Statistical Foundation for Machine Learning

Evaluation Pattern:

1. Assignments / Quiz (10)
2. Class Test – 2 (20)
3. Mid sem (30)
4. End sem (40)

Linear Algebra in Machine learning

$\underline{x}$ → vector (either – or bold)
x → scalar

★ $\underline{a}^T \underline{x} = b$ $\quad \underline{a}^T \rightarrow 1 \times n \quad \underline{x} \rightarrow n \times 1$ & $b \Rightarrow 1 \times 1$
T → Transpose (ie converting rows into columns or vice versa)
Here a & x both are vectors / matrices
b is a scalar

★ Array is a vector 1D → 1×n & 2D → m×n (List of numbers) → Matrix

★ Linear Algebra is used throughout ML

Scalar

- single number (1×1)

Vector

- Array of numbers arranged in order
- Each identified by an index
- written in lower case
- Vectors as points in space
  - ↳ can be represented
- 1×n (1 axis)

Matrices

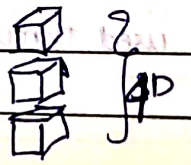
- 2D array
- Denoted by bold typeface $\mathbf{A}$
- If A has shape of height m & width n with real values then
  $A \in R^{m \times n}$
  R → Real number
  C → complex

Tensor

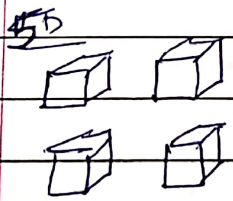
- 2+ dimension array (more than 2 axes)
  Eg: RGB color images has 3 axes
  Represented by $\underline{C}_{i,j,k}$
- It is an array of nbrs arranged in a regular grid ~~with~~ variable no. of axes

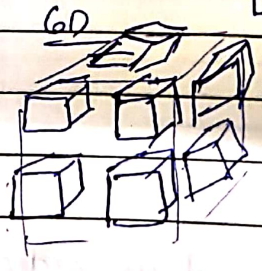
Shape of Tensors

1D ⇒ vector (1D - tensor = vector)
2D ⇒ Matrix (2D - tensor = matrix)
3D ⇒ Tensor (3D - tensor = tensor)
4D - tensor ⇒ tensor



5D/6D/





Eg: 3D-tensor = images
4D- tensor = video

* When we say 1D / 2D
  D → no. of dependent variables

Transpose of Matrix

→ Mirror image across diagonal

$$A = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \\ a_{13} & a_{23} & a_{33} \end{bmatrix}$$

$$A^T = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

- Vectors are special case of matrix
- Scalar is a matrix with 1 element

Matrix Addition

- Add matrices to each other if they have same shape by adding corresponding elements
- If the matrices don't have same shape, we can add 0 to right

Matrix Multiplication

9+15

- $A_{m \times n} \times B_{n \times p} = C_{m \times p}$

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 4 & 5 \end{bmatrix} = \begin{bmatrix} 10 & 13 \\ 22 & 29 \end{bmatrix}$$

- Zeroes can be added to adjust shape of matrix & is called as up sampling

★ Up sampling improves resolution by adding no. of samples

Dot product

- Also called Hadamard product
- Represented by $A \odot B$
- Application: cosine similarity

Matrix Product Properties

- Distributivity over addition

$$A(B + C) = AB + AC$$

- Associativity:

$$A(BC) = (AB)C$$

- Not commutative:

$$AB \neq BA \Rightarrow \text{Not true always}$$

- True in case of identity matrix

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} 4 & 6 \\ 5 & 7 \end{bmatrix} = \begin{bmatrix} 14 & 20 \\ 27 & 46 \end{bmatrix}$$

$$\neq$$

$$\begin{bmatrix} 4 & 6 \\ 5 & 7 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 22 & - \\ - & - \end{bmatrix}$$

- Dot product between vectors is commutative

$$x^T y = y^T x$$

Eg: $x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ $y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$

$$x^T = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

$$(x_1y_1 + x_2y_2 + x_3y_3)$$

$$y^T = \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} \quad x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \Rightarrow \begin{bmatrix} x_1y_1 + x_2y_2 + x_3y_3 \end{bmatrix}$$

- $(AB)^T = A^T B^T$

Applications of Linear Algebra in Machine learning

1. While designing a classifier (linear) using Neural network.

Eg:

- Consider there is a labeled data set which consists of 3 types of images - Dogs, cats & ships from which we derive weights & bias (2D tensor & 1D)
- Now image being 2D matrix is first converted into 1D ie into vector (testing phase) & weights & biases are applied to it to get its relevant class

Formula : $W^T x + b = \text{output}$

↓

Linear equation

Linear Transformation:

- $Ax = b$

  $A \in R^{n \times n}$ & $b \in R^n$

  $A^{-1} A x = A^{-1} b$

Identity & Inverse Matrices

- $A A^{-1} = A^{-1} A = I$

I dimension is obtained for A in case of $A^{-1} A$

- $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ $\quad A^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

There is an issue with this method of calculating inverse ie if determinant $(ad - bc)$ becomes 0.

Linear Equations: Closed form solution

→ Since calculation of A^{-1} has shortcoming hence this method is not used by Gaussian Elimination method followed by back substitution.

Disadvantage:

- numerical instability due to divison by small no.
- $O(n^3)$ for $n \times n$ matrix

How many solutions exist for $Ax = b$

1. Unique solution

$2x + 3y = 9$

$9x + 2y = 11$

x & y are unique

different slope & intercept → unique for both eqns

2. Infinite solution

$2x + 3y = 4$

$4x + 6y = 8$

Both the eqns are same & slope of eqn 1 = slope of eqn 2

Intercept of eqn 1 = Intercept of eqn 2

3. No solution

$2x + 3y = 4$

$2x + 3y = 9$

No solution since slope of eqn 1 = slope of eqn 2

Intercept of eqn 1 ≠ Intercept of eqn 2

Span of set of Vector

- Linear combination of set of vectors

$$Ax = \sum_i x_i A_i$$

- A linear combination of vectors $\{v^{(1)} \dots v^{(n)}\}$ then coefficients c_i is $\sum c_i$
- Linearly Independent of v

$$c_1 v_1 + c_2 v_2 + \dots + c_n v_n = 0$$

This is possible only if $c_1 = c_2 = \dots = c_n = 0$

cosine similarity : Tells how close 2 vectors are
If 2 vectors are close or same, angle is 0
If 2 vectors are different, they're orthogonal ie angle is 90

- Linearly Dependent

$c_1 v_1 = -(c_2 v_2 + \cdots + c_n v_n)$

- Application of span of vectors : Noise cancellation

Use of vector in Regression

L^p Norm

$$||x||_p = \left(\sum_i |x_i|^p\right)^{\frac{1}{p}}$$

L^2 Norm (Euclidean norm):
- distance between origin & point x
- same as euclidean distance

Use of norm in Regression

1. Linear Regression

x: a vector, w: weight vector

$$y(\underline{x}, \underline{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\underline{x})$$

$$y(\underline{x}, \underline{w}) = w_0 + w_1 x_1 + \cdots + w_d x_d \quad \phi \cdot = \underline{w}^T \underline{x}$$

$\underline{\phi}_j$ = Non linear basis fn

2. Loss function

$$\tilde{E}(\underline{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(\underline{x}_n, \underline{w}) - t_n\}^2 + \frac{\lambda}{2} ||\underline{w}^2||$$

Second term is a weighted norm; called as regularizer to prevent overfitting

Angle between Vectors

- Dot product of 2 vectors can be written in terms of L^2 norms & angle θ between them

$$\underline{x}^T \underline{y} \Rightarrow ||x||_2 ||y||_2 \cos\theta$$

- cosine between 2 vectors is measure of similarity between them

Efficiency of diagonal Matrix

- $\text{diag}(v)\, x = v \odot x$
- Easily calculable inverse matrix

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 9 & 0 \\ 0 & 0 & 3 \end{bmatrix} \quad A^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/9 & 0 \\ 0 & 0 & 1/3 \end{bmatrix}$$

Symmetric Matrix

- Transpose of symmetric matrix = ^{DG} matrix

$$A = A^T$$

Eg: $A = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 3 \\ 2 & 3 & 0 \end{bmatrix}$

Unit Vector

- A vector with unit norm

$$\|x\|_2 = 1$$

Orthogonal Vector

$$x^T y = 0$$

If vectors have nonzero norm, vectors are 90° to each other

Orthonormal vectors

$$A^T = A = AA^T = I$$

$$A^T = A^{-1}$$

- orthogonal & have unit norm

Eigen value decomposition / Matrix decomposition

- Decomposition of integer into prime factors
  Eg: 12 → 2×2×3

- In case of square matrix, we follow Eigen value decomposition
  $A = V \, diag(\lambda) \, V^{-1}$
  $V \rightarrow$ Eigen vectors (orthogonal matrix ie $V^{-1} = V^T$ or $VV^{-1} = I$)
  $\lambda \rightarrow$ eigen values

