Probability for Machine Learning

Sadeep Jayasumana

June 28, 2024

Random Variables

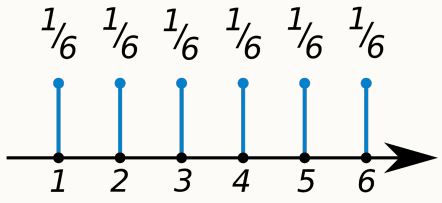
- Discrete Random Variables
  - ✔ X = Number you get when you throw a dice
  - ✔ $X \in \{1, 2, 3, 4, 5, 6\}$

- Continuous Random Variables
  - ✔ X = Height in cm, of a randomly picked student
  - ✔ $X \in [0, 250]$

Probability Density Functions

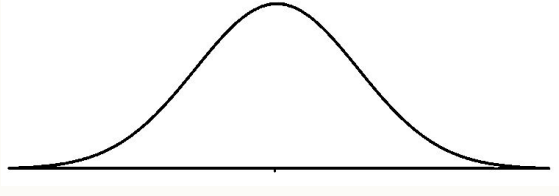
- Discrete Random Variables – Probability Mass Function
  - ✔ X = Number you get when you throw a dice
  - ✔ X ∈ {1, 2, 3, 4, 5, 6}

$$f_X(x) = P(X = x)$$



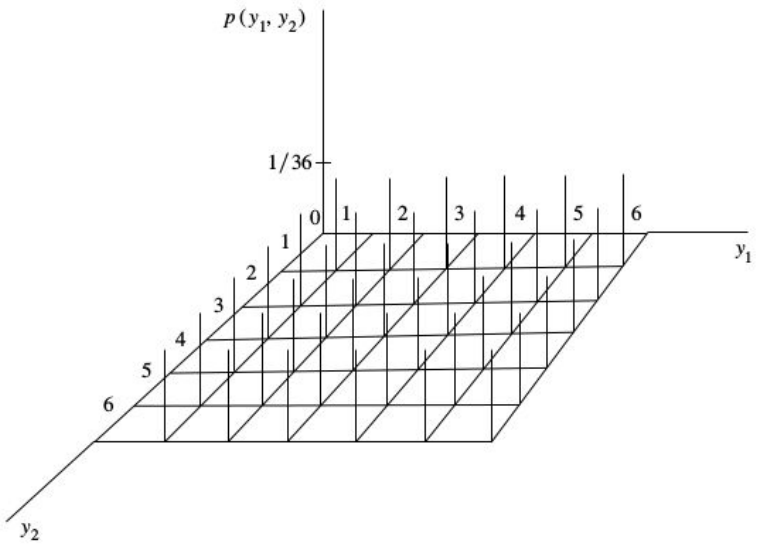
- Continuous Random Variables – Probability Density Function
  - ✔ X = Height in cm, of a randomly picked student
  - ✔ X ∈ [0, 250]

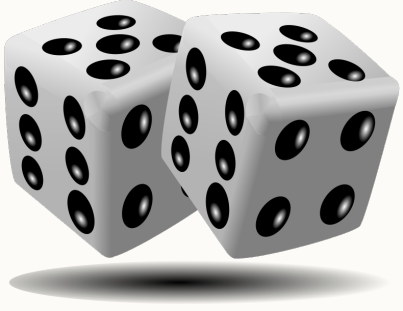
$$f_X(x) = P(x \leq X \leq x + \delta x)$$



Multivariate Probability Distributions

FIGURE **5.1**
Bivariate probability function; y_1 = number of dots on die 1, y_2 = number of dots on die 2



