{R}$. We can do this with MLE as follows:

$$\begin{aligned}
\hat{\theta} &= \arg\max_{\theta}\ \log \mathrm{P}(Y \mid X, \theta) \\
&= \arg\max_{\theta} \sum_{i=1}^{n} \log \mathrm{P}(y_i \mid x_i, \theta) \\
&= \arg\max_{\theta} \sum_{i=1}^{n} \log \mathrm{P}(\epsilon = y_i - \langle \theta, x_i \rangle) \\
&= \arg\max_{\theta} \sum_{i=1}^{n} \log \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i - \langle \theta, x_i \rangle}{\sigma}\right)^2} \\
&= \arg\max_{\theta} \left(n \log \left(\frac{1}{\sigma\sqrt{2\pi}} \right) - \sum_{i=1}^{n} \frac{1}{2} \left(\frac{y_i - \langle \theta, x_i \rangle}{\sigma} \right)^2 \right) \\
&= \arg\max_{\theta} \left(-\frac{1}{2\sigma^2} \right) \sum_{i=1}^{n} (y_i - \langle \theta, x_i \rangle)^2 \\
&= \arg\min_{\theta} \sum_{i=1}^{n} (y_i - \langle \theta, x_i \rangle)^2
\end{aligned}$$

And we have arrived at the mean squared error loss! (Technically, we're missing the $\frac{1}{n}$, but for fixed n this is constant). Thus, we have shown that maximizing log likelihood under a Gaussian noise model directly corresponds to minimizing the mean squared error. Note that the variance of the noise, σ^2, does not directly appear in the result, so it doesn't matter when finding the maximum likelihood estimate.

[1] Note that linear models can be trivially extended to include higher order terms by computing and adding these terms as a new feature – this derivation holds in these cases as well.

Regularization as a Prior

Similar to how mean squared error can be viewed as a natural loss function arising from assuming Gaussian noise, it can also be shown that many forms of regularization can be seen as choosing a prior distribution over parameters, with MAP estimation. For example, L2 weight regularization comes from choosing a Gaussian prior for the weight vector.

If we let our prior be

$$\mathrm{P}(\theta) =$$

The rest is left as an exercise for the reader.