Eigen vector : It is a non zero vector v such that multiplication by A (original vector) only changes scale of v
(scales vector A)

$$Av = \lambda v$$

$\lambda \rightarrow$ scalar value / eigen value

If $A_{2\times2} \rightarrow$ we get 2 λs & λ can be +ve, -ve & 0
λ as 0 → no vector is present for decomposition

$Av = \lambda v$ can be written as $Av - \lambda v = 0$

$$(A - \lambda I)v = 0$$

λ being scalar value hence ~~by~~ we multiply by I

$Av = \lambda v$ has a non zero solution if $|A - \lambda I| = 0$

$|A - \lambda I| = (\lambda_1 - \lambda)(\lambda_2 - \lambda) \cdots (\lambda_n - \lambda)$

$n \rightarrow$ dimension of A

$$\sqrt{\left(\frac{1}{\sqrt{14}}\right)^2 + \left(\frac{2}{\sqrt{14}}\right)^2 + \left(\frac{3}{\sqrt{14}}\right)^2} = 1$$

Eg :

$$|A - \lambda I| = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$$

$$= \begin{bmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{bmatrix} \Rightarrow (2-\lambda)^2 - 1$$

$|A - \lambda I| = 3 - 4\lambda + \lambda^2$

$= \lambda^2 - 4\lambda + 3$

$= (\lambda - 3)(\lambda - 1) \Rightarrow (\lambda - 1)(\lambda - 3) = 0$ since we assume it has non zero solution

$\lambda = 1, 3$

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} V = \lambda V$$

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \lambda \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

$2v_1 + v_2 = v_1 \Rightarrow v_1 + v_2 = 0$

$v_1 + 2v_2 = v_2$

$v_1 + v_2 = 0 \Rightarrow v_1 = -v_2$

If $v_1 = 1, v_2 = -1$

When $v_1 = 1$

$$V_{\lambda=1} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$V_{\lambda=3} = \begin{bmatrix} 3 \\ -3 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 3v_1 \\ 3v_2 \end{bmatrix}$$

$2v_1 + v_2 = 3v_1$

$v_1 + 2v_2 = 3v_2$

$-v_1 + v_2 = 0$

$v_1 = v_2$

When $v_1 = 1$

$$V_{\lambda=3} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Eigen decomposition of A

$$A = V \, diag(\lambda) V^{-1}$$

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

$$diag(\lambda) = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix} \Rightarrow \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

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

$$= \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 3 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

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

$V \cdot V^T \neq I$

V is not orthogonal

$$V^T = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

so $V \, diag(\lambda) V^T \neq A$.

Decomposition of symmetric Matrix

If A is a symmetric matrix

$A A^T = I$. ⇒ orthogonal

- $A = V \, diag(\lambda) V^T$
- The eigen vectors in this case will be unit vectors

- Eigen decomposition is not unique ie λ_1 & λ_2 are same or all eigen values are same

Application / what does eigen decomposition tell us?

1. If $\lambda = 0 \Rightarrow$ Matrix is singular

Matrix can't be inversed if matrix is singular

2. Useful to optimize quadratic expression of form

$f(x) = x^T \Lambda x \rightarrow \|x\|_2 = 1$

↓

eigen vector

3. A matrix whose eigen values are +ve ⇒ +ve definite
4. If all eigen values are -ve ⇒ -ve definite
5. If some eigen values are +ve & some are -ve or 0 → positive semidefinite

Singular Value Decomposition (SVD)

Shortcomings of EVD

- only defined for square matrix
- Requires calculation of inverse matrix which has $O(n^3)$

$A = UDV^T$

- U & V are orthogonal
- A is any matrix

U → left singular vector $\quad U = AA^T$ or A^TA

V → right singular vector $\quad V = A^TA$

If A is $m \times n$

U → $m \times m$

D → $n \times n$

V → $n \times n$

Eg:

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

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

$|A^TA - \lambda I| = 0$

$\lambda = 10$ & 40

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

$$\begin{bmatrix} 25 & -15 \\ -15 & 25 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 10v_1 \\ 10v_2 \end{bmatrix}$$

$25v_1 - 15v_2 = 10v_1 \quad v_1 = v_2$

$15v_1 = 15v_2$

$v_{\lambda=10} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ → $v_{\lambda=10}$ after l_2 norm ⇒ $\begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$

l_2 norm $= \sqrt{1+1} = \sqrt{2}$

When $\lambda = 40$

$25v_1 - 15v_2 = 40v_1$

$-15v_1 = +15v_2$

$v_1 = -v_2$ If $v_2 = 1$ or $v_1 = 1$ either

$v_{\lambda=40} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

(Note: v_1 or v_2 can be taken & any value can be taken)

l_2 norm $= \sqrt{2}$

$v_{\lambda=40} = \begin{bmatrix} -1/\sqrt{2} \\ +1/\sqrt{2} \end{bmatrix}$ If

~~Diagonal~~ Σ (DI) $= \begin{bmatrix} \sqrt{40} & 0 \\ 0 & \sqrt{10} \end{bmatrix}$

e.g.

This should be in descending order of eigen values whereas we don't care in case of EVD

We used $\sqrt{}$ of λ values since

$A^T \cdot A = \|A\|^2 = \Sigma^2$

Hence we take square root to take Σ

we don't bother doing $\sqrt{\lambda}$ in EVD since eigen values are obtained directly from Λ.

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

- We need to place $v_{\lambda = \text{highest nbr}}$ first then $v_{\lambda = \text{next highest nbr}}$ ...

$A \cdot A^T = \begin{bmatrix} 4 & 0 \\ 3 & -5 \end{bmatrix} \begin{bmatrix} 4 & 3 \\ 0 & -5 \end{bmatrix} = \begin{bmatrix} 16 & 12 \\ 12 & 34 \end{bmatrix}$

$|AA^T - \lambda I| = 0$

$AA^T - \lambda I = \begin{bmatrix} 16 & 12 \\ 12 & 34 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$

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

$= (16-\lambda)(34-\lambda) - 144$

$\lambda_1 = 10 \quad , \quad \lambda_2 = 40$

$AA^T \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 10 \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$

$\begin{bmatrix} 16 & 12 \\ 12 & 34 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 10v_1 \\ 10v_2 \end{bmatrix}$

$16v_1 + 12v_2 = 10v_1$

$6v_1 = -12v_2$

$v_1 = -2v_2$

Let $\Rightarrow v_1 = -2 \; ; v_2 = 1$

$v_{\lambda=10} = \begin{bmatrix} -2/\sqrt{5} \\ 1/\sqrt{5} \end{bmatrix}$

$\begin{bmatrix} 16 & 12 \\ 12 & 34 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 40v_1 \\ 40v_2 \end{bmatrix}$

$16v_1 + 12v_2 = 40v_1$

$24 \qquad 2v_1 = -12v_2$

$2v_1 = -v_2$

Let $v_1 = 1 \; ; \; v_2 = -2$

$v_{\lambda=40} \begin{bmatrix} 1/\sqrt{5} \\ -2/\sqrt{5} \end{bmatrix}$

$\Sigma = \begin{bmatrix} \sqrt{40} & 0 \\ 0 & \sqrt{10} \end{bmatrix}$

$V = \begin{bmatrix} 1/\sqrt{5} & -2/\sqrt{5} \\ -2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix}$

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

$V^T = \begin{bmatrix} 1/\sqrt{5} & -2/\sqrt{5} \\ -2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix}$

$U\Sigma V^T = \begin{bmatrix} 16 & 12 \\ 12 & 34 \end{bmatrix} \begin{bmatrix} \sqrt{40} & 0 \\ 0 & \sqrt{10} \end{bmatrix} \begin{bmatrix} 1/\sqrt{5} & -2/\sqrt{5} \\ \cdots & \cdots \end{bmatrix}$

$= \begin{bmatrix} 16\times\sqrt{40} & 12\sqrt{10} \\ 12\sqrt{40} & 34\sqrt{10} \end{bmatrix} \times V^T$

$= \begin{bmatrix} 16\times\sqrt{40} & 12\sqrt{10} \\ 12\sqrt{40} & 34\sqrt{10} \end{bmatrix} \begin{bmatrix} 1/\sqrt{5} & -2/\sqrt{5} \\ -2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix}$