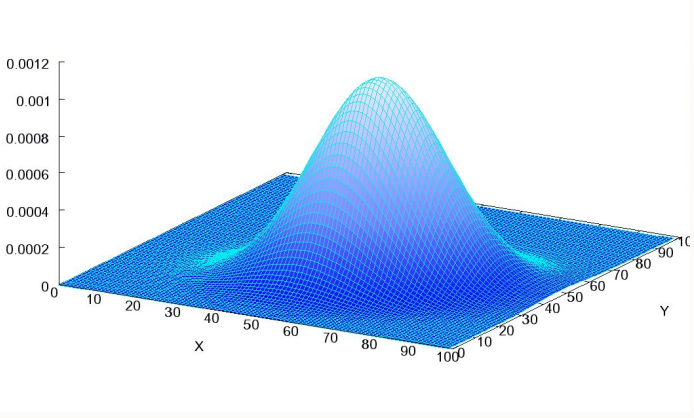


Let Y_1 and Y_2 be discrete random variables. The *joint* (or bivariate) *probability function* for Y_1 and Y_2 is given by

$$p(y_1, y_2) = P(Y_1 = y_1, Y_2 = y_2), \qquad -\infty < y_1 < \infty, -\infty < y_2 < \infty.$$

Discrete R.V.s

Multivariate Probability Distributions



$$F(u, v) = P(U \leq u, V \leq v)$$

$$F(u, v) = \int_{-\infty}^{u} \int_{-\infty}^{v} f(t_1, t_2)\, dt_2\, dt_1$$

Continuous R.V.s

Conditional Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

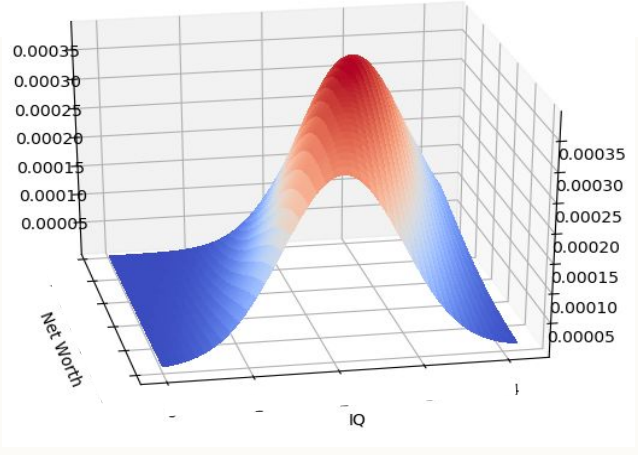
- Probability of an event A happening ***given*** that the even B already happened.

