Topic 19: Maximum *a posteriori* Estimation

02-680: Essentials of Mathematics and Statistics

November 21, 2024

1 The Frequentist versus Bayesian Schools

Both schools of statistics start with probability. For Bayesian inference, we take H to be a hypothesis (parameters) and D some data. Different people will have different a priori beliefs — but we would still like to make useful inferences from data.

When $p(H)$ is known, there is no disagreement, we will all just follow Bayes' Rule as written.
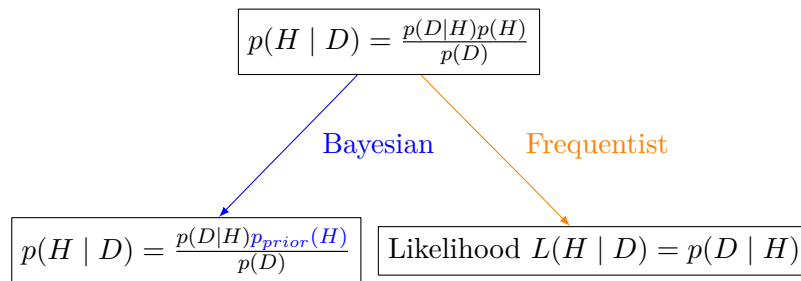
When the prior is *not* known, **Bayesian** logic requires us to develop a prior based on some information we have (intuition); while **Frequentist** logic uses only information in the provided data.

Bayesians and frequentists take fundamentally different approaches to this challenge. The reasons for this split are both practical (ease of implementation and computation) and philosophical (subjectivity versus objectivity and the nature of probability).

The main philosophical difference concerns the **meaning of probability**.

Bayesians	Frequentists
the idea that probability is an abstract concept that measures a state of knowledge or a degree of belief in a given proposition	the idea that probabilities represent long-term frequencies of repeatable random experiments
Subjective interpretation	Objective interpretation
you "believe" that you will get tails 50% of the time	the relative frequency of tails goes to $1/2$ as the number of flips goes to infinity
they consider a range of values each with its own probability of being true	

2   Bayesians' Approach to Parameter Estimation

Lets look at the coin flip example we had in the last topic: $\mathcal{D} = X_1, X_2, \cdots, X_n$ where $X_i \sim Bernouli(\alpha)$. We can further summarize $\mathcal{D}$ into c_H and c_T representing the counts of heads and tails respectively.

We saw last time that

$$\hat{\alpha}_{MLE} = \frac{c_H}{c_H + c_T}.$$

But this is assuming we know nothing about α ahead of time. What if we *believe* that its $50/50$, so we can add what are called **pseudocounts** to the input c_{H_0} and c_{T_0}. And thus compute

$$\hat{\alpha}_{MLE-PC} = \frac{c_H + c_{H_0}}{c_H + c_T + c_{H_0} + c_{T_0}}.$$

Example.   Lets assume we have some an experiment where we throw a coin 100 times, we want to know α, $c_H = 0$ and $c_T = 100$.

Vanilla MLE would say that

$$\hat{\alpha}_{MLE} = \frac{c_H}{c_H + c_T} = \frac{0}{0 + 100} = 0$$

But we have a small belief this is a fair coin, so lets assume we add the pseudocounts $c_{H_0} = c_{T_0} = 1$, in that case

$$\hat{\alpha}_{MLE-PC} = \frac{c_H + c_{H_0}}{c_H + c_T + c_{H_0} + c_{T_0}} = \frac{0 + 1}{0 + 100 + 1 + 1} = \frac{1}{102}.$$

If we're more confident in our prior and set $c_{H_0} = c_{T_0} = 100$ then

$$\hat{\alpha}_{MLE-PC} = \frac{c_H + c_{H_0}}{c_H + c_T + c_{H_0} + c_{T_0}} = \frac{0 + 100}{0 + 100 + 100 + 100} = \frac{100}{300} = \frac{1}{3}.$$

Pseudocounts are a way of exerting your *belief*

> **Larger pseudocounts** – represent a strong prior belief
>> Will have a greater effect on the posterior estimate

> **Small pseudocounts** – represent a weak prior belief
>> Will have a smaller effect on the posterior estimate

As the sample size goes to infinity, data will dominate the estimate.