Topic 19: Maximum ***a posteriori*** Estimation

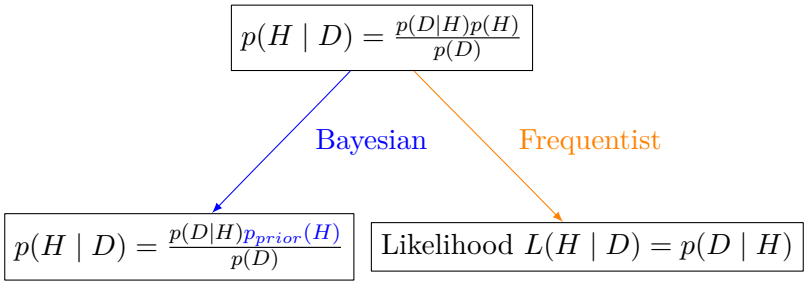
02-680: Essentials of Mathematics and Statistics

November 26, 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 they consider a range of values each with its own probability of being true	the relative frequency of tails goes to 1/2 as the number of flips goes to infinity

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*

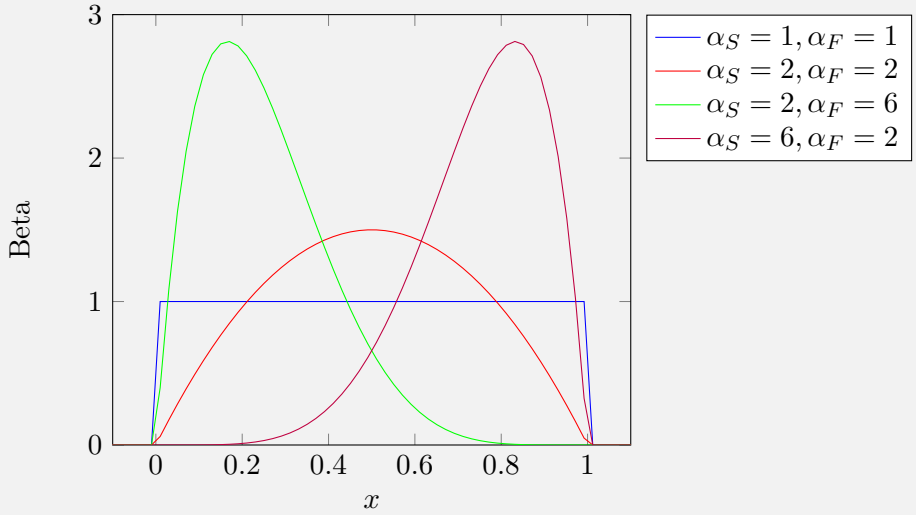
Larger pseudocounts – represent a strong prior belief
Will have a greater effect on the posterior estimate

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.

An additional distribution: Beta

The Beta distribution visually looks like an un-balanced normal.



$$f(x; \alpha_S, \alpha_F) = \frac{x^{(\alpha_S - 1)}(1-x)^{(\alpha_F)}}{B(\alpha_S, \alpha_F)}$$

where

$$B(\alpha_S, \alpha_F) = \frac{\Gamma(\alpha_S, \alpha_F)}{\Gamma(\alpha_S)\Gamma(\alpha_F)} \text{ and } \Gamma(a) = (a-1)!.$$

When $X \sim Beta(\alpha_S, \alpha_F)$, $\mathbb{E}[X] = \frac{\alpha_S}{\alpha_S + \alpha_F}$.

If we model the posterior as a ***Beta*** distribution on c_{H_0} and c_{T_0} (that is $p(\theta) \sim Beta(c_{H_0}, c_{T_0})$).