Conditional Distributions

U = IQ of the person

V = Net worth of the person

$f_{UV}(u,v)$ = Joint PDF

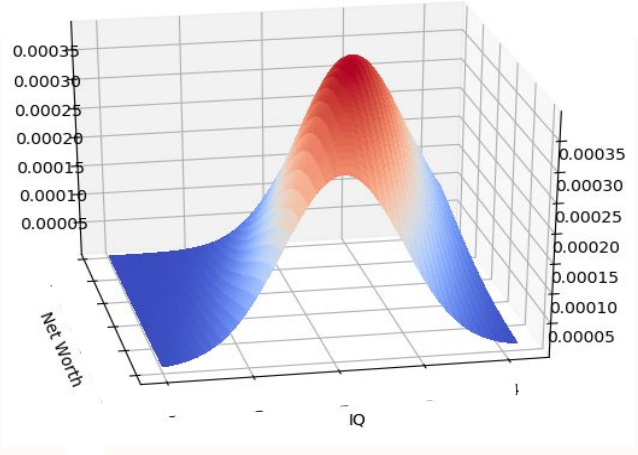


Conditional Distributions

U = IQ of the person

V = Net worth of the person

$f_{UV}(u,v)$ = Joint PDF

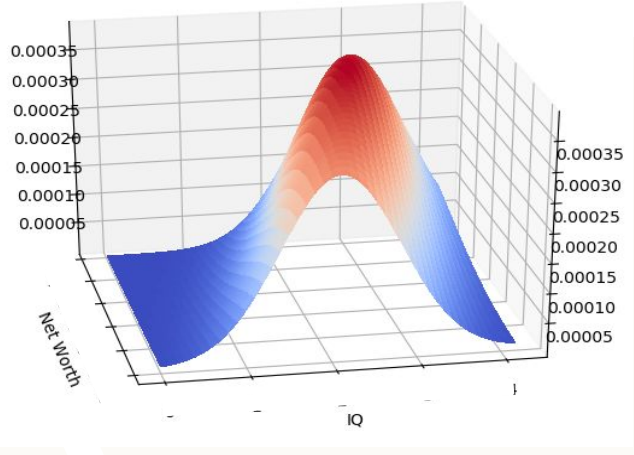


$$f(u|v) = \frac{f_{UV}(u,v)}{f_V(v)}$$

f(u|V=10 million) : Distribution of IQ *given that* net worth is 10 million

f(v|U = 150) : Distribution of net worth *given* that IQ is 150 units

Marginal Distributions



U = IQ of the person

V = Net worth of the person

$f_{UV}(u,v)$ = Joint PDF

$f_U(u)$ = Marginal PDF for U

$$f_U(u) = \int_{-\infty}^{\infty} f_{UV}(u,v)\,dv$$

= Distribution of IQ regardless of net worth

= Distribution of IQ among all people (not filtered by net worth)

= Integrate out V, then you have only U

Important Laws of Probability

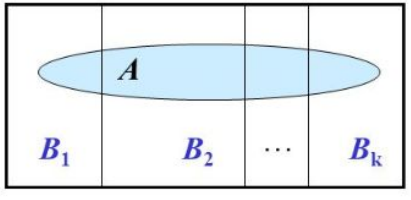
Let $X \in \{x_1, x_2, \ldots, x_i, \ldots, x_m\}$ and $Y \in \{y_1, y_2, \ldots, y_j, \ldots, y_n\}$ be random variables. Then,

Sum rule:

$$P(X = x_i) = \sum_j P(X = x_i, Y = y_j)$$

Product rule:

$$P(X = x_i, Y = y_j) = P(Y = y_j | X = x_i) P(X = x_i)$$



S

If $P(B_i) \neq 0$, then

$$P(A \cap B_i) = P(A|B_i)P(B_i)$$

Hence, for the partition $\{B_1, B_2, \ldots, B_k\}$ of the sample space S, we have $P(A) = P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_k)$

or equivalently,

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_k)P(B_k).$$

The law of total probability

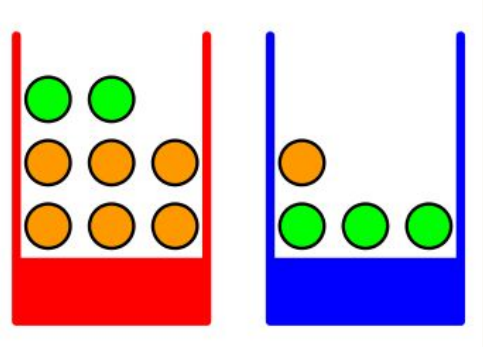
Bayes Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- This lets you invert conditional probabilities.

Example

Behind a curtain there are two buckets containing fruits: a red one and a blue one. The red one contains 2 apples and 6 oranges, the blue one contains 3 apples and one orange. A magician throws a fair dice and if the value is less than or equal to 4, he randomly picks a fruit from the red bucket, otherwise from the blue bucket.



The magician goes behind the curtain and picks a fruit using the above method. You're shown the fruit, but covered with a cloth.

i.) What is the probability that it was picked from the red bucket?

ii.) What is the probability that it is an orange?

iii.) The magician uncovers the fruit and it turns out to be an orange.

What is the probability that it was picked from the red bucket?

Source: Pattern Recognition and Machine Learning, Christopher M. Bishop

Example

A randomly selected lady undergoes a test for detecting a rare disease that exists in only 1/10,000th of the population. The sensitivity of the test (probability of the test saying "positive" when one actually has the disease) is 0.95. The false alarm rate of the test is 0.1 (probability of the test saying "positive" when one actually does NOT have the disease). The test outputs "positive". What's the probability that the lady has the disease?

