Marginal Probability :-

Probability of specific event occuring independent of other events

$$P(A) = \sum_b P(A, B=b) \text{ or } \sum_b P(A \cap B=b)$$

Condtional probability

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

Given B already happened what is probability of A

Joint probability :— two or more events occurring at same time

$P(A \cap B)$ or $P(A, B)$

Sum Rule

For disjoint events

$$P(X=x_i) = \sum_{j=1}^{L} P(X=x_i, Y=y_j)$$

Product Rule

$$P(X=x_i, Y=y_j) = P(Y=y_j \mid X=x_i) \cdot P(X=x_i)$$

or

$$P(X=x_i, Y=y_j) = P(X=x_j \mid Y=y_j)\, P(Y=y_j)$$

Bayes theorem

$$P(X=x_j \mid Y=y_j) = \frac{P(Y=y_j \mid X=x_j) \cdot P(X=x_j)}{P(Y=y_j)}$$

By sum Rule $P(Y=y_j) = \sum_{i=1}^{L} P(Y=y_j \mid X=x_i) \cdot P(X=x_i)$

Probability distribution function

function that describes the likelihood of a continious random variable taking a particular value

PDF → Continious random variable

PMF → disrete random variable.

key characteristics

(1) non negative.

$$f(x) \geq 0 \quad \text{for all } x \text{ (random variable)}$$

(2) Total area under the curve.

total area under the PDF curve across the entire range of possible values equals 1

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

(3) Probability of an interval :- probability of that the random variable falls within a specific interval $[a, b]$ can be found by $P(a \leq X \leq b) = \int_a^b f(x)\,dx$

Cumulative distribution function (CDF)

~ is a function that describes the probability that a random variable takes on a value less than or equal to a specific value.

For random variable X

$$F_X(x) = P(X \leq x)$$

Small x represent specific value

→ As x approach negative infinity CDF approach 0

$$\lim_{x \to -\infty} F(x) = 0$$

→ As x approach positive infinity CDF approach 1

$$\lim_{x \to \infty} F(x) = 1$$

Applications

CDFs allow for the calculation of probability of a random variable falling within a range

$$\boxed{P(a \leq X \leq b) = F(b) - F(a)}$$

Expectation :-

~ provides a measure of "center" or "average" of a random variable's distribution. it give an idea of what value you can expect to obtain from random variable after many trials

⇒ **Discrete Random variable :-**

if X is a discrete random variable with possible values $x_1, x_2, \dots x_n$ and corresponding probabilities

$P(X = x_i) = p_i$ then Expectation $E(x)$ is

$$E(X) = \sum_{i=1}^{n} x_i * p_i$$

⇒ **Continous Random variable :-**

if X is continous random variable with a PDF $f(x)$ then

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

Properties of Expectation

① Linearity

$$E(aX+bY) = aE(X) + bE(Y)$$

a and b are constants, X is any random variable

② non negative :-

if X is non negative random variable

then $E(X) \geq 0$

③ Expectation of Constant : if c is a constant

then $E(c) = c$

Variance

it measure how much the values of random variable X differ its from the expected value(mean)

$$Var(X) = E[(X - E(X))^2]$$

$$= E(X^2) - (E(X))^2$$

properties of variance

① non negativity

$$Var(X) \geq 0 \quad \text{for all } X$$

② Units :- variance is in squared unit of random variable

③ Linearity :- For any random variable X and Y and constant a and b

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2ab\, Cov(X,Y)$$

if X and Y are independent then $Cov(X,Y) = 0$

④ variance of constant is zero

Covariance

measure of relationship btw two random variables it indicate extent to which variable change together.

positive covariance :- both increase together.
negative covariance :- one increase other decrease
zero covariance :- no linear relationship

$$\boxed{Cov(X,Y) = E[(X - E(X))(Y - E(Y))]}$$

if X takes value $x_1, x_2, x_3 \cdots x_n$

with probability $P_1, P_2, P_3 \cdots P_n$

Y takes value $y_1, y_2, y_3 \cdots y_n$ with same probability

Expand

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

$$Cov(X,Y) = \sum_{i=1}^{n} (x_i - E(X))(y_i - E(Y)) \cdot P_i$$

or

$$\boxed{Cov(X,Y) = E(XY) - E(X)\,E(Y)}$$

For sample size n using ☆

$$\boxed{Cov(X,Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}$$

Gaussian Distribution or Normal distribution

bell shaped curve.

