CE 211
Mathematics for Engineers

Lecture 11

All About that Bayes

Previously on Mathematics for Engineers

There are multiple ways in which the term probability can be interpreted. For example,

- Suppose an unbiased coin is being tossed. The probability of observing heads is 0.5. One way to interpret 0.5 is that it is the frequency of occurrence of heads when we conduct a large number of trials.
- On the other hand, consider a statement "Based on historical records, there is an 80% chance that Subhas Chandra Bose died in a plane crash". The frequency argument won't work here since there is no repetition. In such instances, probability can be viewed as a subjective belief.

Previously on Mathematics for Engineers

Theorem (Bayes' Theorem)

Suppose $A_1, \ldots, A_n$ represents a partition of the sample space Ω and $\mathbb{P}(A_i) > 0 \, \forall \, i = 1, \ldots, n$. Then, for any event B with $\mathbb{P}(B) > 0$

$$\begin{aligned}\mathbb{P}(A_i|B) &= \frac{\mathbb{P}(A_i)\mathbb{P}(B|A_i)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(A_i)\mathbb{P}(B|A_i)}{\mathbb{P}(A_1)\mathbb{P}(B|A_1) + \ldots + \mathbb{P}(A_n)\mathbb{P}(B|A_n)}\end{aligned}$$

For two events A and A^c, Bayes' theorem can be rewritten as

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A)\mathbb{P}(B|A)}{\mathbb{P}(A)\mathbb{P}(B|A) + \mathbb{P}(A^c)\mathbb{P}(B|A^c)}$$

Previously on Mathematics for Engineers

Beta distributions are common in situations in which the realizations of the random variable falls in an interval.

$X \sim \text{Beta}(\alpha, \beta)$	
Parameters	$\alpha > 0, \beta > 0$
Support	$x \in (0, 1)$
PDF	$\begin{cases} \dfrac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$
CDF	No closed form
Expectation	$\dfrac{\alpha}{\alpha+\beta}$
Variance	$\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

where $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$

Previously on Mathematics for Engineers

The specific choice of the estimator $\hat{\Theta}$ for a parameter θ is governed by a few desirable properties.

Definition (Bias)

An estimator $\hat{\Theta}_n$ of θ is unbiased if the bias $B(\hat{\Theta}_n) = \mathbb{E}(\hat{\Theta}_n) - \theta$ is 0.

Definition (MSE)

Given an estimator $\hat{\Theta}_n$ of θ, the mean squared error of the estimator is defined as $MSE(\hat{\Theta}_n) = \mathbb{E}\Big((\hat{\Theta}_n - \theta)^2\Big)$

Definition (Consistency)

An estimator $\hat{\Theta}_n$ of θ, is said to be consistent if $\hat{\Theta}_n \xrightarrow{p} \theta$

A **consistent** estimator with **zero bias** and **lower MSE** is always preferred.

Previously on Mathematics for Engineers

Imagine the travel time on a street was lognormal and we are interested in its mean and variance. Suppose, we make the following measurements 5,7,3,4, and 7.

To estimate the mean of the random variable μ, we will define $\hat{\Theta}_n$ as

$$\hat{\Theta}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

$\hat{\Theta}_n$ is a random variable since it is a function of random variables. Every set of measurements of Xs we make will give us a realization of $\hat{\Theta}_n$.

Likewise, to estimate the variance σ^2, we define

$$\hat{\Theta}_n = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \frac{X_1 + X_2 + \ldots + X_n}{n} \right)^2$$

Previously on Mathematics for Engineers

For continuous random variables, we use the PDF function at each data point and multiply them to derive the likelihood function.

As an example, suppose we want to estimate the parameters μ and σ of a lognormal distribution. Assume that we observe realizations $x_1, x_2, \ldots, x_n$.

The likelihood objective can then be written as

$$\mathcal{L}(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{x_i \sigma \sqrt{2\pi}} \exp\left(-\frac{(\ln x_i - \mu)^2}{2\sigma^2}\right)$$

The log-likelihood takes the form

$$\mathcal{LL}(\mu, \sigma) = \sum_{i=1}^{n} \ln\left(\frac{1}{x_i \sigma \sqrt{2\pi}} \exp\left(-\frac{(\ln x_i - \mu)^2}{2\sigma^2}\right)\right)$$

The objective is to thus maximize $\mathcal{LL}(\mu, \sigma)$ by changing both μ and σ subject to $\sigma > 0$.

Lecture Outline

1. The Problem
2. Bayesian Inference
3. Point Estimation
4. A Solution

Lecture Outline

The Problem

The Problem

Simon Newcomb, an astronomer, found in 1881 that the first few pages in logarithm tables in a library were more worn out than the others. Why?

Lecture Outline

Bayesian Inference

Bayesian Inference

Introduction

The estimation methods that we saw earlier are called **frequentist** or **classical** approaches.

In these methods, we assume that the true parameter value (θ) is deterministic but are unknown and we use realizations of estimators $\hat{\Theta}$ to guess their values.

Bayesian methods on the other hand treat the parameters as a random variable Θ. They begin by assuming a **prior** probability distribution $p_{\Theta}(\theta)$ and update it as more information X becomes available to construct a **posterior distribution** $p_{\Theta|X}(\theta|x)$.

Bayesian Inference

Bayes' Theorem Revisited

Assume that we observe data/measurements $X = (X_1, X_2, \ldots, X_n)$, which can be viewed as a realization of a random vector.

The posterior distribution can thus be written as

$$p_{\Theta|X}(\theta|x) = \frac{p_\Theta(\theta)p_{X|\Theta}(x|\theta)}{\sum_{\theta'} p_\Theta(\theta')p_{X|\Theta}(x|\theta')}$$

Does $p_{X|\Theta}(x|\theta)$ in the numerator resemble something that we saw before? An alternate way of writing the posterior is

$$p_{\Theta|X}(\theta|x) \propto \mathcal{L}(\theta)p_\Theta(\theta)$$

which is the product of the prior and the likelihood function.