Mean and Variance

$$\mu = E[X] = \int_{-\infty}^{\infty} x f(x)\, dx$$

$$\sigma^2 = V[X] = E[(X - \mu)^2]$$

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx$$

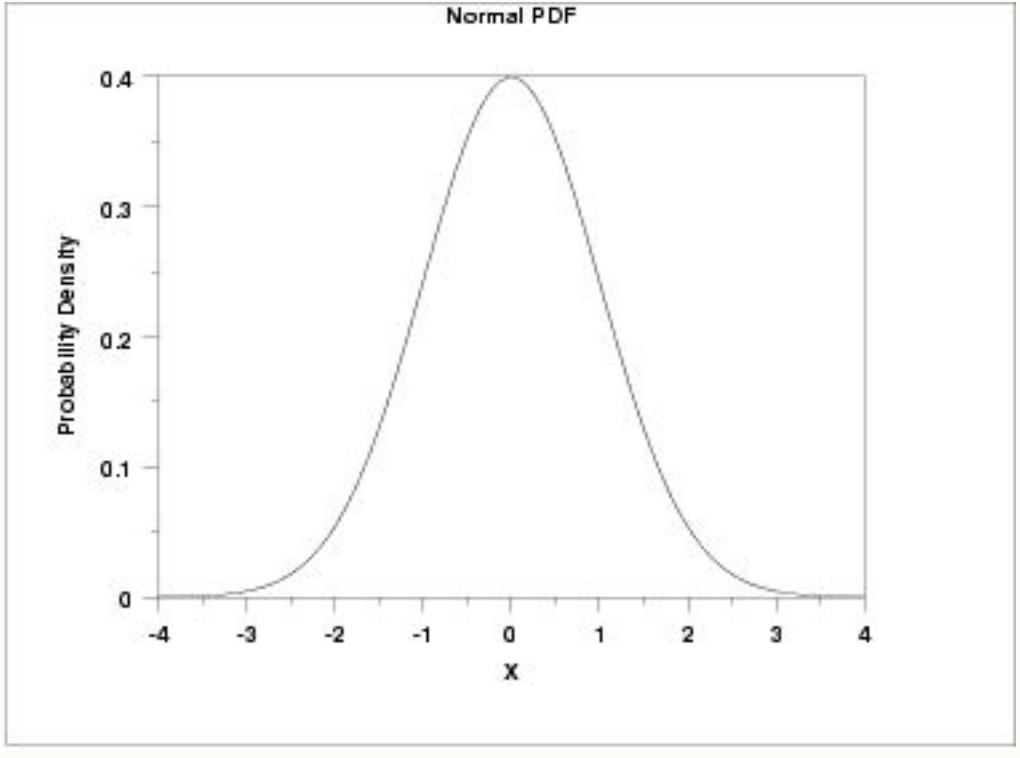
Exercise

- Prove the following where X is a generic r.v. and a, b are real-valued constants.

