CSE 465 / L-01 / 19.01.2025 /

Quizzes ⇒ 10%. (Lowest one will be dropped)

Project ⇒ 30%. (NN based)

Midterm ⇒ 30%.

Final ⇒ 30%.

Probability Theory

⇒ mathematical framework for representing uncertain statements.

- quantifying uncertainty.

Key concepts of probability theory:

- Experiment : any process or action that results in an outcome
- Sample Space (S):
  - the set of all possible outcomes of an experiment.
- Event (E)
  - A subset of the sample space, representing specific outcomes of interest.
- Probability (P):
  - a numerical value between 0 and 1 that represents the likelihood of an event occurring. $P(E) = \frac{\#\text{ favorable outcome}}{\#\text{ all outcomes}}$

Frequentist Probability:

- probability is defined as the long-run relative frequency of an event occurring in repeated, identical trials.

  object can not be changed

Bayesian Probability:

- Probability is a measure of belief or certainty about an event, based on prior knowledge or evidence.

Random variable:

- variable that holds some random values.

(i) Discrete Random Variable:

- finite set of distinct values.

(ii) Continuous Random Variable:

- infinite set of values over an interval.

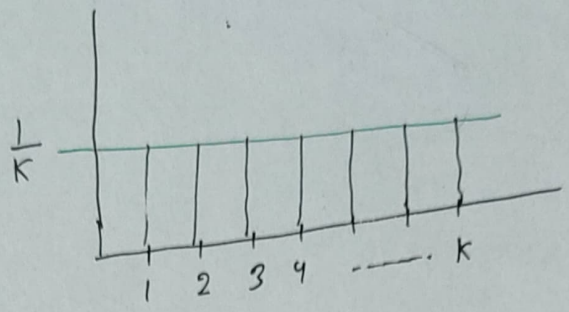
Probability Mass Function (PMF):

- describes the probability distribution of a discrete random variable.

$\Rightarrow \quad 0 \leq P(X = x) \leq 1 \quad | \quad \Rightarrow \sum_x P(X = x) = 1$

(X: random variable; x: ~~random or~~ specific value)

Uniform distribution:



$$P(X = x_i) = \frac{1}{k}$$

$$\sum_i P(X = x_i) = \sum_i \frac{1}{k} = 1$$

Example - 1:

$$P(X = x) = \begin{cases} \frac{1}{4} & ; x = 0, 2 \\ \frac{1}{2} & ; x = 1 \\ 0 & ; \text{otherwise} \end{cases}$$

Example-2:

$$P_Y(1) = P(Y = 1) = P(H) = p \quad (P(H))$$

$$P(Y = 2) = P(TH) = (1-p)\,p \quad (P(T),\ P(H))$$

$$P(Y = 3) = P(TTH) = \frac{(1-p)}{P(T)} \frac{(1-p)}{P(T)} \frac{p}{P(H)} = (1-p)^2 p$$

$$P(Y = k) = P(TTT\ldots TH) = (1-p)^{k-1} p$$

$$\therefore P_Y(y) = \begin{cases} (1-p)^{y-1} p & ; y = 1, 2, 3, \ldots \\ 0 & ; \text{otherwise} \end{cases}$$

Continuous Variable:

- random variable that can take on an infinite number of possible values within a given range.

Probability Density Function: (PDF)

- probability distribution of a continuous random variable.
- probability at a single point is zero.

$\Rightarrow f(x) \geq 0 \quad ; \forall x$

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

$\Rightarrow P(a \leq X \leq b) = \int_{a}^{b} f(x)\,dx$

Conditional Probability:

- probability of an event occurring given that another event has already occurred.

$\Rightarrow P(A|B) = \frac{P(A \cap B)}{P(B)} \quad ; P(B) > 0$

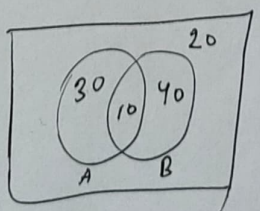
$\Rightarrow$ Given,

$A = 30$

$A \& B = 10$

$B = 40$

Nothing = 20



$P(A) = \frac{40}{100} = 0.4$

$P(B) = \frac{50}{100} = 0.5$

$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.1}{0.5} = 0.2$

Prior belief

L-02 / 21.01.2025

Chain Rule of Conditional Probabilities:

- break down the joint probability of several events.

  joint probability: two or more events occurring simultaneously.

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \text{ ---}$$

Expectation:

- expected value of a random variable.
- average uncertainity.

