Probability and Statistics

Probability is a measure of how likely an event is to occur.

→ Complement

Given probability of an event happening is P, then the probability of it not happening is 1−P

→ Sum of probabilities (Disjoint events)

$P(A \cup B) = P(A) + P(B)$ given that $A \cap B = 0$ (Disjoint sets)

→ Sum of probabilities (Joint events)

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

→ Independence

When the occurrence of one event does not affect the occurrence of another event

$P(A \cap B) = P(A) \cdot P(B)$

→ Birthday problem

Let's find the probability that every one has a different birthday

1st person → $P = \frac{365}{365}$

2nd person → $P = \frac{364}{365}$

3rd person → $P = \frac{363}{365}$

4th person → $P = \frac{362}{365}$

$P(\text{no match}) = \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{362}{365} \cdot \frac{362}{365} \cdots \frac{365 - n + 1}{365}$

$$= \frac{365!}{(365-n)! \cdot 365^n}$$

$$P(\text{at least one match}) = 1 - \frac{365!}{(365-n)! \, 365^n} \geq 0.5$$

$$\Rightarrow \prod_{k=0}^{n-1} \frac{365-k}{365} \leq 0.5$$

For small x, $\ln(1-x) \approx -x$, so:

$$\ln P(\text{no match}) = \sum_{k=0}^{n-1} \ln\left(1 - \frac{k}{365}\right) \approx -\sum_{k=0}^{n-1} \frac{k}{365} = -\frac{n(n-1)}{2 \cdot 365}.$$

Set $P(\text{no match}) = 0.5$

$$\Rightarrow -\frac{n(n-1)}{730} = \ln(0.5) \approx -0.693$$

$$n(n-1) \approx 506.$$

$$\Rightarrow n \approx \frac{1+\sqrt{1+4 \cdot 506}}{2} \approx \frac{1+\sqrt{2025}}{2} = 23$$

So for $n = 23$, we will have at least two people with the same birthday.

→ Conditional probability

Calculating the probability of an event happening given that another event has already happened.

$P(A \cap B) = P(A) \cdot P(B|A)$ → When independent $P(B|A) = P(B)$

→ Bayes Theorem

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

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

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{and} \quad P(B|A) = \frac{P(A \cap B)}{P(A)} \Rightarrow P(A \cap B) = P(B|A)P(A)$$

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

$P(A|B)$ — the posterior. What you want to know: probability of A after seeing evidence B.

$P(B|A)$ — the likenhood. How probable the evidence is, assuming A is true.

$P(A)$ — the prior. What you believed about A before seeing any evidence.

$P(B)$ — the marginal likelihood or evidence. Total probability of seeing B across all posibilities.

$P(B)$ can be expanded using law of total probability:

$$P(B) = P(B|A) \cdot P(A) + P(B|A') \cdot P(A')$$

$$\text{and so } P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A')P(A')}$$

Classical example: A disease affects 1% of the population. A test is 99% accurate (99% true positive rate, 99% true negative rate). You test positive. What is the probability you actually have the disease?

* A = has disease, $P(A) = 0.01$
* B = test positive.
* $P(B|A) = 0.99$ (true positive)
* $P(B|A') = 1 - P(B'|A') = 1 - 0.99 = 0.01$ (false positive)

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(A')P(B|A')} = \frac{0.01 \times 0.99}{0.01 \times 0.99 + 0.99 \times 0.01} = 0.5$$

even with a 99% accurate test, a positive result only means a 50% chance of having the disease. That's because the disease is rare, false positive

chance of having the disease. [illegible]
from the 99% healthy people swamp the true positive from the 100 sick.

→ Monty Hall problem

E_1 = the car is behind door 1, E_2 = door 2, E_3 = door 3., $E_i = 0$, $i = 1, 2, 3$

These are mutually exclusive events, in order words, you cannot simultaneously have a car in two doors, because of the rule of the game.

$\Rightarrow P(E_1 \cap E_2) = P(E_2 \cap E_3) = P(E_1 \cap E_3) = 0$, $P(E_i \cap E_j) = 0$ for $i \neq j$

Another fact is that the car is behind one of the three doors, so

$$P(E_1 \cup E_2 \cup E_3) = 1$$

$P(E_1) = 1/3$, $P(E_1') = \frac{2}{3}$. (choose door 1).

then we have $E_n' = E_2 \cup E_3$ and so $P(E_2 \cup E_3) = 2/3$

If host opens door 3 revealing a goat and asks if we want to switch doors. If we don't switch, the probability of winning remains because this is the initial choice. If we do switch, then, we can notice that the host gave an additional information. They showed us door 3 does not have a car which means $P(E_3) = 0$

But $P(E_2 \cup E_3) = P(E_2) + P(E_3) - P(E_2 \cap E_3) = 2/3 \Rightarrow P(E_2) = 2/3$

→ Generalized Monty Hall problem

* There are n doors and you must choose one door
* Host opens k doors and revealing goats
* You may or may not change your previously chosen door

E_i = the car is behind door i, $i = 1, 2, \cdots, n$.