$$E[aX + b] = aE[X] + b$$

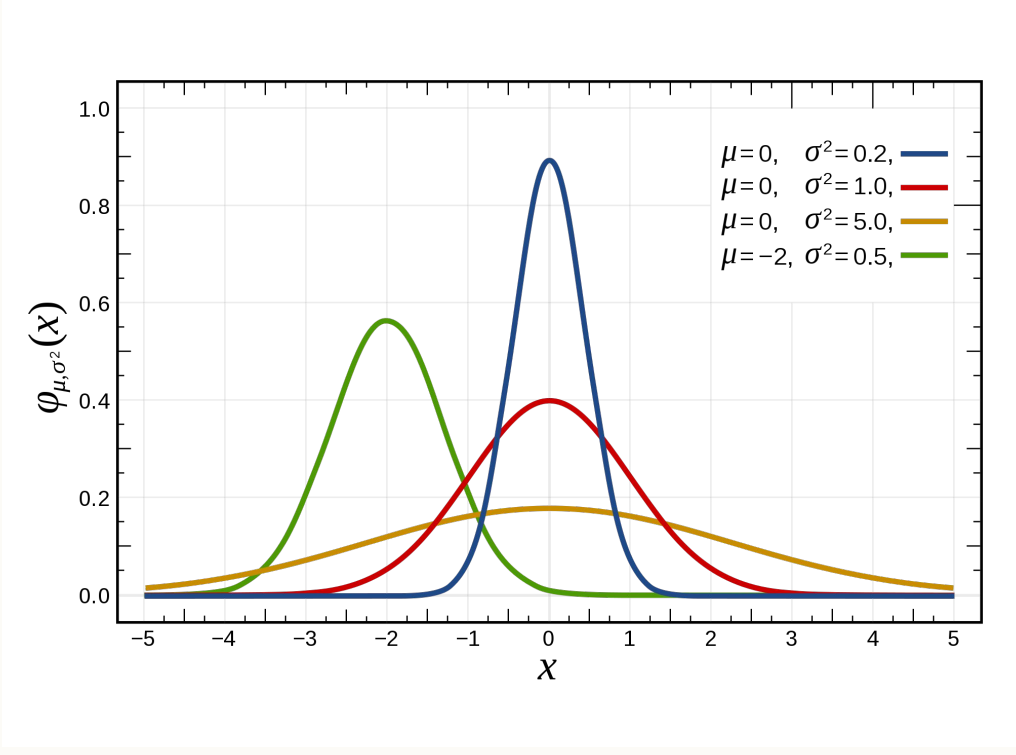
$$V[aX + b] = a^2 V[X]$$

Normal Distribution



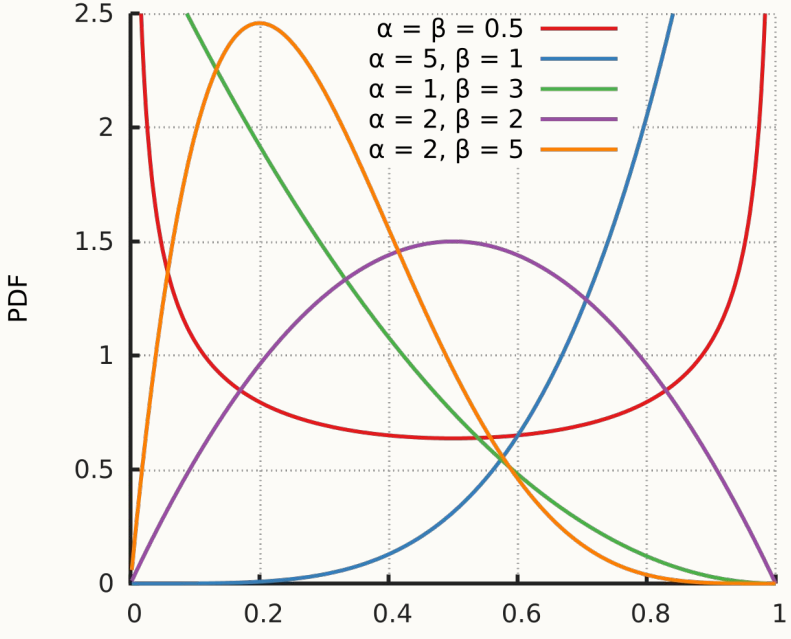
$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Normal Distribution



$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Beta Distribution



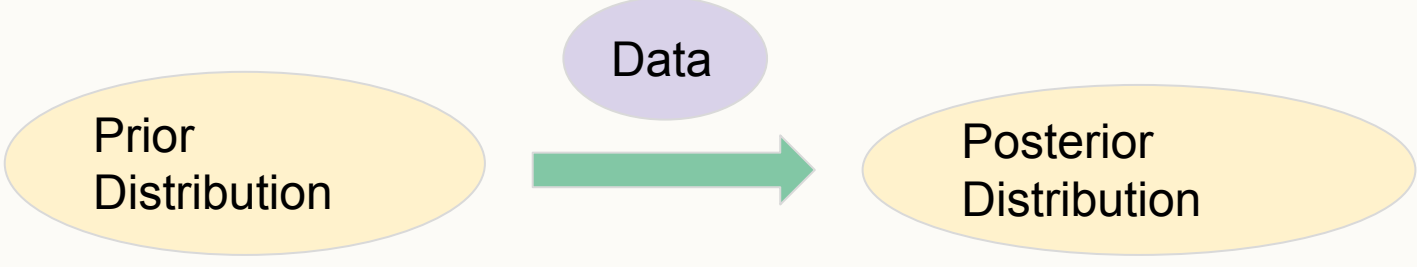
$$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$

where $\mathrm{B}(\alpha,\beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

$$\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\,dx$$

Bayesian Inference

- Frequentist view
  - Parameters are fixed. We can estimate them using data
  - There is only one true real value for a given parameter, but there are can be many datasets
  - Therefore, our estimation of the parameter has some uncertainty
- Bayesian view
  - Parameters are random variables that have distributions
  - There is only one dataset, we need to improve our knowledge about the parameters' distributions using that