If we then want to find the *prior*, $p(\theta \mid \mathcal{D})$?

$$\begin{aligned} p(\theta \mid \mathcal{D}) &= \frac{p(\mathcal{D} \mid \theta)p(\theta)}{p(\mathcal{D})} \\ &\propto p(\mathcal{D} \mid \theta)p(\theta) \\ &= \binom{\alpha_H + \alpha_T}{\alpha_H} \cdot p^{\alpha_H} \cdot (1-p)^{\alpha_T} \cdot Beta(\alpha_{H_0}, \alpha_{T_0}) \\ &= \binom{\alpha_H + \alpha_T}{\alpha_H} \cdot p^{\alpha_H} \cdot (1-p)^{\alpha_T} \cdot \frac{p^{(\alpha_{H_0}-1)}(1-p)^{(\alpha_{T_0})}}{B(\alpha_{H_0}, \alpha_{T_0})} \\ &\sim Beta(\alpha_H + \alpha_{H_0}, \alpha_T + \alpha_{T_0}) \end{aligned}$$

3 Maximum *a Posteriori* (MAP) Estimation

As a reminder

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, p(\mathcal{D} \mid \theta).$$

On the other hand if we want to include information of our prior knowledge then we have MAP:

$$\hat{\theta}_{MAP} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, p(\theta \mid \mathcal{D}) = \underset{\theta \in \Theta}{\operatorname{argmax}}\, p(\mathcal{D} \mid \theta) \cdot p(\theta).$$

Note that in both cases we're making a ***point estimate*** of θ. We still don't have the whole picture, we're using data to estimate our model, but MAP is ***partially Bayesian***.

3.1 Example: *Known* Prior

There are three types of coins which have different probabilities of landing heads when tossed

Type A coins are fair and have probability 0.5 of heads

Type B coins are bent and have probability 0.6 of heads

Type C coins are bent and have probability 0.9 of heads

Suppose you have a drawer containing 5 coins: 5 of type A, 3 of type B, and 2 of type C. You reach into the drawer and pick a coin at random. Without showing you the coin is flipped **once** and get **tails**. What is the probability it is type A? Type B? Type C?

Hypothesis θ	Prior $p(\theta)$	Likelihood $p(\mathcal{D} \mid \theta)$	Bayes Numerator $p(\mathcal{D} \mid \theta) \cdot p(\theta)$	Posterior $p(\theta \mid \mathcal{D})$
A	0.5	0.5	0.4	0.641
B	0.3	0.6	0.4	0.308
C	0.2	0.9	0.1	0.051

Thus $\hat{\theta}_{MAP} = A$.

What if you then flip the same coin another 9 times, so including the first coin we have $\alpha_H = 6$ and $\alpha_T = 4$. Notice in the table below, the prior does not change.

Hypothesis θ	Prior $p(\theta)$	Likelihood $p(\mathcal{D} \mid \theta)$	Bayes Numerator $p(\mathcal{D} \mid \theta) \cdot p(\theta)$	Posterior $p(\theta \mid \mathcal{D})$
A	0.5	$0.97x10^{-3}$	$4.88x10^{-4}$	0.570
B	0.3	$1.19x10^{-3}$	$3.58x10^{-4}$	0.418
C	0.2	$0.05x10^{-3}$	$0.11x10^{-4}$	0.012

In this case, its still true that $\hat{\theta}_{MAP} = A$, but $\hat{\theta}_{MLE} = B$.

3.2 *Unknown* Prior

Assume we have a similar scenario but this time we don't know the prior, we follow a similar procedure to that for MLE: take the zero point of the log probability.

$$\frac{d}{d\theta} \ln p(\mathcal{D} \mid \theta) p(\theta) = 0$$

Conjugate Priors

Definition: If the prior and the posterior belong to the same parametric family, then the prior is said to be conjugate for the likelihood. Parameters of the prior distribution are often called "hyperparameters".
Example: α_S, α_F in $Beta(\alpha_S, \alpha_F)$.

How do you set the hyperparameters?

- Prior knowledge (from experienced domain-specific experts)
- If no prior knowledge, use "uninformative" prior (e.g., $Beta(1, 1)$)

Some common conjugate prior sets are in the table below:

Prior	Likelihood	Posterior
Beta	Bernoulli	Beta
Beta	Binomial	Beta
Gamma	Poisson	Gamma
Normal	Normal	Normal

Useful References

Wasserman. "All of Statistics: A Concise Course in Statistical Inference" §11 Degroot and Schervish. "Probability and Statistics" §§7.1-7.4