E_i's are independent from each other

Since host never opens the same door player ensue and also never opens the

winning door, there is an upper bound for k which is $n-2$

Two facts can be assumed:

* Player chooses door 1
* Host opens door $2, \ldots, k+1$

$$P(E_1) = \frac{1}{n}, \quad P(E') = 1 - \frac{1}{n} = \frac{n-1}{n}$$

$$E_1' = E_2 \cup E_3 \cup \cdots \cup E_n$$

$$P\left(\bigcup_{i=2}^{n} E_i\right) = \frac{n-1}{n}$$

If player switches to a random available door, then they must choose one of the following $k+2, k+3, \ldots, n-1, n$, $(n-k-1$ elements$)$

$$P(\text{win} \mid \text{switch}) = \frac{n-1}{n} \cdot \frac{1}{n-k-1}$$

$$P(\text{win} \mid \text{switch}) = \frac{n-1}{n} \cdot \frac{1}{n-k-1} = \frac{1}{n} \cdot \frac{n-1}{n-k-1} \geq \frac{1}{n} = P(E_1) = P(\text{win} \mid \text{not switch})$$

Always good to switch.

→ Naive Bayes Model

Assume the appearance of the words $w_1, w_2, \ldots, w_n$ are independent

$$P(A \mid w_1, \ldots, w_n) = \frac{P(A)\, P(w_1, \ldots, w_n \mid A)}{P(A) P(w_1, \ldots, w_n \mid A) + P(B) P(w_1, \ldots, w_n \mid B)}$$

$$= \frac{P(A)\, P(w_1 \mid A) \cdots P(w_n \mid A)}{P(A) P(w_1 \mid A) \cdots P(w_n \mid A) + P(B) \cdot P(w_1 \mid B) \cdots P(w_n \mid B)}$$

→ Probability in machine learning

Bayes theorem can be used in classifier for ML

Sentiment analysis too

→ Random variables

Random variables allow us to model the who experiment at once

Types of random variables

→ Discrete (can take countable number of intervals)
→ Continous (takes the whole value in an interval, not countable)

→ Discrete probability distribution

* $P_X(x) \geq 0$
* $\sum_x P_X(x) = 1$

↑ probability mass function (PMF)

→ Binomial Distribution

$$P_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n$$

→ Bernoulli Distribution

→ Continous probability distribution

Discrete:
→ Sum of heights equal to 1

Continous:
→ Area under the curve equals 1

→ Probability Density function (PDF) $f_X(k)$ are only defined for continous variables.

$P[a < X < b] =$ area under $f_X(x)$

$f_X(x)$ needs to satisfy:

* It is defined for all numbers
* $f_X(x) \geq 0$
* Area under $f_X(x) = 1$

→ Cummulative Distribution Function

$CDF(x) = F_X(x) = P(X \leq x)$ ← defined for every real number.

Properties:

* $0 \leq F_X(x) \leq 1$
* Left endpoint is 0
* Right endpoint is 1
* Never decreases

→ Uniform distribution

A continous random variable can be modeled with a uniform distribution if all possible values lie in an interval and have the same frequency of occurence.

Parameters:

* a — beginning of the interval
* b — end of the interval

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & x \notin (a;b) \end{cases}$$

$$F_X(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x < b \\ 1 & b \leq x \end{cases}$$