$= \begin{bmatrix} 8\sqrt{2} & \cdots \end{bmatrix}$

$12\sqrt{10}\times 2$

$16\times\sqrt{10}\times 3$

$16\times 2\sqrt{2}$

4×10

8×5

$2\sqrt{10}$

$32\sqrt{2}$

Applications of SVD

1. We can calculate inverse of matrix / Moore Penrose Pseudo inverse

$$A = UDV^T$$

$$A^+ = (UDV^T)^+ \qquad + \rightarrow \text{pseudoinverse}$$

$$= (V^T)^+ D^+ U^+$$

$$A^+ = V D^+ U^T$$

Since U & V are orthogonal

$$UU^T = I$$

$$VV^T = I$$

Using SVD, we can calculating A^{-1} very easily

2. PCA or dimension reduction

Book References

SVD – Pattern Recognition – Bishop (examples given)

Linear Algebra – Schaum Freeze

8/9 → Quiz
(Till today's session)

DATE

⇒ Populus Book

Probability Theory in ML

- Quantify the uncertainity from 0-1
- Used in classification within ML
- Bayes theorem is used for evaluating probability in ML

Random Variables

- Enable us to associate numerical quantities with some experiment.
- Random variables are associated always for uncertainty
- Deterministic variable - variable whose outcome is already known. The probability of such is always 1
- Random variables are independent to each other ie outcomes don't affect each other
- Random Process: associated with Random variable with time x(w)

Eg: E - Tossing a coin
$U = \{H, T\}$

Marginal Probability

Eg: consider there are 2 R.V. box & fruit
Box - Red, blue
Fruit - apple, orange

Marginal Probability: (Individual Probability)
Probability of an apple (only 1 probability)
Eg: $P(F = a)$

Conditional Probability:
Given we've orange, what is probability we choose blue box
eg: $P(B = b \mid F = 0)$

Joint Probability:
Probability of orange & blue box
Eg: $P(B = b, f = 0)$

→ Joint

$$P(X=x_i, Y=y_j) = \frac{n_{ij}}{N}$$

$$= \frac{n_{ij}}{c_i} \times \frac{c_i}{N}$$

$\frac{n_{ij}}{c_i}$ → $P(Y=y_j | X=x_i)$ Joint conditional

$\frac{c_i}{N}$ → Marginal probability

$$= P(Y=y_j | X=x_i) \times P(x_i)$$

$$\therefore P(Y=y_j | X=x_j) = \frac{P(X=x_i, Y=y_j)}{P(x_i)}$$

⇩

Bayes Theorem

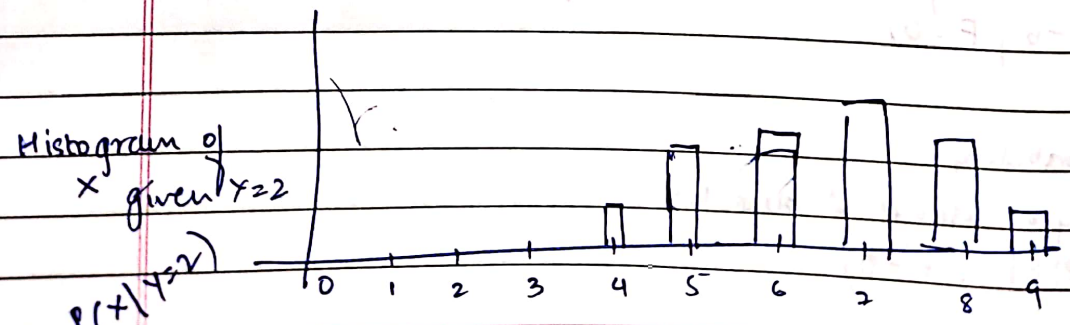
$$P(x, y) = \frac{P(x|y)}{P(y)} = \frac{P(y/x)}{P(x)}$$

$$P(x|y) = \frac{P(y/x)\,P(x)}{P(y)} = \cancel{P(x,y)}$$

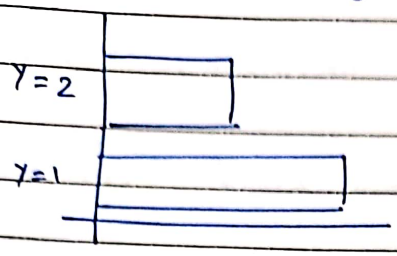
$$P(x) = \sum_y P(x, y)$$

$$= \sum_y \frac{P(x|y)}{P(y)} = \sum_x \frac{P(y|x)}{P(x)}$$

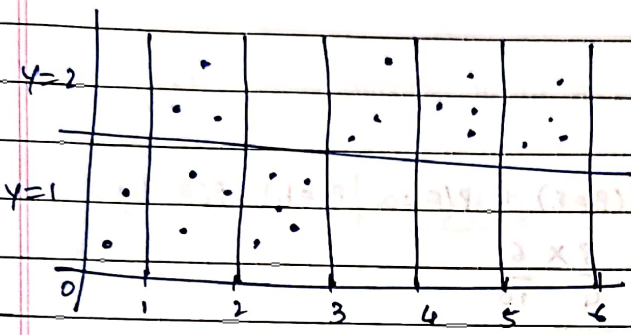
Joint Distribution over 2 variables



Histogram of Y / Marginal Probability of Y



P(X, Y)



Example: Find probability of box as red & orange using Bayes theorem

$P(B=r \mid F=0)$

Given:

$P(F=0 \mid B=r) = \frac{3}{4}$ $\quad P(B=r) = \frac{4}{10}$ $\quad P(F=0 \mid B=b) = \frac{1}{4}$ $\quad \frac{12}{48}$

$P(F=a \mid B=r) = \frac{1}{4}$ $\quad P(B=b) = \frac{6}{10}$ $\quad P(F=a \mid B=b) = \frac{3}{4}$ 4=

$$P(B=r \mid F=0) = \frac{P(F=0 \mid B=r) \times P(B=r)}{P(F=0)} = \frac{3}{4} \times \frac{4}{10}$$

3 - apples
1 - orange

$$P(F=0) = \sum P(F=0, B=r) + P(F=0, B=b)$$

$$= \underbrace{P(F=0 \mid B=r)}_{P(B=r)} \times + \underbrace{P(F=0 \mid B=b)}_{P(B=b)} \times$$

$$= \frac{3}{4} \times \frac{4}{10} + \frac{1}{4} \times \frac{6}{10} = \frac{3}{\frac{4}{10}} + \frac{1}{\frac{6}{10}}$$

$$\frac{15 \times 6}{8 \times 6} + \frac{5 \times 4}{12 \times 4} = \frac{90+20}{48} = \frac{110}{48}$$

$= P(F=0 \mid B=r) \times P(B=r) + P(F=0 \mid B=b) \times P(B=b)$

$= \frac{9}{20}$

$P(B=b \mid F=a) = ?$

$= \frac{P(F=a \mid B=b) \times P(B=b)}{P(F=a)}$

$= \frac{\frac{3}{4} \times \frac{6}{10}}{P(F=a)}$

$P(F=a) = P(F=a \mid B=r) \times P(B=r) + P(F=a \mid B=b) \times P(B=b)$

$= \frac{1}{4} \times \frac{4}{10} + \frac{3}{4} \times \frac{6}{10}$

$= \frac{22}{40} = \frac{11}{20}$

$= \frac{\frac{3 \times 6 / 3}{\cancel{4} \cancel{10}}}{\frac{11}{\cancel{20}\ 2}} = \frac{9}{11}$

Independent Variables

$P(X \neq Y) = P(X) \cdot P(Y)$

X & Y are independent

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

Mutually Exclusive :

$P(A \cap B) = \phi = \{\}$ $\quad P(A \cap B) = P(A) \cdot P(B)$

Not ~~independent variables are necessarily mutually exclusive~~

Sum, Product, Bayes for continuous Random variable

$p(x) = \int p(x,y)\, dy$ ⇒ Marginal

$p(x,y) = p(y|x)\, p(x)$ ⇒ Joint

$p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(x|y)\, p(y)}{p(x)}$ ⇒ conditional

Expectation : It is average value of function $f(x)$ or x, represents weighted average of all possible values of R.V where weights are probabilities of each value

- Denoted by E
- $E[x] = \sum_x x\, p(x)$ — discrete R.V

Eg: x – 10 values from 0 to 9

Each value is equiprobable (uniform distribution)

$P(0) = P(1) = \cdots = P(9) = \frac{1}{10}$

$E[x] = (0)\left(\frac{1}{10}\right) + \left(\frac{1}{10}\right) + 2\left(\frac{1}{10}\right) + 3\left(\frac{1}{10}\right) + 4\left(\frac{1}{10}\right) + 5\left(\frac{1}{10}\right) + 6\left(\frac{1}{10}\right) + 7\left(\frac{1}{10}\right) + 8\left(\frac{1}{10}\right) + 9\left(\frac{1}{10}\right)$

$= \frac{45}{10} = 4.5$

★ Expectation of a deterministic signal remains constant

- $E[f(x)] = \sum_x f(x)\, p(x)$

eg: $f(x) = x^2$

- If $f(x)$ is continuous

$E[f(x)] = \int f(x)\, p(x)\, dx$

- If x is continuous

$E[x] = \int x\, p(x)\, dx$

★ Conditional Expectation?