Bayesian Inference

- Likelihood function: P(Data | θ) viewed as a function of θ
- Maximum likelihood estimate (MLE): θ value that maximises the likelihood function
- Prior distribution: Prior (before seeing data) belief about the distribution of θ
- Posterior distribution: Posterior (after seeing data) belief about the distribution of θ
- Maximum a posteriori probability (MAP): θ value that maximises posterior probability

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

Example

You find a coin that looks normal. But you want to find out whether it has a bias. You flip the coin 3 times and it lands head in all three times.

Let P(coin lands on head) = h

i.) Write down the likelihood function L(h) = P(Data | h).

ii.) What is the MLE estimate for h ?

iii.) Think of h as a random variable H. Before flipping the coin and observing the results, propose a suitable probability distribution for H (prior distribution).

iv.) Given the flipping results, derive the posterior distribution of H.

v.) What is the MAP estimate for h ?

vi.) You're going to flip the coin two more times, what's the probability of it landing head in both those times? Think in both the frequentist perspective and the Bayesian perspective.

Example

h = Head probability = Pr(Side = H)

Bayesian Inference
Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available.

Frequentist View
Parameters are Fixed: In the frequentist view, parameters (e.g., the mean or variance of a population) are considered fixed but unknown values. These parameters are not random; they are fixed properties of the population.
Estimate Using Data: We use data from samples to estimate these fixed parameters. For example, we might use the sample mean to estimate the population mean.
One True Value: There is only one true value for a given parameter. However, since we can have many different datasets (samples), our estimates will vary depending on the sample.
Uncertainty in Estimation: Because our sample might not perfectly represent the population, our estimate of the parameter will have some uncertainty. This uncertainty is often quantified using confidence intervals.
Bayesian View
Parameters are Random Variables: In the Bayesian view, parameters are not fixed. Instead, they are treated as random variables with their own probability distributions.
Distributions: These distributions (called prior distributions) represent our initial beliefs about the parameters before seeing the data.
One Dataset: There is only one dataset, and we use this data to update our knowledge about the parameters' distributions.
Updating Knowledge: Bayes' theorem allows us to update our prior distributions with the observed data to obtain posterior distributions. These posterior distributions reflect our updated beliefs about the parameters after considering the data.
Comparing the Two Views
Frequentist:
Parameters are fixed.
There is one true value for each parameter.
Estimates have uncertainty due to sampling variability.
Bayesian:
Parameters are random variables with distributions.
We use the data to update these distributions.
Posterior distributions represent our updated knowledge about the parameters.

Example
Consider estimating the mean of a population:

Frequentist Approach:

Assume the mean is a fixed value.
Collect a sample and calculate the sample mean.
Construct a confidence interval around the sample mean to express the uncertainty.
Bayesian Approach:

Assume the mean has a prior distribution (e.g., normal distribution with some mean and variance).
Collect a sample and update the prior distribution using Bayes' theorem to get the posterior distribution.
The posterior distribution provides a full description of our uncertainty about the mean after considering the data.
Conclusion
Both approaches aim to make inferences about unknown parameters based on data, but they differ fundamentally in their treatment of parameters and the interpretation of probability. Frequentists view probability as the long-run frequency of events, while Bayesians view it as a measure of belief or certainty about events.

Example

h = Head probability = Pr(Side = H)

Likelihood function $L(h) = Pr(Data \mid h) = h^3$

MLE estimate $h = \text{argmax}_h\ L(h) = 1.0$

Likelihood Function
P(Data |): This is the likelihood function, which represents the probability of the observed data given a specific value of the parameter \theta.
Function of : When viewed as a function of \theta, it tells us how likely different values of \theta are, given the observed data.