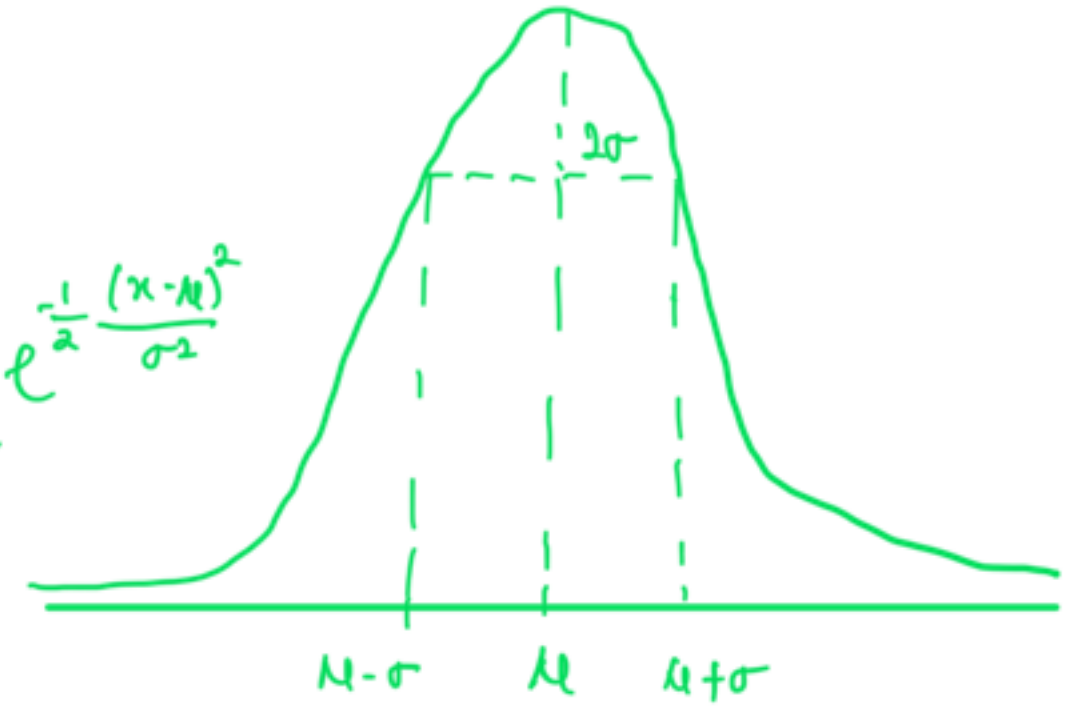
→ Normal distribution (Gaussian distribution)

μ = mean

σ = standard deviation

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

$$X \sim N(\mu, \sigma^2)$$



When $\mu = 0$ and $\sigma = 1$, we have standard normal distribution

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}x^2}, \quad X \sim N(0, 1^2)$$

We can standardize any normal distribution also

If $\mu = 2$ and $\sigma = 2.5$, we let $Z = \frac{X-2}{2.5}$

→ Sampling from a distribution

Step 1: generate random numbers

Step2: find out which interval the number belongs to

Step 3: Assign an outcome based on the interval.

→ Naive Bayes

The algorithm makes a naive assumption that each feature is independent of the others. It is important to note that the Naive Bayes is a Supervised algorithm, meaning it requires data that's already labeled to function effectively.

→ Naive Bayes for Spam Detection

probability of interest for a given email is denoted as: p(spam|email)

The higher this probability, the more likely the email is to be classified as spam. Bayes' theorem is used in the calculation in the following way:

$$P(\text{email}|\text{spam})\,P(\text{spam})$$

$$P(\text{spam}|\text{email}) = \frac{P(\text{email}|\text{spam})\,P(\text{spam})}{P(\text{email})}$$

$P(\text{spam})$: probability of a randomly selected email being spam, equivalent to the proportion of spam emails in the dataset.

$P(\text{email}|\text{spam})$: Probability of a specific email occuring given that it is known to be spam.

$P(\text{email})$: Overall probability of the email occuring

An interesting early shortcut we can take is just ignore the premail) term. The goal of the calculation will be to compare the probability of an email is spam to the probability it is ham.

$$P(\text{spam}|\text{email}) = \frac{P(\text{email}|\text{spam})\,P(\text{spam})}{P(\text{email})}$$

$$P(\text{ham}|\text{email}) = \frac{P(\text{email}|\text{ham})\,P(\text{ham})}{P(\text{email})}$$

Since $P(\text{email}) > 0$ and it appears in both expression, comparing the two probabilities only require comparing the numerators and ignoring the denominators

Representing an email as $\text{email} = \{\text{word}_1, \text{word}_2, \ldots, \text{word}_n\}$?

$$P(\text{email}|\text{spam}) = P(\text{word}_1|\text{spam})\,P(\text{word}_2|\text{spam}) \cdots P(\text{word}_n|\text{spam})$$

$$P(\text{word}_k|\text{spam}) = \frac{\#\text{spam emails with word}_k}{\#\text{spam emails}}$$

$$P(\text{spam}) = \frac{\#\text{spam emails}}{\#\text{total emails}}, \quad P(\text{ham}) = \frac{\#\text{ham emails}}{\#\text{total emails}}$$

→ Describing probability distributions

→ Expected Value

Expected value $E(X)$ also known as mean.