Conditional Expectation:

$$E[f] = \sum_x p(x|y) f(x)$$

Variance:

- Define: variability in $f(x)$ around its mean value.
- $\text{var}[f] = E\left[\{f(x) - E[f(x)]\}^2\right]$ $\quad E[(x-\mu)^2] \Rightarrow E[f(x)]$

$$= E\left[\{f(x)\}^2 - 2f(x)E[f(x)] + E[f(x)]^2\right]$$

$$= E[f(x)^2] - E[2f(x)E[f(x)]] + E\left[E[f(x)]^2\right]$$

This will be a constant

$$= E[f(x)^2] - 2E[f(x)]^2 + E[f(x)]^2$$

$$= E[f(x)^2] - E[f(x)]^2$$

Covariance: (notation)

- Define between 2 different random variables
- $\text{cov}[x,y] = E_{x,y}\left[\{x - E[x]\}\{y - E[y]\}\right]$

$$= E_{x,y}\left[xy - xE[y] - yE[x] + E[x]E[y]\right]$$

$$= E_{x,y}[xy] + E_{x,y}[-xE[y] - yE[x]] + E_{x,y}[E[x]E[y]]$$

$$= E_{x,y}[xy] - E[x]E[y] - E[x]E[y] + E[x]E[y]$$

$$= E_{x,y}[xy] - E[x]E[y]$$

$E_{x,y}[-xE[y]] \Rightarrow -E_{x,y}[x]\,E_{x,y}[E[y]]$

$E[y]$ ⇒ since $E[y]$ is an expectation & expectation on expectation remains same & stays same

$-E[x]\,E[y]$

Bayesian Probability

- Probability is quantification of uncertainty
- Likelihood - observe the data
  Prior - Having predetermined knowledge
  Posterior - find answer which is correct

Properties of probability

- No probability can be > 1 ie P(sample space) = 1
- P(Event) Any event) ≥ 0 * $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$
  If E_1 & E_2 are mutually exclusive $P(E_1 \cap E_2) = 0$
- ~~Bayes~~ Posterior $\propto$ Likelihood $\times$ Prior $\Rightarrow$ Bayes theorem
- Sum Rule

$$P(D) = \int P(D|w)\, P(w)\, dw$$

continuous

continuous

Since $P(D|w)$ & $P(w)$ are continuous, hence $\int$

Gaussian Distribution

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

$\mu \Rightarrow$ Mean

$\sigma^2 \Rightarrow$ variance

$\sigma \Rightarrow$ S.D.

Single variate as x has 1 feature

Expectation of x under Gaussian

$$E[x] = \int N(x|\mu, \sigma^2)\, x\, dx$$

$$E[x^2] = \int N(x|\mu, \sigma^2)\, x^2\, dx = \mu^2 + \sigma^2$$

$$\text{var}[x] = E[x]^2 - E[x^2]$$

* If x is multi dimensional

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

$\mu \rightarrow$ Mean

$\Sigma \rightarrow$ covariance matrix $= D \times D$

D - dimensional vector

Krusha Doshi

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

Prove: $\log p(x|\mu,\sigma^2) = \frac{-1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\log\sigma^2 - \frac{N}{2}\log(2\pi)$

$$p(x|\mu,\sigma^2) = \prod_{n=1}^{N} \frac{1}{(2\pi\sigma^2)^{1/2}} \times e^{\left(\frac{-1}{2\sigma^2}(x-\mu)^2\right)}$$

Taking log on both sides

$$\log p(x|\mu,\sigma^2) = \ln\left(\sum_{n=1}^{N}\frac{1}{(2\pi)^{1/2}\sigma}\right) \times \left(\frac{-1}{2\sigma^2}(x-\mu)^2\right)$$

$$= \frac{-1}{2\sigma^2}\sum_{n=1}^{N}(x-\mu)^2$$

$$\ln p(x|\mu,\sigma^2) = \log\left(\prod_{n=1}^{N}\frac{1}{(2\pi\sigma^2)^{1/2}} \times e^{\left(\frac{-1}{2\sigma^2}(x-\mu)^2\right)}\right)$$

By property $\log(a \times b) = \log a + \log b$

$$= \log\left(\sum_{n=1}^{N}\frac{1}{(2\pi\sigma^2)^{1/2}}\right) + \left(\frac{-1}{2\sigma^2}(\tilde{x}-\mu)^2\right)$$

$$= \frac{-1}{2\sigma^2}\left(\sum_{n=1}^{N}(x_n-\mu)^2\right) - \log\left(\sum_{n=1}^{N}(2\pi\sigma^2)^{1/2}\right)$$

(By property of $\log\left(\frac{1}{a}\right) = -\log a$)

$$= \frac{-1}{2\sigma^2}\left(\sum_{n=1}^{N}(x_n-\mu)^2\right) - \frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2$$

Independent Events

$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2) = P(E_1, E_2)$

$P(E_2 | E_1) = P(E_1) \cdot P(E_2) / P(E_1) = P(E_2)$

- 2 mutually exclusive events are not independent

$E_1 \cap E_2 = 0$

$P(E_1 \cap E_2) = 0$

But $P(E_1 \cap E_2) = P(E_1) \cdot P(E_2) \neq 0 \Rightarrow P(E_1 \cup E_2) = P(E_1) + P(E_2)$

(Independent event condition)

- 2 identical events are not independent

$P(E_1 \cap E_2) = P(E_1)$

But by independent event defn,

$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$ $\{P(E_1 \cap E_2) = P(E_1, E_2)\}$

But $P(E_1) \cdot P(E_2) \neq P(E_1)$

since $P(E_2) \neq 1$ $P(E_2) < 1$

- If E_1 & E_2 are independent, are their complement too?

$$P(E_1' \cap E_2') = 1 - P(E_1 \cup E_2)$$
$$= 1 - P(E_1) - P(E_2) + P(E_1 \cap E_2)$$
$$= 1 - P(E_1) - P(E_2) + P(E_1) \cdot P(E_2)$$
$$= (1 - P(E_1))(1 - P(E_2))$$
$$= P(E_1') \cdot P(E_2')$$

which means they're independent

Example

$P(\text{clear, raising}) = P(\text{Rising} | \text{clear}) \times P(\text{clear})$

$P(z_1, z_2) = P(z_1 | z_2) \times P(z_2)$

~~$P(\text{clear, raising}) = P(\text{clear} | \text{raising}) \times P(\text{raising})$~~

$P(\text{dry} | \text{clear, rising}) = \dfrac{P(\text{dry, clear, rising})}{P(\text{clear, rising})} = \dfrac{0.40}{0.47} \approx 0.85$

$P(z_1) = \sum P(z_1 | z_3) \times P(z_3) + \cancel{P(z_1)}$

0.07 + 0.1 + 0.11 + 0.12 = 0.33 ; 0.4 + 0.08 + 0.09 + 0.03 = 0.12 ; 0.48

0.11
0.12
23

! P(Cloudy) = P(Cloudy | Dry) P(Dry) + P(Cloudy | wet) P(wet)

= 0.7

Cloudy, wet P(wet)
P(wet)

★ P(E1|E2) ≠ P(E2|E1) ⇒ Inverse of probability

0.47

P(clear|rising) = P(rising, clear) / P(rising)
= 0.47 / 0.67 = 0.7

0.11
0.09
0.2

★ P(rising ∪ falling) = P(rising) + P(Falling) = 1

P(Rising | clear) + P(Falling | clear) = 1
P(clear | falling) + P(clear | rising) ≠ 1

★ P(• | E1) = P(E2 ∪ E3 ∪ E4 | E1) (since there are 4 events E1, E2, E3, E4)

⇓
Any event

P(E3 ∪ E2 ∪ E4 | E1) = P(E2|E1) + P(E3|E1) + P(E4|E1)

E2, E3, E4 ⇒ disjoint

★ P(E1 | E2 ∪ E3 ∪ E4) ≠ P(E1|E2) + P(E1|E3) + P(E1|E4)

⇓
P(E1 | •)

