Bayesian framework

February 13, 2025

Thomas Bayes work:

"Divine Benevolence, or an Attempt to Prove That the Principal End of the Divine Providence and Government is the Happiness of His Creatures" (1731)
"An Introduction to the Doctrine of Fluxions, and a Defence of the Mathematicians Against the Objections of the Author of The Analyst" (published anonymously in 1736)
"An Essay towards solving a Problem in the Doctrine of Chances" which was read to the Royal Society in 1763 after Bayes' death

Trivia question:
In which of the three works, the famed Bayes theorem was formulated?

Unobservable variable

1. Suppose that there is a pair (X, Y) of random variables, of which only X is observable.
2. It is assumed that everything is known about the mathematical model for (X, Y), i.e. their joint distribution $F(x, y)$ is known.
3. The goal is to make inference about Y, i.e. to make interpolation (prediction) of Y.
4. The Bayesian set-up simply means to replace, in the notation, Y by θ, plus some shady terminology and interpretation to cover up that the model is completely specified - no parameters to adjust the model.
5. To cover up this lack of the parameterization, θ is used to pretend to be a parameter.
6. To summarize, the Bayesian approach is assuming a fully specified model, with no parameters, but more complex stochastic structure that makes possible to account for any behavior of observations by conditioning arguments and 'blaming' invisible observation Y, called θ, for the behavior.

Bayesian statistics? – confusing the audience

1. The Bayesian methods are part of the classical statistics, where one has some unobservable variables but essentially no parameters.
2. There is no such a thing as Bayesian statistics per se, from the methodological point of view.
3. Hidden variables plus actual parameters is even more general methodology and is a part of classical statistics.
4. Sometimes Bayesian methods are used to quantify the concept of 'belief'. This is especially common in social sciences, then there may be justified not to interpret the 'uncertainty' of θ as randomness.
5. This brings subjectivity to the statistical method but, in a sense, any choice of the model for the data is subjective.
6. The Bayesian approach through conditioning and the posterior distribution of unobserved variable give a tool to accept any subjective choice with no control of it being wrong or right.

Bayesian religion – 'parameters' as random variables

The main and conceptually only difference between classical statistical model (frequentist) and the Bayesian framework is that the 'parameter' θ in the Bayesian setup is a random variable with a certain distribution described by, say, $\pi(\theta)$ is attached to this variable. This one change is very consequential:

1. The random variable θ is not observable (hidden).
2. It is not possible to make classical inference about the original mechanism producing this variable and thus $\pi(\theta)$ is assumed to be known.
3. The distribution has to be assumed prior to data collection thus it is called the **prior distribution** or simply prior.
4. We can imagine that θ is drawn form $\pi(\theta)$, X from $p(x|\theta)$ is observed and we can infer about the unobserved value of θ given observed x, i.e. based on $p(\theta|x)$.

Prior, likelihood, posterior

Typically the Bayesian setup is presented in the following terms.

- Prior to collection of data, the parameter of interest $\theta \in \Theta$ is coming from some subjective and known **prior** given by the prior distribution (pdf, pmf,...)
$$\pi(\theta)$$
- The data x are observed and generated from the distribution given by the distribution (pdf,pmf,...)
$$p(x|\theta),$$
where θ is unknown to us. Since x is observed we treat the above as the **likelihood**
$$\theta \mapsto p(x|\theta).$$
- Our prior believe are now modified by the data and **posterior** distribution for θ is obtained using **the Bayes theorem**
$$\pi(\theta|x) \sim p(x|\theta)\pi(\theta),$$
where $\sim$ means the proportionality relation (up to a multiplicative constant, which can be obtained by integrating, summing in θ).

Bayesian statistics in a snapshot

In a surprisingly accurate summary of the Bayesian approach, one can write the Bayesian paradigm as

Posterior $\sim$ Likelihood * Prior

Example1: Hypergeometric model with an expert solicited prior

- Data (observation) k and the statistical model

$$P_\theta(X = k) = \frac{\binom{N\theta}{k}\binom{N-N\theta}{n-k}}{\binom{N}{n}},$$

where θ is the proportion of the defective items in the shipments so θ can take any of i/N, $i = 0, \ldots, N$.