If X is a discrete random variable with PMF $P_X(x) = P(X = x)$

$$E(X) = \sum_x x P_X(x)$$

If X is a continous random variable with PDF $f_X(x) = P(X = x)$

$$E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx$$

$$E(g(X)) = \sum_x x\, g(x)$$

$$E(aX + b) = a E(X) + b$$

$$E(X_1 + X_2) = E(X_1) + E(X_2)$$

In general

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$$

→ Variance

Variance is used to measure spread.

$$\text{Variance} = E[(X - E[X])^2]$$

$$\text{Var}(X) = E((X - E(X))^2) = E(X^2) - E(X)^2$$

$$\begin{aligned}\text{Var}(X) &= E(X^2 - 2XE(X) + E(X)^2) \\ &= E(X^2) - 2E(X)E(X) + E(X)^2 \\ &= E(X^2) - 2E(X)^2 + E(X)^2 \\ \text{Var}(X) &= E(X^2) - E(X)^2\end{aligned}$$

$$\begin{aligned}\text{Var}(aX + b) &= E(a^2X^2 + 2abX + b^2) - E(aX + b)^2 \\ &= a^2E(X^2) + 2abE(X) + b^2 - (aE(X) + b)(aE(X) + b) \\ &= a^2E(X^2) + 2abE(X) + b^2 - a^2E(X)^2 - 2abE(X) - b^2\end{aligned}$$

$$Var(aX+b) = a^2(E(X^2) - E(X)^2) = a^2 Var(X)$$

$$\Rightarrow Var(aX+b) = a^2 Var(X)$$

→ Standard deviation

Variance has one drawback which is its unit, it has the square of the base unit.

$$std(X) = \sqrt{var(X)}$$

→ Sum of Gaussians

$$R = T + L$$

$$\mu_R = E(R) = E(T+L) = E(T) + E(L) = \mu_T + \mu_L$$

$$\sigma_R^2 = Var\,R = Var(T+L) = Var(T) + Var(L) = \sigma_T^2 + \sigma_L^2$$

In general: $W = aX + bY$

Independent $\begin{cases} X \sim N(\mu_X, \sigma_X^2) \\ Y \sim N(\mu_Y, \sigma_Y^2) \end{cases}$

$$\rightarrow W \sim N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2)$$

→ Skewness and Kurtosis

$E(X)$ → First moment

$E(X^2)$ → Second moment.

$E(X^3)$ → Third moment

...

$E(X^k)$ → kth moment.

$$Skewness = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$ → which direction there is more value in.

$$Kurtosis = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$$ → Thickness of the tails of a distribution

→ Joint Probability Distribution

$$P_{XY}(x,y) = P(X=x, Y=y)$$

When the variables are independent $P_{XY}(x,y) = P(x) \cdot P(y)$

→ Marginal and Conditional Distribution

Marginal Distribution is a distribution of one variable while ignoring the other. (sum the joint probability distribution over all value of the other variable).

$$P_Y(y_j) = \sum_i P_{XY}(x_i, y_j)$$

Conditional Distribution is when we want to observe a variable given that the other is known.

$$P_{Y|X=x}(y) = \frac{P_{XY}(x,y)}{P_X(x)} \rightarrow \text{Discrete}$$

$$f_{Y|X=x}(y) = \frac{f_{XY}(x,y)}{f_X(x)} \rightarrow \text{Continuous}$$

marginal distribution of X

→ Covariance (Relationship between two random variables)

$$Cov(X,Y) = \frac{1}{n}\sum_i (x_i - \mu_x)(y_i - \mu_y)$$

→ Coveriance of a Probability Distribution

$$Cov(X,Y) = \sum P_{XY}(x_i,y_i)(x_i - \mu_x)(y_i - \mu_y)$$

$$Cov(X,Y) = E(XY) - E(X)E(Y)$$

$\text{Cov}(X,Y)$

→ Covariance Matrix

$$C = \begin{bmatrix} \text{Var}(x) & \text{Cov}(X,Y) \\ \text{Cov}(X,Y) & \text{Var}(y) \end{bmatrix}$$

→ Correlation Coefficient (lies between -1 and 1)

$$\text{Correlation coefficient} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)} \cdot \sqrt{\text{Var}(Y)}}$$

→ Multivariate Gaussian Distribution

for a single variable, X

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

Parameters: $\mu \Rightarrow$ center of the bell
$\sigma \Rightarrow$ spread of the bell

What if we have $X_1, X_2, \ldots, X_n$?

for W and H:

If W and H are independent.