★ If $\bar{E}$ is complement of E

$P(\bar{E_1} | E_2) = 1 - P(E_1|E_2)$

but

$P(E_1|\bar{E_2}) \neq 1 - P(E_1|E_2)$

$\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\left(\log\left(\frac{1}{2}\right)\right) = -\log\left(\frac{1}{2}\right) = \log_2 2 = 1$

$-\frac{1}{2}\left(\log\left(\frac{1}{2}\right)\right)$

$P(\text{Positive} \mid \text{sick}) = P(\text{Positive, sick}) / P(\text{sick}) = 0.009 / 0.01 = 0.9$

$P(- \mid \text{Healthy}) = 0.891 / 0.99 = 0.9$

$P(\text{sick} \mid +) = 0.009 / 0.108 = 0.0833$

$P(\text{Healthy} \mid -) = 0.891 /$

Entropy

- Entropy is measure of randomness of random variable Z.
- Represented by H(Z), for single event / random variable

$$H(Z) = -\sum_{r \in Z} P_Z(Z) \log_2 P_Z(Z)$$

- Entropy is maximum if $P_Z(Z) = \frac{1}{M}$ ie equiprobable for all values of Z
  Eg: Sun rises can rise from any direction / Toss of coin it can be H or T
- It is minimum : $P(Z) = 1$ for a value of Z (others are null)
  Eg: Sun rises in east, there is no uncertainity

- If there several random variables it is called as ensemble of entropy

$$\sum_i p_i = 1$$

$$H = -1 = -\sum_i p_i \log_2 p_i$$ → Marginal probability

- Entropy of an ensemble is simply average of all elements in it it can be computed by their average entropy by weighing each of $\log p_i$ contributions by its probability p_i occuring :

- In case of several events entropy is multiply by weight/probability in ensemble but not in case of entropy with 1 event since there is only 1 event & average would be 1

Example:

Let Z be an alphabet from A to Z but all prob of all alphabets is equiprobable

$$H = -\left[\frac{1}{26} \times 26 \log_2 \frac{1}{26}\right]$$

$= 4.7$

$$H = -\Big[0.105 \times \log_2 0.105 + 0.072 \log_2 0.072 + 0.066 \log_2 0.066 + 0.063 \log_2 0.063 + 0.059 \log_2 0.059 + 0.055 \log_2 0.055 + 0.054 \log_2 0.054 + 0.052 \log_2 0.052 + 0.047 \log_2 0.047 + 0.035 \log_2 0.035 + 0.029 \log_2 0.029 + 0.023 \log_2 0.023\Big]$$

$$= -\Big[-0.105 \times 3.2515 - 0.072(3.8) - 0.066(3.92) - 0.063(3.9) - 0.059(4.08) - 0.055(4.16) - 0.054(4.21) - 0.052(4.26) - 0.047(4.4) - 0.035(4.8) - 0.029(5.1) - 0.023(5.42)\Big]$$

$$= \underline{\underline{2.69}}$$

Conditional Entropy:

Joint Entropy:

- $$H = -\sum_{x,y} p(x,y) \log_2 p(x,y)$$

$$H = \sum_{x,y} p(x,y) \log_2 \frac{1}{p(x,y)}$$

- Joint entropy is additive

$$H(X,Y) = H(X) + H(Y) \qquad p(x,y) = p(x) \cdot p(y)$$

Given: $$H(X,Y) = \sum_{x,y} p(x) \cdot p(y) \log_2 \left(\frac{1}{p(x) \cdot p(y)}\right)$$

$$= -\sum_{x,y} p(x) \cdot p(y) \log_2 \left(p(x) \cdot p(y)\right)$$

$$= -\sum_{x,y} p(x) \cdot p(y) \left[\log_2 p(x) + \log_2 p(y)\right]$$

$$= -\sum_{x,y} p(x) \cdot p(y) \log_2 p(x) + p(x) \cdot p(y) \log_2 p(y)$$

$$H(x) = -\sum_{x} p(x) \log_2 p(x) \qquad H(y) = -\sum_{y} p(y) \log_2 p(y)$$

$$H(x) + H(y) = -\left[\sum_{x} p(x) \log_2 p(x) + \sum_{y} p(y) \log_2 p(y)\right]$$

$$= -\sum_x P(x) \log_2 P(x) \sum_y P(y) - \sum_y P(y) \log_2 P(y) \sum_x P(x)$$

$$\sum_x P(x) = 1 \quad \& \quad \sum_y P(y) = 1$$

$$= -\left[\sum_x P(x) \log_2 P(x) + \sum_y P(y) \log_2 P(y)\right]$$

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

Conditional Entropy

Measures uncertainty about random variable X when R.V Y has taken some random value

$$H(X \mid Y = b_j) = \sum_x P(x \mid Y = b_j) \log_2 \frac{1}{P(x \mid Y = b_j)}$$

↳ This is for single event

$$H(X|Y) = \sum_y P(y) \left[\sum_x P(x|y) \log_2 \frac{1}{P(x|y)}\right]$$

$$= \sum_{x,y} P(y) \cdot P(x|y) \log_2 \frac{1}{P(x|y)}$$

$$= \sum_{x,y} P(x,y) \log_2 \frac{1}{P(x|y)}$$

$$P(x|y) = \frac{P(x,y)}{P(y)} \quad \text{or} \quad P(y|x) \cdot P(x) = P(x,y) \Rightarrow P(x) = \frac{P(x,y)}{P(y|x)}$$

Chain Rule for Entropy

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

$$H(X) + H(Y|X) = -\left[\sum_x P(x) \log_2 P(x) + \sum_{x,y} P(x,y) \log_2 P(y|x)\right]$$

$$= -\left[\sum_x P(x) \cdot \sum_{x,y} P(x,y) \log_2 P(x)\right]$$

$$= -\left[\sum_x 1\right.$$

Krusha Doshi

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

$$= -\sum P(Y|X) P(X) \log_2 [P(Y|X) P(X)]$$

$$= -\sum P(Y|X) P(X) [\log_2 P(Y|X) + \log_2 P(X)]$$

$$= -\left[\sum_{x,y} P(Y|X) \cdot P(X) \log_2 P(Y|X) + \sum_{x,y} P(Y|X) P(X) \cdot \log_2 P(X)\right]$$

$$= -\left[-\sum_{x,y} P(Y|X) \cdot P(X) \log_2 P(Y|X) - \sum_{x} P(X) \log_2 P(X) \sum_{y} P(Y|X)\right]$$

$$= -\left[-\sum_{x,y} P(x,y) \log_2 P(Y|X) + H(X) \sum_{y} P(Y|X)\right]$$

$\sum_{y} P(Y|X) \Rightarrow 1$ since its summation of $P(Y|X)$ & x is already given

$$= -\left[-\sum_{x,y} P(x,y) \log_2 P(Y|X) + H(X)\right]$$

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

Assignment

Prove $H(X_1, X_2, \cdots X_n) \leq \sum_{i=1}^{n} H(X_i)$

$$H(X) = -\sum_{x} P(x) \log P(x)$$

$$\sum_{i=1}^{n} H(X_i) = \sum_{i=1}^{n} -\sum_{x} P(x_i) \log P(x_i)$$

$H(X_1, X_2, \cdots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \cdots + H(X_n | X_n, X_2, \cdots, X_{n-1})$ ⇒ By chain rule for entropy

By formula, $H(X,Y) = -\sum_{x,y} P(x,y) \log P(x,y)$

$\therefore H(X_1, X_2, \cdots, X_n) = -\sum_{x} P(x_1, \cdots, x_n) \log P(x_1, \cdots x_n)$

Each conditional entropy $H(X_i | X_1, \cdots X_{i-1})$ is non negative, indicating that joint entropy will not exceed total entropy of individual components. This follows the rule that $H(X|Y) \leq H(X)$ as existence of known value for Y tend to reduce randomness in X

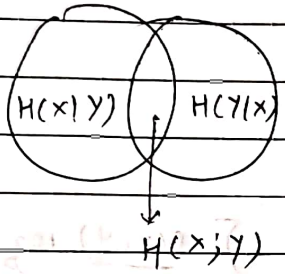
Hence,

$$H(X_1, X_2, \cdots, X_n) \leq \sum_{i=1}^{n} H(X_i)$$

Mutual Information between x & Y

Mutual info between 2 R.V measures amount of information that one conveys about other. Equivalently measures avg reduction in uncertainty about x that results from learning Y

$$H(X;Y) = \sum_{x,y} P(x,y) \log_2 \frac{P(x,y)}{P(x) \cdot P(Y)}$$



- When X & Y are independent, mutual entropy is 0

$$I(X;Y) = \sum_{x,y} P(x,y) \cdot \log_2 \frac{P(x) \cdot P(Y)}{P(x) \cdot P(Y)} \hookrightarrow 1$$

$$= 0 \quad \{\log_2 1 = 0\}$$

Relative Entropy Mutual Info:

- It is always non negative ie ≥ 0

$P(X,X) = P(X)$

$$I(X;X) = \sum_{x} P(x) \log_2 \frac{P(x)}{P(x) \cdot P(x)}$$

self info

$$= \sum_{x} P(x) \log_2 \frac{1}{P(x)}$$

$$= -\sum P(x) \log_2 P(x)$$

$$= H(X)$$

, ①

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

$= I(Y;X)$

$= H(X) + H(Y) - H(X,Y) \rightarrow ②$

Prove:

① $I(X;Y) = H(X) - H(X|Y)$

$= -\sum_x P(x) \log_2 P(x) - \left[-\sum_{x,y} P(x|y) \log_2 P(x|y) \right]$

By Bayes theorem,

$P(x|y) = \frac{P(x,y)}{P(y)}$

$= + \sum_x P(x) \log_2 \frac{1}{P(x)} + \sum_{x,y} \frac{P(x,y)}{P(x)P(y)} \log_2 \frac{P(x|y)}{P(x,y)}$

$= \sum_x P(x) \log_2 \frac{1}{P(x)} + \sum_{x,y} P(x,y) \log_2 \left(\frac{P(x,y)}{P(y)} \right)$

$= \sum_x P(x) \sum_x P(x) \log_2 \frac{1}{P(x)} + \sum_{x,y} P(x,y) \log_2 \left(\frac{P(x,y)}{P(y)} \right)$

$= \sum_x P(x) \log_2 \frac{1}{P(x)} + \sum_{x,y} P(x,y) \log_2 \left(\frac{P(x,y)}{P(x)} \right)$

$I(X;Y) = \sum_{x,y} P(x,y) \log_2 \frac{P(x,y)}{P(x) \cdot P(y)}$

$= \sum P(x|y) \cdot P(y) \log_2 \frac{P(x|y) \cdot P(y)}{P(x) \cdot P(y)}$

$= \sum P(x|y) \cdot P(y) \log_2 \frac{P(x|y)}{P(x)}$

$= \sum P(x|y) \cdot P(y) \log_2 P(x|y) - \sum P(x|y) P(y) \log_2 P(x)$

$= -\left[\sum P(x,y) \log_2 P(x|y) \right] + \left[\sum P(x|y) P(y) \log_2 P(x) \right]$

$$= \left[-H(X|Y) + H(X) \right]$$

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

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

$$= H(X) - \left[-H(Y) + H(X,Y) \right]$$

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

Cross Entropy

If there are 2 distributions $P(x)$ & $q(x)$ over same set of outcomes

$$H(P,q) = -\sum_x P(x) \log q(x)$$

If $P(x) = q(x)$

$$H(P) = \sum_x P(x) \log P(x)$$

- Cross entropy is asymmetric

$$H(P,q) \neq H(q,P)$$

Distance between X & Y (R.V)

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

$D(X,Y) \geq 0$ ∵ distances are non-negative

$D(X,X) = 0$

$D(X,Y) = D(Y,X)$ ∋ symmetric

$D(X,Z) \leq D(X,Y) + D(Y,Z)$ (Triangle Inequality rule)

KL divergence

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

Here $p(x)$ & $q(x)$ are diff probability distributions defined over x

Application :

A model say $x \to y_x$ is a training data has a model, this model cannot be applied on $z \to y_z$ if $p(x)$ & $q(z)$ are very different. They can be determined as different using ~~KL~~ distance between them & we can use KL divergence

$$D_{KL}(P||q) = H(P,q) - H(P)$$

If variables are independent, they are always uncorrelated

If variables are uncorrelated, not necessary are independent. Under condition variables follows Gaussian joint distribution

Correlation & covariance Matrix & Wiener Filter

1. Hermitian Matrix

$C^H = C$

{ what is conjugate
$y = a + jb \Rightarrow y^* = a - jb$ }

$H \rightarrow$ Hermitian operator
$\rightarrow$ conjugate + transpose
$C \rightarrow$ conjugate matrix

2. Covariance Matrix

$$C_{p\times p} = \begin{bmatrix} (x_1 - \mu x_1) \\ (x_2 - \mu x_2) \\ \vdots \\ (x_p - \mu x_p) \end{bmatrix}_{1\times p} \left[\begin{bmatrix} (x_1 - \mu x_1), \cdots (x_p - \mu x_p) \end{bmatrix}^* \right]_{p\times 1}$$

Auto covariance Matrix

$$= \begin{bmatrix} C_{x_1 x_1} & C_{x_1 x_2} & \cdots C_{x_1 x_p} \\ \vdots & & \\ & & C_{x_p x_p} \end{bmatrix}_{p\times p}$$

Prove: covariance matrix is always symmetric

$$\begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

$C = E[(\underline{x} - \mu \underline{x})(\underline{x} - \mu \underline{x})^H]$

Given $C^H = C \Rightarrow$ To prove symmetric

$$C = \begin{bmatrix} C_{x_1 x_1} & C_{x_1 x_2} & \cdots & C_{x_1 x_p} \\ \vdots & & & \\ & & & C_{x_p x_p} \end{bmatrix}_{p\times p}$$

$$C^H = \begin{bmatrix} C_{x_1 x_1} & \cdots & C_{x_p x_1} \\ C_{x_1 x_2} & & \\ \vdots & & \\ C_{x_1 x_p} & & C_{x_p x_p} \end{bmatrix}$$

$$(AB)^H = [(AB)^T]^* = [B^T A^T]^* = B^H A^H$$

$$C^H = E[((\underline{x} - \mu \underline{x})(\underline{x} - \mu \underline{x}))^H]$$

with $A = (\underline{x} - \mu \underline{x})$, $B = (\underline{x} - \mu \underline{x})$

$$= E[(\underline{x} - \mu \underline{x})^H (\underline{x} - \mu \underline{x})^H]$$

$$= \underline{\underline{C}}$$

covariance matrix is always hermitian matrix

If mean is 0 Then

$\Rightarrow E[\underline{x}\,\underline{x}^H]$

$\Downarrow$

correlation / variance / auto-covariance

$\mu_x = E[x]$

$\sigma_x^2 = E\left[|(x-\mu_x)|^2\right]$

$C_{\underline{x},\underline{y}} = E\left[(\underline{x}-\mu_{\underline{x}})(\underline{y}-\mu_{\underline{y}})^H\right]$

Here $\underline{x}$ & $\underline{y}$ are vectors

$C_{x,y} = E\left[(x-\mu_x)(y-\mu_y)^H\right]$

Here x & y are scalars vectors

$= E\left[(x-\mu_x)(y-\mu_y)^*\right]$

$= E[xy^*] - E[\mu_x y^*] - E[x\mu_y^*] + E[\mu_x \mu_y^*]$

$\because \mu_x = E[x]$ Expectation & expectation on expectation is always same ($\mu_x = E[x]$)

$= E[xy^*] - \mu_x E[y^*] - \mu_y^* E[x] + \mu_x \mu_y^*$

$= E[xy^*] - \mu_x \mu_y^* - \mu_y^* \mu_x + \mu_x \mu_y^*$

$= E[xy^*] - \mu_x \mu_y^*$

If $C_{x,y} = 0$ i.e.

$E[xy^*] = \mu_x \mu_y^* \Rightarrow$ uncorrelated

[correlation between 2 random variables equals to product of means]

Random process → Random variable becomes fn of time

DATE

Stationarity

- The data is now wrt time

First order stationary

$E[x(n)] = \mu_x = \mu$ ⇓ No Power = 1 ⇒ First order

$x(n)$ → Discrete time random process

Eg: So if there are 500 samples are taken & it results into mean μ, now again another 500 samples are taken & mean is taken again which comes out to μ & so on
Then the series is stationary

Second order stationary / wide sense stationarity (WSS)

1. $E[x(n)] = \mu$
2. $E[\underbrace{x(n)}_{A}\ \underbrace{x^*(n+k)}_{B}] = r(k)$ — k : lag/gap

$r^*(-k) = E[x(n+k)\, x^*(n)]$ — $n+k = m$

$= E[x(m)\, x^*(m-k)]$ — $n = m-k$

$= r(k)$

Hence, it is Hermitian/conjugate symmetry

If $k = 0$,

$r(0) = E[|x(n)|^2] \geq 0 \Rightarrow$ Average power of signal

Auto correlation :

- Correlation with same random variable
- $\mu = 0$

$$R_{p\times p} = E[\underline{x}\ \underline{x}^H]$$

$$= \begin{bmatrix} E[|x(n)|^2] & E[x(n)x^*(n-1)] & \cdots & E[x(n)x^*(n-p)] \\ \vdots & & & \\ \cdots & & & E[|x(n-p)|^2] \end{bmatrix}$$

To get this value ⇒ n-(n-k)

$$= \begin{bmatrix} r(0) & r(1) & \cdots & r(p) \\ \vdots & & & \\ r(-p) & r(1-p) & \cdots & r(0) \end{bmatrix}$$

$\because r^*(-k) = r(k)$

$$= \begin{bmatrix} r(0) & r(1) & \cdots & r(p) \\ \vdots & & & \\ r^*(p) & r^*(p-1) & \cdots & r(0) \end{bmatrix}$$

Toeplitz Matrix ⇒ If upper diagonal values are known then we can compute lower diagonal values.

