02-680 Module 18

Essentials of Mathematics and Statistics

Aidan Jan

November 25, 2025

Parameter Estimation and Maximum Likelihood Estimation

Consider a dataset $X_1, X_2, \ldots, X_n$ of independent and identically distributed random varialbes from the same unknown distribution. That is, for each X_i, the underlying distributions have the same μ and σ.

Let $\bar{X_n}$ be the average of the actual value of the observations:

$$\bar{X_n} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Note this is different from the expected value of the underlying distribution μ. Note also that $\bar{X_n}$ is also a random variable in an of itself.

Statistical Inference

We will use statistical inference or **learning** to try to make the models match the observations we've made about the world.

Example: Coin Flip Simulator

Consider a coin flip simulator which usually outputs "heads" or "tails", but once in every 1000 flips, it outputs a nonsense number. The simulator acts as a data generator, producing observations we can see and measure.

Model vs. Reality

On the other side, we have a set of models - mathematical representations of how we think the data might have been generated. Each model has its own set of parameters, e.g., the probability of heads in a biased coin.

What is Statistical Inference?

Statistical Inference is the task of:

- using the data (observations)
- to estimate or "learn" the parameters of the model
- so that the model better matches the observed data

It's about bridging the gap between theory (models) and reality (data).

Assume you have a set $D = X_1, X_2, \ldots, X_n$ of independent and identically distributed variables for some

unknown distribution. Let F be a statistical model defined as a set of probability distributions. We will focus on parametric models defined as

$$F\{f(x|\theta) \ : \ \theta \in \Theta\}$$

where $\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$ are unknown parameters $\theta \in \Theta$.

The task is to find a distribution $\hat{f} \in F$ that models the phenomenon well, which includes modeling the associated θ.

We want to do this in a way that has:

- the ability to generalize well
- the ability to incorporate prior knowledge and assumptions
- the ability to scale.

Data Likelihood

Bayes theorem (written slightly differently) states:

$$p(hypothesis|data) = \frac{p(data|hypothesis)p(hypothesis)}{p(data)}$$

If we can figure out all the values on the right, we can get the probaiblity of interest on the left (how good our model parameters are given the data).

Let's start by looking at the most complicated component of the right side: $p(data|hypothesis)$, or in our case $p(D|\theta)$. Remember that the elements of D are identically and independently distributed, thus we can rewrite it:

$$p(\mathcal{D}|\theta) = p(X_1, X_2, \ldots, X_n|\theta) = \prod_{i=1}^{n} p(X_i|\theta)$$

Log Likelihood

Taking products of probabilities of the values can get very small. Therefore, we often take the log likelihood of the function. This turns the product into a sum.

$$\log(p(X_1, X_2, \ldots, X_n|\theta)) = \sum_{i=1}^{n} \log(p(X_i|\theta))$$

Maximum Likelihood Estimation (MLE)

The maximum likelihood estimate (MLE) is a way to estimate the value of a parameter of interest θ. This is the most frequently used method for parameter estimation.

Thinking about how we **maximize a function** (or really find the parameter that maximizes the function), we need to take its derivative and check the 0 points.

$$\frac{\mathrm{d}}{\mathrm{d}\theta} p(\mathcal{D}|\theta) = 0$$

(but we also need to ensure its a max and not a min or saddle.) If there is no critical point, then we will use the maximum allowable parameter value.