$$f_{HW}(h,w) = f_H(H) f_W(W) = \frac{1}{2\pi\sigma_H\sigma_W} e^{\frac{-1}{2}\left(\frac{(h-\mu_H)^2}{\sigma_H^2} + \frac{(w-\mu_W)^2}{\sigma_W^2}\right)} \rightarrow \left\| \begin{bmatrix} \frac{h-\mu_H}{\sigma_H} \\ \frac{w-\mu_W}{\sigma_W} \end{bmatrix} \right\|_2^2$$

$$= \begin{bmatrix} \frac{h-\mu_H}{\sigma_H} & \frac{w-\sigma_W}{\sigma_W} \end{bmatrix} \begin{bmatrix} \frac{h-\mu_H}{\sigma_H} \\ \frac{w-\mu_W}{\sigma_W} \end{bmatrix}$$

$$= \left(\begin{bmatrix} h & w \end{bmatrix} - \begin{bmatrix} \mu_H & \mu_W \end{bmatrix} \right) \begin{pmatrix} \frac{1}{\sigma_H^2} & 0 \\ 0 & \frac{1}{\sigma_W^2} \end{pmatrix} \left(\begin{bmatrix} h \\ w \end{bmatrix} - \begin{bmatrix} \mu_H \\ \mu_W \end{bmatrix} \right)$$

$$= \left(\begin{bmatrix} h \\ w \end{bmatrix} - \begin{bmatrix} \mu_H \\ \mu_W \end{bmatrix} \right)^T \begin{bmatrix} \sigma_H^2 & 0 \\ 0 & \sigma_W^2 \end{bmatrix}^{-1} \left(\begin{bmatrix} h \\ w \end{bmatrix} - \begin{bmatrix} \mu_H \\ \mu_W \end{bmatrix} \right)$$

μ — Σ Covariance matrix — μ

$$f_{HW}(h,w) = \frac{1}{2\pi\sigma_H\sigma_W} \exp\left(-\frac{1}{2}\left(\begin{bmatrix}h\\w\end{bmatrix}-\begin{bmatrix}\mu_H\\\mu_W\end{bmatrix}\right)^T\begin{bmatrix}\sigma_H^2 & 0\\0 & \sigma_W^2\end{bmatrix}^{-1}\left(\begin{bmatrix}h\\w\end{bmatrix}-\begin{bmatrix}\mu_H\\\mu_W\end{bmatrix}\right)\right)$$

$$= \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix}h\\w\end{bmatrix}-\mu\right)^T\Sigma^{-1/2}\left(\begin{bmatrix}h\\w\end{bmatrix}-\mu\right)\right)$$

When W and H are dependent:

still the same formula but now $\Sigma = \begin{bmatrix}\sigma_H^2 & Cov(W,H)\\Cov(H,W) & \sigma_W^2\end{bmatrix}$

$$f_X(x_1,x_2,\dots,x_n) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{\frac{-1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

where $x = [x_1\ x_2\ \dots\ x_n]^T$ and $\mu = [\mu_1\ \mu_2\ \dots\ \mu_n]^T$

and $\Sigma = \begin{bmatrix}\sigma_{x_1}^2 & Cov(x_1,x_2) & \dots & Cov(x_1,x_n)\\Cov(x_2,x_1) & \sigma_{x_2}^2 & \dots & Cov(x_2,x_n)\\\vdots & & \ddots & \vdots\\Cov(x_n,x_1) & Cov(x_n,x_2) & \dots & \sigma_{x_n}^2\end{bmatrix}$

→ Sampling and Point Estimation

→ Population and Sample

A population is the group items we want to study. A sample is a smaller subset that we actually observe or measure. In machine learning, we often use samples to train models and make predictions.

→ Population and Sample in ML

* Every dataset we work with in ML is a sample not the population

→ N

Population size: N

Sample size: n

The greater the sample size, the better the estimate of the population we will get

$$\text{Population proportion } p = \frac{\text{number of items with a given characteristics } (x)}{\text{population } (N)}$$

Sample proportion $\hat{p}$ = estimate of the population proportion

$$\text{Var}(X) = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

let $Y = (X-\mu)^2$

$$E(Y) = \mu_Y = \frac{1}{N}\sum_{i=1}^{N} y_i \longrightarrow \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Population mean of Y — Sample mean of Y

$$\text{So } \widehat{\text{Var}(X)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad ???$$

($\bar{x}$: sample mean)

If we want sample variance to be unbiased, we have

$$\hat{\text{var}}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x - \bar{x})^2$$

$\longrightarrow$ Law of large numbers.

As the sample size increases, the average of the sample tends to get closer to the average of the entire population.

n: number of samples

X_i: is the i-th sample from the population

Each X_i are independent and identically distributed (i.i.d)

$$\text{as } n \to \infty \quad \frac{1}{n}\sum_{i=1}^{n} X_i \to E[X] = \mu_X$$

Under the following conditions:

under the following

* Sample is drawn randomly
* Sample size must be sufficiently large
* Individual observations in the sample must be independent of each other

→ Central Limit Theorem

If you take independent, identically distributed random variables $X_1, X_2, \ldots, X_n$ from any distribution with finite mean μ and finite variance σ^2, then as n grows, the distribution of their sample mean approaches a normal distribution:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{d} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Equivalently, the standardized sum converges to a standard normal:

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

→ Point Estimation

Point estimation is using sample data to compute a single value (point) that serves as the best guess for unknown population parameters.