DATE 03 11 24

Hermitian Matrix

⇒ If $T_{p\times p}$ is a hermitian matrix or a toeplitz matrix $\{\bar{T}_{p\times p} = T^H_{p\times p}\}$

$$T\underline{x} = \lambda \underline{x} \quad (\underline{x} \neq 0) \rightarrow \text{Eigen value decomposition}$$

(T: Matrix, λ: Eigen value, $\underline{x}$: Eigen vector)

$$(T\underline{x})^H = \lambda^* \underline{x}^H \quad \{\text{order doesn't matter in } \lambda x \text{ since } \lambda \text{ is a scalar}\}$$

$$\underline{x}^H T^H = \lambda^* \underline{x}^H$$

$$\underline{x}^H T = \lambda^* \underline{x}^H$$

$$\underline{x}^H T \underline{x} = \lambda^* \underline{x}^H \underline{x} \quad \text{(Multiplying } \underline{x} \text{ both the sides)}$$

$$\underline{x}^H \lambda \underline{x} = \lambda^* \underline{x}^H \underline{x} \quad \text{We multiply after } \underline{x}^H T \text{ since multiplication might get messed up)}$$

$$\lambda \underline{x}^H \underline{x} = \lambda^* \underline{x}^H \underline{x}$$

$$(\lambda - \lambda^*) \underline{x}^H \underline{x} = 0$$

$$\therefore (\lambda - \lambda^*) = 0$$

$$\therefore \lambda = \lambda^*$$

Hence, eigen values are always real

{By property, $\underline{x}^H \underline{x} = \sum_{i=1}^{p} |x_i|^2$ $\therefore \underline{x}^H \underline{x} \neq 0$}

⇒ Prove $\underline{x}_1^H$ & $\underline{x}_2$ are orthogonal if $\lambda_1 \neq \lambda_2$

$$T\underline{x}_1 = \lambda_1 \underline{x}_1 \quad \text{—①}$$

$$T\underline{x}_2 = \lambda_2 \underline{x}_2 \quad \text{—②}$$

Apply Hermitian on eqn ①

$$(T\underline{x}_1)^H = (\lambda_1 \underline{x}_1)^H$$

$$\underline{x}_1^H T^H = \lambda_1^* \underline{x}_1^H \quad \text{—③}$$

Multiply $\underline{x}_2$ on eqn ③

$$\underline{x}_1^H T^H \underline{x}_2 = \lambda_1^* \underline{x}_1^H \underline{x}_2$$

$$\underline{x}_1^H T \underline{x}_2 = \lambda_1^* \underline{x}_1^H \underline{x}_2$$

$$\underline{x}_1^H \lambda_2 \underline{x}_2 = \lambda_1^* \underline{x}_1^H \underline{x}_2$$

$$\lambda_2 \underline{x}_1^H \underline{x}_2 = \lambda_1^* \underline{x}_1^H \underline{x}_2$$

$$(\lambda_2 - \lambda_1^*) \underline{x}_1^H \underline{x}_2 = 0 \quad \{\because \lambda_1 = \lambda_1^*\}$$

since $\lambda_1 \neq \lambda_2$ $\therefore \underline{x}_1^H \underline{x}_2 = 0 \Rightarrow$ orthogonal

Example:

$$A = \begin{bmatrix} 1+2j & 2 \\ 3 & 4+2j \end{bmatrix} \quad A^H = \begin{bmatrix} (1+2j)^* & 3^* \\ 2^* & (4+2j)^* \end{bmatrix}$$

$$A^H A = \begin{bmatrix} (1+2j)^* & 3 \\ 2 & (4+2j)^* \end{bmatrix} \begin{bmatrix} 1+2j & 2 \\ 3 & 4+2j \end{bmatrix}$$

$$= \begin{bmatrix} (1+2j)^*(1+2j) + 9 & +4 + (4+2j)^*(4+2j) \end{bmatrix}$$

$$= (1-2j)(1+2j) + 13 + (4-2j)(4+2j)$$

$$= (1-(4j^2)) + 13 + 16 - 4j^2$$

$$= 30 - 8j^2$$

$$= 38$$

$$= \begin{bmatrix} 1+4+9 & 2-2j+12+6j \\ 2+4j+12-6j & 4+16+4 \end{bmatrix}$$

$$= \begin{bmatrix} 14 & 14+4j \\ 14-2j & 24 \end{bmatrix}$$

$\Rightarrow$ If all eigen values & vectors are combined into matrix

$$T \begin{bmatrix} e_1 & e_2 & e_3 & \cdots & e_p \end{bmatrix} = \begin{bmatrix} \lambda_1 e_1 & \lambda_2 e_2 & \cdots & \lambda_p e_p \end{bmatrix}$$

lets call as E

$$TE = \begin{bmatrix} e_1 & e_2 & \cdots & e_p \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & & 0 \\ \vdots & & & & \\ 0 & \cdots & & & \lambda_p \end{bmatrix}$$

$$TE = ED$$

$$E^H E = \begin{bmatrix} e_1^H \\ e_2^H \\ \vdots \\ e_p^H \end{bmatrix} \begin{bmatrix} e_1 & e_2 & \cdots & e_p \end{bmatrix} \quad \text{--- (4)}$$

Apply E^H both the sides on eqn (1)

$$I E E^H = E D E^H$$

As we proved in previous theorem, x_i^H & x_i are orthogonal

$$I I = E D E^H$$

$$\boxed{I = E D E^H}$$ Decomposition of Toeplitz matrix

* $\{ E E^H = I \}$ or $\{ E^{-1} = E^H \}$ We use Hermitian since it avoids inverse matrix computation
* Computation of inverse matrix is difficult for non square matrix
* If $|A| = 0$ then it might be problematic

Weiner Filters

- Important for optimization in image processing → considers both degradation fn & noise
- Uses supervised learning
- Also called Minimum Mean sq errors or least square errors filter
  ⇒ It removes additive noise & inverts blurring simultaneously
- Minimizing error

$$e(n) = y(n) - d(n)$$

error — o/p — actual value

$E^2 = E[e^2(n)]$ ⇒ We apply statistical operation since there will be randomness in signals

Mean squared error

$$= E[e(n)\, e^H(n)]$$
$$= E[e(n)\, e^T(n)] \quad \{\because e^H(n) \text{ is } e^T(n) \text{ as } e(n) \text{ has real values}\}$$
$$= E[(d(n) - y(n))\, e^T(n)]$$
$$= E[(d(n) - \underline{w}^T \underline{x}(n))\, e^T(n)] \quad \{\because y(n) = \underline{w}^T \underline{x}(n)\}$$
$$= E[(d(n) - \underline{w}^T \underline{x}(n))(d(n) - \underline{w}^T \underline{x}(n))^T]$$

⇓

Weiner filter is linear estimation of original image

$[w_0\ w_1]\begin{bmatrix} x(n) \\ x(n-1) \end{bmatrix} = w_0 x(n) + w_1 x(n-1)$

$x^T(n) = [x(n)\ x(n-1)]\begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$

$= E\left[d^2(n) - W^T x(n) d(n) - W^T x(n) d(n) + W^T x(n) x^T(n) W\right]$ ∵ $d(n)$ is a scalar, ∴ $d^T(n) = d(n)$

$= E\left[d^2(n) - \underline{W}^T \underline{x}(n) d(n) - d(n)(\underline{W}^T \underline{x}(n))^T + \underline{W}^T \underline{x}(n)(\underline{W}^T \underline{x}(n))^T\right]$

$= E\left[d^2(n) - \underline{W}^T \underline{x}(n) d(n) - d(n) \underline{x}^T(n) \underline{W} + \underline{W}^T \underline{x}(n)(\underline{x}^T(n) \underline{W})\right]$

