Probability and Statistics – Summary

Han-Miru Kim

May 11, 2021

1 Likelihood Analysis

Say we are given data $x_1, \ldots, x_n$ that were generated by random variables $X_1, \ldots, X_n$ and we are interested in finding the correct model of their distributions to explain the data.

Since there are multiple possible distributions (usually infinitely many) that could have produced the data, it is impossible to tell how exactly the random variables are distributed.

Moreover, unless we are given further information we can't even say which model was "most likely" to have generated this data because that would require us to turn the set of all possible P-spaces into a P-space itself.

What we do to solve this problem is to only look at some some *candidate distributions* $\mu_{X_1}, \ldots, f_{X_N}$ associated to the random variables that could have generated the data and we pick the one that has the highest chance of them to produce the results $x_1, \ldots, x_n$.

We let Θ to be set of candidate distributions, which are indexed by by some variable θ.

1.1 Maximum Likelihood Method

In the discrete case, we define the **Likelihood function**

$$\mathcal{L} : \Theta \times \mathbb{R}^n \to \mathbb{R}, \quad \mathcal{L}(\theta, x_1, \ldots, x_n) := \mathbb{P}_\theta[X_1 = x_1, \ldots X_n = x_n]$$

which tells us the likelihood that the model with index θ realised the outcomes $x_1, \ldots, x_n$.

In the absolutely continous case, that would be

$$L(\theta, x_1, \ldots, x_n) = f_\theta(x_1, \ldots, x_n)$$

where f_θ is the collective density of $\boldsymbol{X}$.

A natural approach to guess the model θ is to take the model that maximizes $\mathcal{L}(\theta, X$