→ Setup:

* We have a population with some unknown parameter θ – could be a mean, variance, proportion, etc.
* We take a sample from that population
* We compute a statistics from the sample, called a point estimator
* $\hat{\theta}$ is our estimator of θ

→ Properties of a good estimator:

* Unbiased: On average, $\hat{\theta}$ equals the true θ. $E(\hat{\theta}) = \theta$
* Consistent: As $n \to \infty$, $\hat{\theta}$ converges to θ (this is what Law of Large Numbers gives us for the sample mean)
* Efficient: It has the smallest variance among unbiased estimators

* Efficient: ...

* Sufficient: Uses all relevant information from the sample

→ Common methods

* Method of moments: Match sample moments to population moments
* Maximum likelihood estimation (MLE): Pick the θ that makes observed data most probable
* Method of least squares: Pick θ that minimizes squared errors (used in regression)
* Bayesian estimation: Use the posterior mean/median/mode

MLE is the workhorse, it is what we are implicitly doing when we let most models in statistics and machine learning.

→ Maximum Likelihood Estimation (MLE)

Picking the most probable from the observed data.

Likelihood is the probability of seeing a particular data based on the model

→ Bernoulli example:

$$L(p; 8H) = p^8(1-p)^2$$

$$\log L = 8\log p + 2\log(1-p)$$

$$\frac{1}{L}\frac{dL}{dp} = \frac{8}{p} - \frac{2}{1-p} \Rightarrow \frac{8}{p} - \frac{2}{1-p} = 0 \Rightarrow 8 - 10p = 0 \Rightarrow p = \frac{8}{10}$$

n coins and k heads where each coin is a bernouni variable.

$$L(p; x) = P_p(X = x) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x} = p^{\sum_{i=1}^{n} x_i}(1-p)^{(n-\sum_{i=1}^{n} x_i)}$$

$$\ell(p|x) = \log(L(p; x)) = \sum_{i=1}^{n} x_i \log p + \left(n - \sum_{i=1}^{n} x_i\right)\log(1-p) \quad / \text{ let } S = \sum_{i=1}^{n} x_i$$

$$\frac{d\ell}{dp} = \frac{S}{p} - \frac{n-S}{1-p} = 0 \Rightarrow S(1-p) - p(n-S) = 0 \Rightarrow S - pn = 0$$

$$\Rightarrow p = \frac{S}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

→ Gaussian example

$S_1, \ldots, S_n$ are random variables and they are the numbers -1 and 1.

Say we have some observations

These observations were sampled from some distribution and the question is what distribution could they have been sampled from?

> The best distribution is the one where the mean of the distribution is the mean of the sample
>
> not only mean-, other measures also like variance

Suppose we have n samples $X = (x_1, x_2, \dots, x_n)$ from a Gaussian distribution with mean μ and variance σ^2. This means that $x_i \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$

Suppose we want MLE for μ and σ, we first define the likelihood.

$$L(\mu, \sigma; x) = \prod_{i=1}^{n} f_{x_i}(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}}$$

$$= \frac{1}{(\sqrt{2\pi})^n \sigma^n} e^{\frac{-1}{2}\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}}$$

$$\ell(\mu, \sigma) = \log(L(\mu, \sigma; x)) = \frac{-n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2}\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}$$

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\mu\right) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\mu\right)$$

$$\frac{\partial \ell}{\partial \sigma} = \frac{-n}{\sigma} + \frac{1}{\sigma^3}\left(\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

Setting both derivatives to 0

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\mu\right) = 0$$

First we know that $\sigma > 0$, the only optimum is that $\sum_{i=1}^{n} x_i - n\mu = 0$

MLE for μ: $\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$

Next for σ:

$$\frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\left(\sum_{i=1}(x_i-\mu)^2\right) = 0$$

also $\sigma > 0$ and $\mu = \bar{x}$

$$-n + \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right) = 0$$

MLE for σ: $\hat{\sigma} = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$

$$\frac{\partial^2 \ell}{\partial \mu^2} = -\frac{n}{\sigma^2}$$

$$\frac{\partial^2 \ell}{\partial \mu \partial \sigma} = \frac{\partial^2 \ell}{\partial \sigma \partial \mu} = \frac{-2}{\sigma^3}\left(\sum_{i=1}^{n} x_i - n\mu\right)$$

$$\frac{\partial^2 \ell}{\partial \sigma^2} = \frac{2n}{\sigma^2} - \frac{3}{\sigma^4}\left(\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