For Sigle real value x

$$N(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} e\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

μ is mean

σ is Standard deviation

Variance σ^2

Multivariate Gaussian Distribution

For single real value variable x

$$N(x|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} e\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

μ = mean

d = dimension

Σ Covariance matrix $D \times D$

Likelihood function for Gaussian distribution

~ expresses the probability of observing the data given specific parameters (mean and variance)

For a set of independent observations $X = [x_1, x_2 \ldots x_n]$ from a gaussian distribution with unknown mean μ and variance σ^2

probability of dataset given by

$$p(X|\mu,\sigma^2) = \prod_{n=1}^{N} N(x_n|\mu,\sigma^2)$$

$\prod$ stand for product

product of gaussian distribution

Log likelihood function is

$$\ln(p(X|\mu,\sigma^2) = \frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2)$$

To estimate parameters μ and σ^2 we maximize the log likelihood functions

$$\mu_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\sigma^2_{ML} = \frac{1}{n} \sum_{i=1}^{n} (x - \mu_{ML})^2$$

Differentiate w.r.t μ

μ — w.r.t σ^2

which is mean

variance

Independence :-

Two events A and B are independent if the occurrence of one doesnot effect the occurrence of other.

Ex Flipping a coin, getting head
Rolling a die and getting 4

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

Mutual Exclusivity

Two events A and B are mutually exclusive if they cannot occur at the same time.

Ex Rolling a die and getting 2
Rolling a die and getting 5

$$P(A \cap B) = 0$$

key differences

Independence :- Events can occur together
mutual exclusivity :- Events cannot occur together.

Entropy :-

Entropy is a measure of uncertainty or randomness associated with a random variable.

Given a discrete random variable X with possible outcomes $x_1, x_2 \cdots x_n$ and corresponding probabilities $P(x_1), P(x_2) \cdots P(x_n)$ then entropy $H(X)$ is defined as

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

Higher Entropy :- indicates more uncertainty. if all outcomes are equally likely, entropy is maxi

Lower Entropy :- indicates less uncertainty. if outcome is certain (probability of 1) then entropy is zero.

properties

① non negative, always greater than or equal zer

② additive

$$H(X,Y) = H(X) + H(Y)$$

Joint Entropy

quantifies measure of uncertainity associated with two random variables considered together, it quantifies the total amount of information required to describe the outcome of both variable simultaneously

For two discrete R.V X and Y

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} P(x,y) \log_2 P(x,y)$$

$P(x,y)$ is joint probability distribution of X and Y

Higher joint entropy :- indicate greater uncertainty or complexity in relationship btw the variables

Lower joint entropy :- suggests a more predictable or constrained relationship.

Properties :-

① Relation to individual Entropies :- The joint entropy is related to the individual entropies

$$H(X,Y) = H(X) + H(Y|X)$$

where $H(Y|X)$ is conditional entropy of Y given X

(2) if x and y are independent

$$H(X, Y) = H(X) + H(Y)$$

Conditional Entropy

Conditional entropy is a measure of the amount of uncertainity remaining about a random variable given that the value of another random variable is known. it quantifies how much additional information is needed to describe the outcome of one variable when the outcome of another variable is already known

For two discrete R.V X and Y, conditional Entropy of Y given X

$$H(Y|X) = -\sum_{x \in X} P(x) \sum_{y \in Y} P(y|x) \log_2 P(y|x)$$

Where

$P(y|x)$ is conditional probability of Y given X

$P(x)$ is probability of X

High conditional entropy:- indicate that knowing X does not provide much information about Y (Y remains uncertain)

low conditional Entropy:- Suggest that knowing X significantly reduces uncertainty about Y.

properties

Relation shipto Joint Entropy

$$H(Y|X) = H(X,Y) - H(X)$$

Conditional entropy of X ondy can be derived from the total uncertainity in X and y minus uncertainity in X alone.

if Y is independent of X :-

$$H(Y|X) = H(Y)$$

in this case knowing X provides no additional information about y

Chain Rule of Entropy

For a sequence of random variable $X_1, X_2, X_3 \cdots X_n$ the joint entropy $H(X_1, X_2, \cdots X_n)$ can be expressed using the chain rule.

$$H(X_1, X_2, \cdots X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \cdots + H(X_n|X_1, X_2, \cdots X_{n-1})$$

Cross Entropy

- measures the difference btw two probability distributions. it quantifies how well one probability distribution approximates another, often used in context such as classification task.

For two discrete probability distribution P and Q the cross entropy $H(P, Q)$ is defined as

$$H(P, Q) = -\sum_{x} P(x) \log_2 Q(x)$$

where, P is true distribution (often ground truth)
Q is the estimated distribution (often model prediction)

The sum is taken over all possible outcomes x.