$\{ W^T x(n) = x^T(n) W$
$(W^T x(n))^T = x^T(n) W$
$= W^T x(n)$

$= E\left[d^2(n) - \underline{W}^T \underline{x}(n) d(n) - d(n) \underline{W}^T \underline{x}(n) + \underline{W}^T \underline{x}(n) \underline{x}^T(n) \underline{W}\right]$

$= E\left[d^2(n) - 2\underline{W}^T \underline{x}(n) d(n) + \underline{W}^T \underline{x}(n) \underline{x}^T(n) \underline{W}\right]$

$= E[d^2(n)] - 2E[\underline{W}^T \underline{x}(n) d(n)] + E[\underline{W}^T \underline{x}(n) \underline{x}^T(n) \underline{W}]$

$= \sigma_d^2 - 2\underline{W}^T E[\underline{x}(n) d(n)] + \underline{W}^T E[\underline{x}(n) \underline{x}^T(n)] \underline{W}$

$$E^2 = \sigma_d^2 - 2\underline{W}^T \underline{p} + \underline{W}^T \underline{R} \underline{W} \quad \text{— (1)}$$

$\underline{p}$ ⇓ Cross correlation b/w $x(n)$ & $d(n)$

{ we don't apply expectation on W since its static in nature & doesn't change with time

Gradient descent optimization

objective : W should be optimized ie coefficients should be optimized

Take first order derivative of W for eqn (1)

$\nabla_W E^2 = \underline{0}$ (∵ we want minima)

Since E^2 is a vector so we need to equate with 0 vector

$$\nabla_W E^2 = \frac{\partial}{\partial w_k}\left(\sigma_d^2 - 2\underline{W}^T \underline{p} + \underline{W}^T \underline{R} \underline{W}\right)$$

let $A = \underline{W}^T \underline{p}$

second order derivative → +ve → minima

second order derivative → 0 → Maxima

$\frac{\partial A}{\partial W_k} =$

$A = W_0 P(0) + W_1 P(1) + \cdots + W_k P(k) + \cdots + W_N P(N)$

$\frac{\partial A}{\partial W_k} = P(k)$

$\nabla_W A = \begin{bmatrix} \frac{\partial A}{\partial W_0} \\ \frac{\partial A}{\partial W_1} \\ \vdots \\ \frac{\partial A}{\partial W_N} \end{bmatrix} = \begin{bmatrix} P(0) \\ P(1) \\ \vdots \\ P(k) \\ \vdots \\ P(N) \end{bmatrix} = P$

let $B = \underset{(N+1)\times 1}{W^T} \underset{(N+1)\times(N+1)}{R} \underset{(N+1)\times 1}{W}$

$B = \sum_{i=0}^{N} W_i \sum_{j=0}^{N} R_{ij} W_j$

$\nabla_W B = \begin{bmatrix} \frac{\partial B}{\partial W_0} \\ \frac{\partial B}{\partial W_1} \\ \vdots \\ \frac{\partial B}{\partial W_k} \\ \vdots \\ \frac{\partial B}{\partial W_N} \end{bmatrix} = \begin{bmatrix} \\ \\ \\ \\ \\ \\ \end{bmatrix}$

$\frac{\partial B}{\partial W_k} B = \sum_{\substack{i=0 \\ i\neq k}}^{N} W_i \sum_{j=0}^{N} R_{ij} W_j + W_k \sum_{j=0}^{N} R_{kj} W_j$

(j does contain k^{th} term)

$\frac{\partial B}{\partial W_k} = \sum_{\substack{i=0 \\ i\neq k}}^{N} W_i R_{ik} + \frac{\partial}{\partial W_k}\left[W_k \sum_{\substack{j=0 \\ j\neq k}}^{N} R_{kj} W_j + W_k R_{kk} W_k \right]$

$= \sum_{\substack{i=0 \\ i\neq k}}^{N} W_i R_{ik} + \sum_{\substack{j=0 \\ j\neq k}}^{N} R_{kj} W_j + 2 W_k R_{kk}$

$= \sum_{i=0}^{N} W_i R_{ik} + \sum_{j=0}^{N} R_{kj} W_j + W_k R_{kk} + W_k R_{kk}$

$$= \sum_{i=0}^{N} R_{ji} W_i + \sum_{j=0}^{N} R_{kj} W_j$$

$\because$ R is hermitian, $R^T = R \therefore R_{ji} = R_{ij}$

$$= \sum_{i=0}^{N} W_i R_{ki} + \sum_{j=0}^{N} R_{kj} W_j$$

$$= \left[W_0 R_{k0} + W_1 R_{k1} + \cdots W_N R_{kN} \right] + \left[R_{k0} W_0 + R_{k1} W_1 + \cdots + R_{kN} W_N \right]$$

$$= 2 \sum_{i=0}^{N} W_i R_{ki}$$

$$\frac{\partial B}{\partial W_0} = 2 \sum_{i=0}^{N} W_i R_{0i} = 2 \left[W_0 R_{00} + W_1 R_{01} + \cdots + W_N R_{0N} \right]$$

$$\frac{\partial B}{\partial W_1} = 2 \sum_{i=0}^{N} W_i R_{1i} = 2 \left[W_0 R_{10} + W_1 R_{11} + \cdots + W_N R_{1N} \right]$$

$$\therefore \nabla_{W} B = \begin{bmatrix} \frac{\partial B}{\partial W_0} \\ \frac{\partial B}{\partial W_1} \\ \vdots \\ \frac{\partial B}{\partial W_k} \\ \vdots \\ \partial B / \partial W_N \end{bmatrix} = 2RW$$

(R is matrix, P is vector)

$$\nabla_W E^2 = 0 - 2P + 2RW = 0$$

$$= -2P + 2RW = 0$$

W is optimal weights / optimal filter

$$\therefore P = RW \qquad \therefore \boxed{W^* = R^{-1} P}$$

(When R is orthogonal then we can write $R^{-1} = R^H$)

Ex: $\underline{R} = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$ $\underline{P} = \begin{bmatrix} 0.2 \\ 0.1 \end{bmatrix}$

$W = ?$ using Gradient descent opti

$|R| = ad - bc$

$= 1 - 0.81$

$= 0.19$

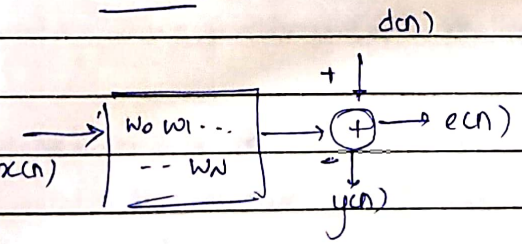
$$R^{-1} = \frac{1}{0.19}\begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix} = \begin{bmatrix} 5.26 & -4.73 \\ -4.73 & 5.26 \end{bmatrix}$$

$W = R^{-1} P$

$$= \frac{1}{0.19}\begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}\begin{bmatrix} 0.2 \\ 0.1 \end{bmatrix}$$

$$= \begin{bmatrix} 5.26 & -4.73 \\ -4.73 & 5.26 \end{bmatrix}\begin{bmatrix} 0.2 \\ 0.1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.579 \\ -0.42 \end{bmatrix}$$

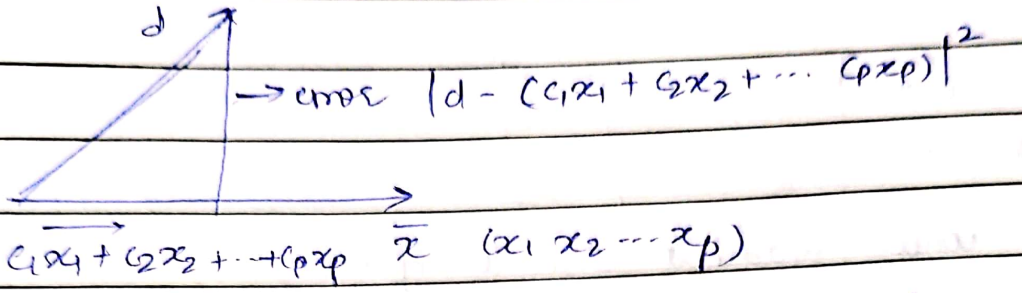


$E^2 = E[e^2(n)]$

Objective: Minimize mean square error.

Let us consider a random variable x whose values are $x_1, x_2 \ldots x_p$. Suppose we want to estimate another random variable who is linear combination of x & it is called linear estimation

$$d = c_1 x_1 + c_2 x_2 + \cdots c_p x_p$$



if weights are optimal then error is minimum or is orthogonal to $\bar{x}$