Hessian Matrix $H(x,y) = \begin{vmatrix} \frac{\partial^2 \ell}{\partial \mu^2} & \frac{\partial^2 \ell}{\partial \mu \partial \sigma} \\ \frac{\partial^2 \ell}{\partial \sigma \partial \mu} & \frac{\partial^2 \ell}{\partial \sigma^2} \end{vmatrix}$

at $\mu = \bar{x}$ and $\sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$

$$\frac{\partial^2 \ell}{\partial \mu^2} = \frac{-n^2}{\sum(x-\bar{x})^2}$$

$$\left(\frac{\partial^2 \ell}{\partial \mu \partial \sigma}\right)^2 = \frac{-2n^3}{\sum(x-\bar{x})}\left(n\bar{x} - n\bar{x}\right) = 0$$

$$\frac{\partial^2 \ell}{\partial \sigma^2} = \frac{2n^2}{\sum(x-\bar{x})^2} - \frac{3n^2}{\left(\sum(x-\bar{x})^2\right)^2}\cdot\left(\sum(x-\bar{x})^2\right) = \frac{-n^2}{\sum(x-\bar{x})^2}$$

$$D = \frac{n^4}{\sum(x-\bar{x})^2} > 0$$

and since $\frac{\partial^2 \ell}{\partial \mu^2} < 0$, we have a maximum value

→ MLE Linear regression

Least square errors using Gaussian distribution.

→ Regularization

Model : $y = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$

Log-loss : ll

L2 regularization error : $a_n^2 + a_{n-1}^2 + \cdots + a_1^2$

Regularizatum parameter : λ

Regularized error : $ll + \lambda(a_n^2 + a_{n-1}^2 + \cdots + a_1^2)$

→ Bayesian Statistics

Frequentists	Bayesians
Probabilities represents long term frequency of events	Probabilities represent the degree of belief (or certainty)
Concept of likelihood	Concept of prior
Goal : Find the model that most likely generated the observed data	Goal : Update prior belief based on observations

→ Maximum a Posteriori (MAP)

→ Updating priors.

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

$P(B|A)$ → likelihood of evidence B appearing, given A happened

$P(A)$ → prior

$P(B)$ → probability of evidence B in any circumstances

→ $P(B|A)P(A) + P(B|A')P(A')$

$P(A|B)$ → posterior

Belief that A will happen after considering B

$$P_{Y|X=x}(y) = \frac{P_{X|Y=y}(x)\,P_Y(y)}{P_X(x)}$$

→ discrete

$$f_{Y|X=x}(y) = \frac{f_{X|Y=y}(x)\, f_Y(y)}{f_X(x)} \rightarrow \text{continous}$$

Summary:

* Bayesians update prior beliefs
* MAP with uninformative priors is just MLE
* With enough data, MLE and MAP estimates usually converge.
* Good for instances when you have limited data or strong prior beliefs.
* Wrong priors, wrong conclusions

→ MAP vs MLE

MLE is a method of estimating the parameters of a statistical model by finding the parameter values that maximize the likelihood of observing given data.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(\text{data}|\theta)$$

MAP is a method of estimating an unknown parameter θ by finding the value that maximizes the posterior distribution - the probability of θ given the observed data.

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta|\text{data})$$

Let's start with Bayes:

$$P(\theta|\text{data}) = \frac{P(\text{data}|\theta)\, P(\theta)}{P(\text{data})}$$

Since P(data) does not depend on θ, it is a constant when maximizing over θ. So:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\text{data}|\theta) P(\theta)$$

* $P(\text{data}|\theta)$ is the likelihood, same thing MLE uses
* $P(\theta)$ is the prior, your belief before seeing data

The posterior $P(\theta|data)$ is the updated belief about θ after observing data

Multiplying probabilities causes underflow (same problem with Naive Bayes)

$$\hat{\theta}_{MAP} = \arg\max_{\theta} [\log P(data|\theta) + \log P(\theta)]$$

→ MAP as regularization

In machine learning, you fit model parameters w by maximizing log-likelihood (MLE):

$$\hat{w}_{MLE} = \arg\max_{w} \log P(data|w)$$

But this overfits - W can grow arbitrarily large to fit noise.

If we put Gaussian prior on W (centered at zero, variance σ^2),

MAP becomes:

$$\hat{w}_{MAP} = \arg\max_{w} \left[\log P(data|w) - \frac{||w||^2}{2\sigma^2} \right]$$

The penalty term $||w||^2/2\sigma^2$ is L2 regularization (ridge regression).

It's literally just the log of a Gaussian prior.

Summary:

MLE asks: What θ best explains the data?

MAP asks: What θ is most plausible, given both the data and what I already believed?

→ Confidence Interval

* Interval is a lower and upper limit
* Confidence level is the probability the interval contain θ.