Higher cross entropy :- indicates a larger difference btw two distributions
meaning Q doesnot accurately represent P

Lower cross entropy :- indicate the Q is a good approximation of P

Mutual Information

mutual information btw two R.V measure the amount of information that one conveys about the other. it measure the average reduction in uncertainity about x that result from learning about y

For two discrete R.V X and Y, mutual information is defined as

$$I(X;Y) = H(X) - H(X|Y)$$

or it can be expressed as Joint entropy

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

Zero mutual information :- if $I(X;Y) = 0$ the variables are independent, knowing one doesnot provide any information about the other

positive mutual information :- indicates a dependency btw the variables, knowing one variable reduces uncertainity about the other.

properties

① Symmetry $I(X;Y) = I(Y;X)$

② chain rule $I(X;Y,Z) = I(X;Y) + I(X;Z|Y)$

~~$I(X;Y,Z) = I(X;Y) + I(X;Z)$~~

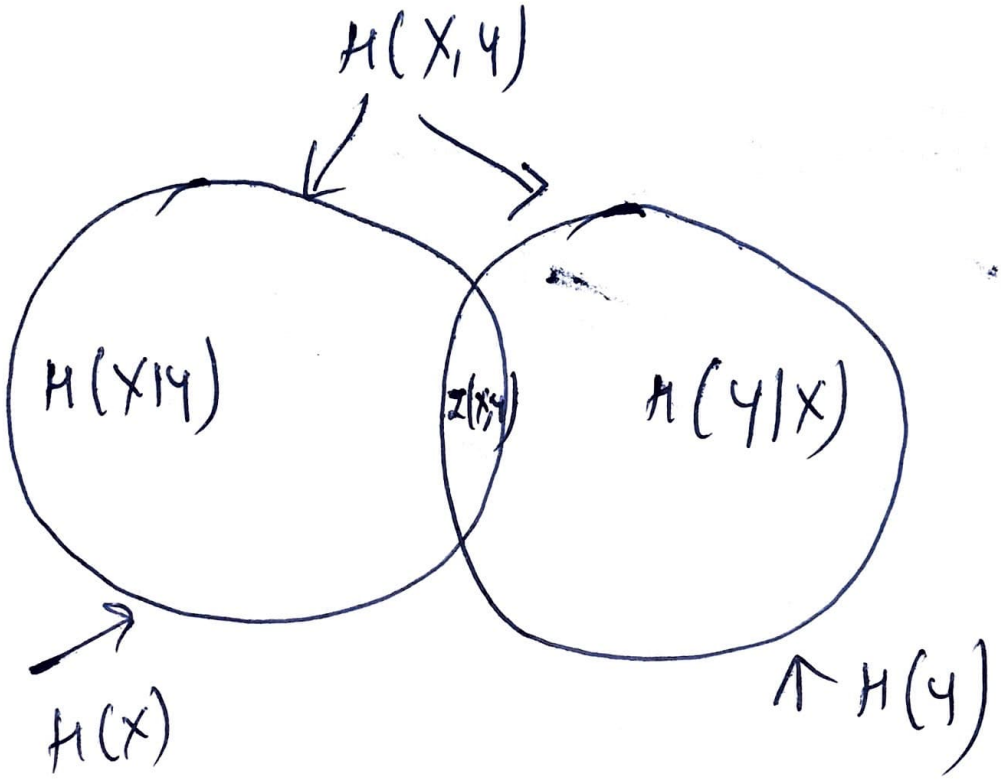
Venn diagram summary of concepts and relationship

Entropy $H(X)$, $H(Y)$

Joint Entropy $H(X,Y) = H(X) \cup H(Y)$

Conditional Entropy $H(X|Y)$, $H(Y|X)$

mutual information $I(X;Y) = H(X) \cap H(Y)$



Distance $D(X,Y)$ Btw two random variable X and Y

The amount by which the joint entropy of two random variable exceeds their mutual information is a measure of the "distance" btw them

$$D(X,Y) = H(X,Y) - I(X;Y)$$

Joint — mutual.

Kullback-Leibler distance, or KL divergence

quantifies the difference btw two probability distributions. Specifically, it measures how one probability distribution diverges from second, expected distribution.

For two discrete probability P (true distribution) and Q (the approximate distribution) the KL divergence from Q to P is defined as

$$D_{KL}(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

$P(x)$ is the probability of event x under distribution P

$Q(x)$ is the probability of event x under distribution Q.