- The prior on θ based on the customers who have provided accurate records of the number of defective items that they have found in their shipments

$$\pi(\theta = i/N) = \pi_i, \;\; i = 0, \ldots, N.$$

where π_i are known frequencies based on the prior customer input.

Example1 (continued): Posterior in the hypergeometric model

- Given the observed k faulty items the posterior corrects prior according to the Bayes rule (when zero is put in the below whenever the Newton symbol $\binom{v}{w}$ does not make sense, i.e. whenever it is not true that $v, w = 0, 1, 2, \ldots, v \geq w$)

$$\pi(\theta = i/N|k) \sim \binom{i}{k}\binom{N-i}{n-k}\pi_i, \ i = 0, \ldots, N.$$

Again we do not worry about the normalization, which can be also obtained through division by the sum of the terms in i.

Homework *Propose a hierarchical extension of this model by assuming that π_i's are obtained from a sample from a binomial distribution with the parameter $\theta = (m, p)$, where m is uniformly distributed over $50, 51, \ldots, 100$ and p is uniformly distributed on $[0, 1]$. Find the posterior distribution of θ.*

Example 2: Bernoulli model with a theoretical prior

- Parameter $\theta \in (0,1)$: the probability of a success in a single Bernoulli trial.
- Sample (observations) $X = (X_1, \ldots, X_n)$: a 0-1 sequence of indicators of the success in n-trials.
- Statistical model:

$$p(x|\theta) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} = \theta^k(1-\theta)^{n-k}, \; k = \sum_{i=1}^n x_i$$

- The Bayesian model: beta prior for θ (note that in the Bayesian setup we usually do not care about constants, as they can be always evaluated if needed, and more often than not they are actually not needed)

$$\pi(\theta) \sim \theta^{r-1}(1-\theta)^{s-1}, \quad s, r > 0$$

- Posterior:

$$\theta^k(1-\theta)^{n-k}\theta^{r-1}(1-\theta)^{s-1} = \theta^{k+r-1}(1-\theta)^{n-k+s-1},$$

Example 2 (continued): Conjugate prior

- We have seen the posterior:

$$\theta^{k+r-1}(1-\theta)^{n-k+s-1},$$

  which is again beta with parameters $k+r$ and $n-k+s$.
- **Prior**: beta with parameters $r > 0$ and $s > 0$.
- **Data**: the observation of k successes.
- **Posterior**: beta with empirically 'corrected' parameters $k+r > 0$ and $n-k+s > 0$.
- We observe a property of the posterior belonging to the same family of distributions as the one from which the prior is taken.
- We refer to this class as a **conjugate family** of priors for the given likelihood.

The Bayesian statistical inference

- In the classical statistics, we could consider different goals: point estimation, confidence intervals, testing hypotheses, or prediction.
- The analogous goals cannot be put forward in the Bayesian framework as they are all assuming a parameter which does ot exist in the Bayesian setup.
- The convenience of the Bayesian approach is that all information about what data tell about the unobserved variable is summarized conveniently in the posterior distribution.
- For example one can take the mean of the posterior and consider it as the Bayesian ‘estimator’ of the ‘parameter’, which is obviously a wrong interpretation.
- However, thinking in the traditional terms of statistical inference is misleading as there are no true values of the parameters.

The decision theory in the Bayesian setup

- How the posterior can be used for the purpose in hand depends on the priorities in the given context.
- This priorities can be summarized in terms of the so-called loss function that is attributed to a given decision.
- The **decision theory framework** in the Bayesian context uses the **Bayesian risk** which is the regular risk averaged over the prior distribution of the parameter

$$r(\delta) = E(R(\theta, \delta)) = E_\theta(E(l(\theta, \delta)|\theta)) = E(l(\theta, \delta(X))),$$

where the first expectation is only with respect the prior $\pi(\theta)$ and the last one is with respect the joint distribution of (X, θ) given by

$$f(x, \theta) = p(x|\theta)\pi(\theta).$$