A confidence interval is a range of values, computed from sample data, that is likely to contain the true population parameter with a specified level of confidence

→ Significance level (α) and confidence level

confidence level = $1-\alpha$

Significance level = α

α is the probability that the procedure fails ie produces an interval that doesn't contain the true parameter.

→ Critical value (z^* or t^*)

The critical value defines how wide the interval is. It comes from the distribution of the estimator (typically normal or t-distribution).

For a two-sided interval at confidence level $1-\alpha$, you split into two tails ($\frac{\alpha}{2}$ each)

$$z^* = z_{1-\alpha/2} \quad (z_{\alpha/2} \leq z \leq z_{1-\alpha/2})$$

→ Margin of error

$$MoE = z^* \cdot \text{Standard error}$$

Standard error (SE) is the standard deviation of the estimator:

* For a mean: $SE = \sigma/\sqrt{n}$ (or $s/\sqrt{n}$ if σ is unknown)
* For a proportion: $SE = \sqrt{\hat{p}(1-\hat{p})/n}$

Margin of error = Critical value × standard error

Putting it all together:

$$CI = \hat{\theta} \pm MoE = \hat{\theta} \pm z \cdot SE$$

For a population mean:

$$CI = \bar{x} \pm z^* \cdot \frac{\sigma}{\sqrt{n}}$$

When we don't know σ?

If the population standard deviation is unknown (almost always the case), we use the sample standard deviation s and the t-distribution instead of normal:

$$CI = \bar{x} \pm t^*_{n-1} \cdot \frac{s}{\sqrt{n}}$$

For large n (typically $n>30$), $t^* \approx z^*$ and the difference doesn't matter much

* Confidence interval are a sample mean with a margin of error added to each side
* Confidence level is the probability a confidence interval contains (for example 95%)
* Ideally you have both high confidence and a narrow interval.
* Large samples (more data) will give a narrower interval
* Decreasing confidence level will also shrink the interval (we rarely see <90%)

→ Calculation steps

* Find the sample mean
* Define a desired confidence level (1-K)
* Get the critical value ($z_{1-\alpha/2}$)
* Find the standard error ($\frac{\sigma}{\sqrt{n}}$)
* Find the margin of error.
* Add/subtract the margin of error to the sample mean to obtain CI

→ Assumptions

* Simple random sample
* Sample size > 30 or population is approximately normal.

→ Difference between Confidence and Probability

Confidence has to do with the success rate of constructing the unfictence Interval and its not the probability that a specific interval contains the population mean.

→ Hypothesis testing

Hypothesis testing is a way to tell if some belief we have about a population is likely to be frue or false.

Hypothesis testing is a method for deciding whether the evidence in a sample supports a specific claim about a populatin. We start with two

competing claims, gather data and use probability to determine whether the data is unusual enough to reject one claim in favor of the other.

→ The two hypotheses:

* Null hypothesis (H_0): The default assumption — usually "nothing is happening", "no effect". This is what we assume true until evidence forces us to abandon it e.g H_0: the coin is fair ($P=0.5$)
* Alternate hypothesis (H_1 or H_a): The claim we are trying to find evidence for — usually "there is an effect" e.g H_1: the coin is biased ($p \neq 0.5$)

We can only reject H_0 or fail to reject H_0. We never prove H_1, we just find data inconsistent with H_0.

→ One-sided vs two-sided

* Two-sided (two-tailed): $H_1: \mu \neq \mu_0$ — testing if the parameter differs in direction.
* One-sided (one-tailed): $H_1: \mu > \mu_0$ or $H_1: \mu < \mu_0$ — testing if it differs in a specific direction

→ Significance level (α)
The threshold for rejecting H_0 — the maximum acceptable probability of falsely rejecting a true H_0

→ P-value is the probability of observing data as extreme or more extreme than what we actually saw, a umin H_0 is true.

p-value = P(data this extreme $|H_0$)

Interpretation:

* Small p-value → data is unlikely under H_0 → reject H_0
* Large p-value → data is consistent with H_0 → fail to reject H_0

Decision rule:

* If p-value $< \alpha$ → reject H_0
* If p-value $\geq \alpha$ → fail to reject H_0

Type I and Type II errors

	H_0 true	H_0 false
Reject H_0	Type I error (α)	Correct (power)
Fail to reject H_0	Correct.	Type II error (β)

Type I error (α): Rejecting H_0 when it is actually true. False positive

Type II error (β): Failing to reject H_0 when it is actually false. False negative.

power $= 1 - \beta$: The probability of correctly rejecting H_0 when it is false.

Higher power = better at detecting real effects.

There is a tradeoff: lowering α (being stricter about rejecting H_0) increases β (we miss more real effects). We can only reduce both simultaneously by increasing sample size n.