Bayesian Inference

Summary

A Bayesian is one who, vaguely expecting a horse, and catching a glimpse of a donkey, strongly believes he has seen a mule (Source: Internet)

Bayesian Inference

Example 1: Coin Tosses

Imagine that we wish to estimate the probability of heads of a biased coin. Suppose we toss it n times and observe k heads. How does the frequentist solve this problem?

The Bayesian would assume that the head probability is a random variable Θ. A potential prior could be $\Theta \sim U(0,1)$. Hence, $f_\Theta(\theta) = 1$ for $\theta \in (0,1)$.

The observations could be viewed as realizations of n Bernoulli trials $X_1, \ldots, X_n \sim \mathit{Bernoulli}(\theta)$. The posterior PDF can thus be written as

$$\begin{aligned} f_{\Theta|X}(\theta|k) &\propto f_\Theta(\theta) p_{X|\Theta}(k|\theta) \\ &\propto \theta^k (1-\theta)^{n-k} \end{aligned}$$

Comparing it with the PDF of Beta distribution $\frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$, we can conclude that

$$\Theta|X \sim \text{Beta}(k+1, n-k+1)$$

Bayesian Inference

Example 2: Gaussian Priors

Suppose we observe $X = (X_1, \ldots, X_n)$, where each $X_i \sim \mathcal{N}(\theta, \sigma^2)$ and σ^2 is known and wish to estimate the mean. How does the frequentist solve this problem?

Suppose the Bayesian assumes that the mean is a random variable Θ with a prior $\sim \mathcal{N}(a, \sigma^2)$, where a is a known constant. The posterior can thus be written as

$$\begin{aligned} f_{\Theta|X}(\theta|x) &\propto f_{\Theta}(\theta) f_{X|\Theta}(x|\theta) \\ &\propto \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\theta-a}{\sigma})^2} \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x_1-\theta}{\sigma})^2} \end{aligned}$$

Simplifying this expression gives $\Theta|X \sim \mathcal{N}\left(\frac{a+x_1+\ldots+x_n}{n+1}, \frac{\sigma^2}{n+1}\right)$
This is an example where the posterior also belongs to the same family as that of the prior, and these are called **conjugate** prior and posterior.

Lecture Outline

Point Estimation

Point Estimation

Introduction

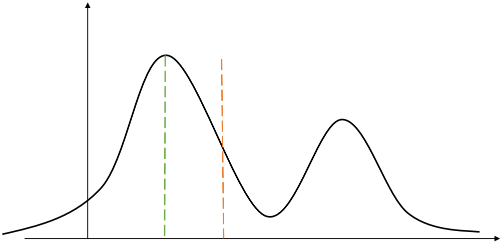
Unlike in the frequentist method, Bayesian inference does not provide a single parameter estimate $\hat{\theta}$ from the data.

While the earlier approach helps us derive the posterior distributions, we might be interested in a single point estimate which gives the best guess of Θ.

Point Estimation

MAP Rule

One popular point estimate is the Maximum a Posteriori (MAP) probability rule.



According to this method, we simply select the realization $\hat{\theta}$ at which the posterior distribution is maximized (see green point). Mathematically,

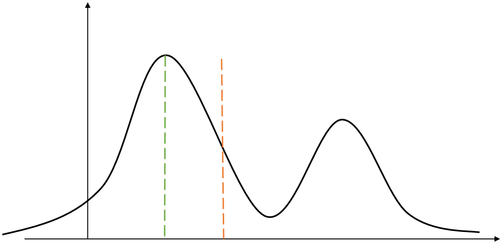
$$\hat{\theta} = \arg\max_{\theta} f_{\Theta|X}(\theta|x)$$

What is the MAP estimator for the Gaussian posterior in the earlier example?

Point Estimation

Conditional Expectation

Another option is to use the expected value of the posterior random variable (see orange point). This value is also called the conditional expectation estimator.



Mathematically, $\hat{\theta} = \mathbb{E}(\Theta | X = x)$. What is the conditional expectation estimator for the Gaussian posterior in the earlier example?

Lecture Outline

A Solution

A Solution

Benford's Law

Most numbers found in practice tend to start with the digit 1. Frank Benford, a physicist discovered this pattern in 1938 on several datasets such as physical constants and surface areas of rivers.

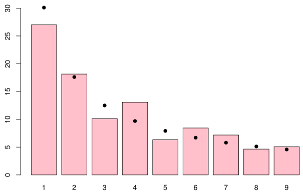
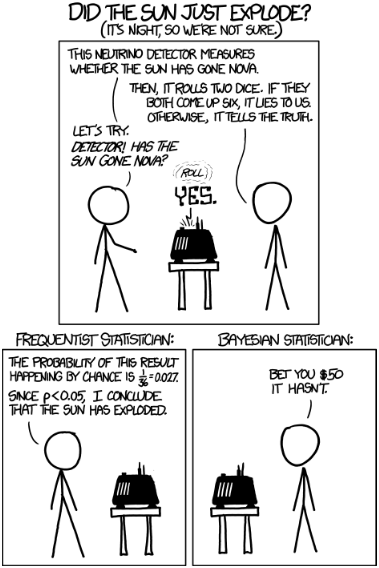


Figure: Distribution of first digits in the population of countries

Since then, this pattern has been found in many other areas such as bank account balances and genome data. It has also been used a screening for fraud detection. Mathematical explanations based on entropy have been proposed for this phenomenon.

Your Moment of Zen


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY. DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36}=0.027$. SINCE $p<0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.