Maximum Likelihood Estimate (MLE)
Definition: The MLE is the value of \theta that maximizes the likelihood function.
Purpose: It provides the parameter estimate that makes the observed data most probable.

Prior Distribution
Definition: The prior distribution represents our belief about the possible values of \theta before observing any data.
Purpose: It incorporates any existing knowledge or assumptions about \theta.

Posterior Distribution
Definition: The posterior distribution represents our updated belief about \theta after observing the data.
Formula:
\text{Posterior} \propto \text{Likelihood} \times \text{Prior}PosteriorLikelihood×Prior
This means the posterior distribution is proportional to the product of the likelihood function and the prior distribution.

Maximum a Posteriori Probability (MAP)
Definition: The MAP estimate is the value of \theta that maximizes the posterior distribution.
Purpose: It provides the parameter estimate that is most probable given both the observed data and the prior information.

Example

h = Head probability = Pr(Side = H)

Posterior ∝ Likelihood x Prior

Conjugate priors make things easier

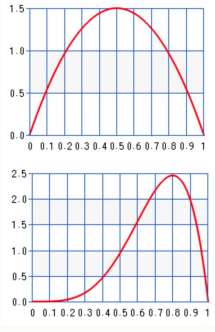
Likelihood function $L(h) = Pr(Data \mid h) = h^3$

MLE estimate $h = \text{argmax}_h\ L(h) = 1.0$

Prior distribution: Beta(2, 2)

Posterior distribution: Beta(5, 2)

MAP estimate $h = \text{argmax}_h\ Pr(h \mid Data) = 0.8$



Example

h = Head probability = Pr(Side = H)

Posterior ∝ Likelihood x Prior

Conjugate priors make things easier

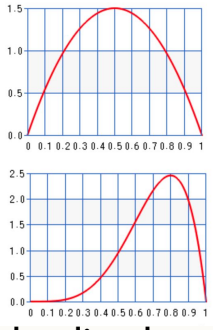
Likelihood function $L(h) = Pr(Data \mid h) = h^3$

MLE estimate $h = \text{argmax}_h\ L(h) = 1.0$

Prior distribution: Beta(2, 2)

Posterior distribution: Beta(5, 2)

MAP estimate $h = \text{argmax}_h\ Pr(h \mid Data) = 0.8$



You're going to flip the coin two more times, what's the probability of it landing head in both those times? Think in both the frequentist perspective and the Bayesian perspective.

Exercise

Prove the following where X and Y are random variables.

$E[X + Y] = E[X] + E[Y]$

$V[X + Y] = V[X] + V[Y] + 2\,COV[X, Y]$

What happens when X and Y are independent?

Detailed Explanation

Likelihood Function
The likelihood function P(\text{Data} | \theta)P(Data) measures the probability of observing the given data for different values of \theta. For example, if you have data points from a normal distribution, the likelihood function would tell you how likely it is to observe those data points for different values of the mean \mu and variance \sigma^2

Maximum Likelihood Estimate (MLE)
The MLE is the value of \theta that maximizes the likelihood function. This means we are finding the parameter value that makes the observed data most likely. Mathematically, it is:

Prior Distribution
Before seeing any data, we might have some beliefs about the parameter \theta. This belief is encoded in the prior distribution P(\theta)P(). For example, if you believe \theta is likely to be around a certain value, you would choose a prior distribution centered around that value.

Posterior Distribution
After observing the data, we update our prior beliefs using Bayes' theorem to get the posterior distribution. The posterior distribution combines the prior distribution and the likelihood of the observed data. Mathematically, it is:
P(\theta | \text{Data}) \propto P(\text{Data} | \theta) \times P(\theta)P(Data)P(Data)×P()
This means the posterior distribution is proportional to the likelihood function multiplied by the prior distribution

Maximum a Posteriori Probability (MAP)
The MAP estimate is the value of \theta that maximizes the posterior distribution. This is similar to the MLE but takes into account the prior distribution as well. Mathematically, it is:

Thank You!