Covariance matrix

~ is a matrix that summarizes the covariances between multiple random variables. each element in matrix represents the covariance btw two variables and the diagonal elements represent the variance of each variable

For n random variables $X_1, X_2, \ldots X_n$ the Covariance matrix Σ is defined as

$$\Sigma = \begin{bmatrix} Cov(X_1, X_2) & Cov(X_1, X_2) \cdots Cov(X_1, X_n) \\ Cov(X_2, X_1) & Cov(X_2, X_2) \cdots Cov(X_2, X_n) \\ \vdots & \\ Cov(X_n, X_1) & Cov(X_n, X_2) \cdots Cov(X_n, X_n) \end{bmatrix}$$

→ if Covariance matrix have only diagonal elements then all random variable $X_1, X_2 \ldots X_p$ are independent to each other. or uncorrelated

Autocovariance Covariance with some R.V

$$C = E\left[(\underline{X} - \mu_{\underline{X}})(\underline{X} - \mu_{\underline{X}})^H\right]$$

Covariance matrix is always Conjugate Symmetric (Hermitian)

Q Calculate SVD, eigen vectors and eigenvalues

$$\begin{bmatrix} 4 & 0 \\ 3 & -5 \end{bmatrix}$$

Soln $A = U \Sigma V^T$

Step 1 Find V^T

$$A^T.A = \begin{bmatrix} 4 & 3 \\ 0 & -5 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 3 & -5 \end{bmatrix} = \begin{bmatrix} 16+9 & 0-15 \\ 0-15 & 0+25 \end{bmatrix}$$

$$= \begin{bmatrix} 25 & -15 \\ -15 & 25 \end{bmatrix}$$

To Find Eigen values

$$|A^TA - \lambda I| = 0 \rightarrow \text{characteristic eqn.}$$

$$\begin{vmatrix} 25-\lambda & -15 \\ -15 & 25-\lambda \end{vmatrix} = 0$$

$$\boxed{\lambda^2 - S1\lambda + S2 = 0}$$

$S_1 = \text{Trace}(A^TA) = 25+25 = 50$

$S_2 = \begin{vmatrix} 25 & -15 \\ -15 & 25 \end{vmatrix} = 625 - 225 = 400$

$\lambda^2 - 50\lambda + 400 = 0$

$\lambda^2 - 40\lambda - 10\lambda + 400 = 0$

$\boxed{\lambda = 40, \ \lambda = 10}$ eigen values

Eigen vector at 40

$$A^TA \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 40 \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

$25v_1 - 15v_2 = 40v_1$ — ①

$-15v_1 + 25v_2 = 40v_2$ — ②

consider ①

$-15v_1 - 15v_2 = 0$

$v_1 = -v_2$

choosing $v_1 = 1$ we get

$$v_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$L_2 \text{ Norm} = \sqrt{(1)^2 + (-1)^2} = \sqrt{2}$

$$\text{Eigen vector} = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$$

eigen vector at 10

$$A^TA \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 10 \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

$25v_1 - 15v_2 = 10v_1$ — ①

$-15v_1 + 25v_2 = 10v_2$ — ②

consider ①

$15v_1 - 15v_2 = 0$

$v_1 = v_2$

choosing $v_1 = 1$ we get

$$v_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$L_2 \text{ Norm} = \sqrt{(1)^2 + (1)^2}$

$$\text{Eigen vector} = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

Eigenvectors matrix V

$$V = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

Step 2 Calculate Σ which is singular values at diagonal in descending order.

$$\delta^1 = \sqrt{40} = 2\sqrt{10} \qquad \delta^2 = \sqrt{10}$$

$$\Sigma = \begin{bmatrix} 2\sqrt{10} & 0 \\ 0 & \sqrt{10} \end{bmatrix}$$

step 3 Calculate U

$$A \cdot A^T = \begin{bmatrix} 4 & 0 \\ 3 & -5 \end{bmatrix} \begin{bmatrix} 4 & 3 \\ 0 & -5 \end{bmatrix} = \begin{bmatrix} 16+0 & 12+0 \\ 12+0 & 9+25 \end{bmatrix}$$

$$= \begin{bmatrix} 16 & 12 \\ 12 & 34 \end{bmatrix}$$

$$|A \cdot A^T - \lambda I| = 0$$

Same as step - ①

$$\lambda^2 - S1\lambda + S2 = 0$$

$S1 = 16+34 = 50$

$S2 = 544 - 144 = 400$

$$\lambda^2 - 40\lambda - 10\lambda + 400 = 0$$

$$\lambda = 40, \lambda = 10$$

Eigenvalues 40, 10

Eigen vector at $\lambda = 40$

$$AA^T \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = 40 \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

$$\begin{bmatrix} 16 & 12 \\ 12 & 34 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 40u_1 \\ 40u_2 \end{bmatrix}$$

$$16u_1 + 12u_2 = 40u_1 \quad -①$$
$$12u_1 + 34u_2 = 40u_2 \quad ②$$

Consider ①

$$12u_2 = 24u_1$$
$$u_2 = 2u_1$$

$u_1 = 1$ we get

$$u_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Normalize u_1 :

$$\|u_1\| = \sqrt{1^2 + 2^2} = \sqrt{5}$$

Eigen vector at $\lambda = 10$

$$\begin{bmatrix} 16 & 12 \\ 12 & 34 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 10u_1 \\ 10u_2 \end{bmatrix}$$

$$16u_1 + 12u_2 = 10u_1 \quad -①$$
$$12u_1 + 34u_2 = 10u_2 \quad -②$$

consider ①

$$-12u_2 = 6u_1$$
$$u_1 = -2u_2$$

Choose $u_2 = 1$

$$u_1 = -2$$
$$u_2 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$$

$$\|u_2\| = \sqrt{(-2)^2 + 1^2} = \sqrt{5}$$

$$U = \begin{bmatrix} 1/\sqrt{5} & -2/\sqrt{5} \\ 2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix} \quad \Sigma = \begin{bmatrix} 2\sqrt{10} & 0 \\ 0 & \sqrt{10} \end{bmatrix} \quad V^* = \begin{bmatrix} \frac{1}{\sqrt{2}} & 1/\sqrt{2} \\ \frac{-1}{\sqrt{2}} & 1/\sqrt{2} \end{bmatrix}$$

$$A = U \Sigma V^T = \begin{bmatrix} 1/\sqrt{5} & -2/\sqrt{5} \\ 2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix} \begin{bmatrix} 2\sqrt{10} & 0 \\ 0 & \sqrt{10} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -1/\sqrt{2} \\ +1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

Correlation

~ measure the strength and direction of a linear relationship between two different variables

For two R.V X and Y Correlation Coefficient (often denoted ρ or r) is

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

$Cov(X,Y)$ is covariance between X and Y

σ_X and σ_Y are standard deviations of X and Y

The Correlation Coefficient range from -1 to $+1$

$+1$: perfect positive linear relationship
-1 : perfect negative linear relationship
0 : No linear relationship

Correlation is widely used in statistics and machine learning to explore how one variable might predict or effect another.

Auto Correlation

a measure correlation of a signal with a delayed copy of itself. it assesses how a signal or time series correlates with itself over various time lags For a time series X_t, the autocorrelation at lag k (denoted $\rho(k)$) is defined as

$$\rho(k) = \frac{Cov(X_t, X_{t+k})}{\sigma_x^2}$$

Where

$Cov(X_t, X_{t+k})$ is covariance btw the time series and itself at a time lag of k

σ_x^2 is variance of time series

key Differences

→ Correlation is btw two different variables; autocorrelation is within a single variable or time series at different time steps

→ Correlation tells you how two variable move in relation to each other while auto correlation tells you how a variable relate to itself over time.

Stationarity

~ is a key concept in time series analysis refering to a time series whose statistical properties do not change over time, in other words, a time series is stationary if its mean, variance and auto correlation structure remain constant overtime. Stationarity is important because many statistical models especially time series forecasting, require data to be stationary for accurate modeling and prediction.

Types

① Strict Stationarity

A time series is said to be strictly stationary if the joint distribution of any subset of observation is the same, no matter when the observations are taken. Mathematically for a time series X_t, it means that for any time points $t_1, t_2, \dots t_n$ and any

any lag k, the joint distribution of $(X_{t1}, X_{t2}, \ldots X_{tn})$ is identical to that of $(X_{t1+k}, X_{t2+k}, \ldots, X_{tn+k})$

② Weak stationarity (Second order stationarity)

~ is a less strict condition and is more commonly used in practice. A time series is weakly stationary if it satisfies the following conditions

i) Constant mean :- mean of series $E[X_t]$ is constant and doesnot depend on time t

$$E[X_t] = \mu \quad \text{for all } t$$

ii) Constant variance :- variance of series $Var(X_t)$ is constant and doesnot depend on time t.

$$Var(X_t) = \sigma^2, \text{ for all } t$$

iii) Constant autocovariance structure :- autocovariance btw values at two points t and t+k depends only on the time lag k, not on specific time t

$$Cov(X_t, X_{t+k}) = \gamma(k)$$