⇒ for descrete random variable

$$E[X] = \sum_i x_i \cdot P(X = x_i).$$

⇒ for continuous random variable

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

q_1	q_2	q_3
30%	40%	30%
7	5	6

Unweighted average = 6

weighted average = $0.3 \times 7 + 0.4 \times 5 + 0.3 \times 6 = 5.9$

✳ For unfair dice:

x	1	2	3	4	5	6
$P(X=x)$	0·3	0·1	0·15	0·15	0·2	0·1

average value $= (0{\cdot}3\times1) + (0{\cdot}1\times2) + (0{\cdot}15\times3) + (0{\cdot}15\times4) + (0{\cdot}2\times5) + (0{\cdot}1\times6)$

$= 3{\cdot}15$ (Expectation value)

$\therefore E(X) = \sum_x x \cdot P(X=x)$

✳ What about two fair dices?

⇒ sum

S	2	3	4	5	6	7	8	9	10	11	12
$P(S=s)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

$E(S) = \sum_s s \cdot P(S=s)$

$= 7$

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

✳ Variance:

- average distance of random variable from the expected value.

$\mu = E[X]$

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

or,

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

Example- slide-22

Covariance:

- correlation
- how two random variable X and Y vary together.
- determine linear relationship

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

or,

$$Cov(X, Y) = E[XY] - E[X]\,E[Y]$$

Example- Slide-25

- Correlation is not always the causation.

Logistic function:

- keep the value between 0 and 1

$$f(x) = \frac{1}{1 + e^{-x}}$$

⇒ Usecase
- logistic regression.
- activation function in neural network.
- growth models.

ReLV : Rectified Linear Unit

$$f(x) = \max(0, x)$$

- activation function in neural network
  - deep learning model.

SoftPlus Function:

- smooth approximation of the ReLU

$$f(x) = \ln(1 + e^x)$$

- activation function in neural networks.

Baye's rule / Baye's theorem:

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

$$\therefore P(H|E) = \frac{P(H \cap E)}{P(E)}$$

$$P(E|H) = \frac{P(E \cap H)}{P(H)}$$

$$P(H \cap E) = P(E|H) \cdot P(H)$$

$$\therefore P(H|E) = \frac{P(E|H)\,P(H)}{P(E)}$$

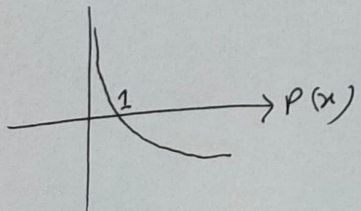
- $P(H|E)$: updating our belief using evidence.
- $P(E|H)$: evidence
- $P(H)$: hypothesis or prior belief

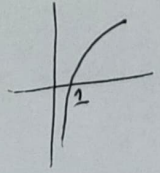
Information Theory

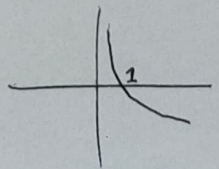
- quantifying how much information is present in a signal.

- unlikely event ⇒ more informative
- likely event ⇒ less informative

$$I(x) = -\log P(x)$$







Unit of information content:

⇒ Bits (Binary Digits)

- logarithm base is 2

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

- most commonly used in information theory

⇒ Nats

- logarithm base is e

$$I(x) = -\ln P(x)$$

- used in mathematical or theoretical contexts.

⇒ Hartleys (or Decimals)

- logarithm base is 10

$$I(x) = -\log_{10} P(x)$$

- named after Ralph Hartley.

Entropy:

- expected information

$$X = x_1 \quad x_2 \quad x_3 \quad \cdots \quad x_n$$

$$P(X=x) = P(x_1) \quad P(x_2) \quad P(x_3) \quad \cdots \quad P(x_n)$$

$$I = -\log P(x_1) \quad -\log P(x_2) \quad \cdots$$

$$E(I) = -P(x_1)\cdot \log P(x_1) - P(x_2)\cdot \log P(x_2) \cdots$$

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

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

- higher entropy ⇒ more uncertainty

Joint Entropy:

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

KL Divergence:

- measures how one probability distribution P diverges from a second probability distribution Q.

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

- Used in machine learning for comparing distributions.

⇒ Properties:

- non negativity
- Asymmetry

$$D_{KL}(P \| Q) \neq D_{KL}(Q \| P)$$

Example-
slide = 37 - 39

L-03 / 26.01.2025 /

Neural Networks

* Matrix Operations are important.

  - addition
  - subtraction
  - multiplication ***
  - scalar multiplication

Slide - 3

* Data represented as matrix and perform many operations.

* Artificial Intelligence > Machine Learning > Deep Learning

* Deep learning:

  - artificial neural networks
    - much similar to human brain.
  - learning process known as representation learning.
  - find out the features on its own and find out the pattern.
  - famous for its performance
  - data hungry.

* DL vs ML:

(i) Data dependency

- for less data, ML works best but for large data DL is best. ML saturates after a certain point. slide->

(ii) Hardware dependency:

- ML works on CPU
- For DL, CPU is not enough, needs GPU and more memory.

(iii) Training time:

- DL models are complex and required more time.
- Prediction time for DL is very fast.

(iv) Feature Selection:

- DL uses representation learning, where features are automatically extracted.
- in ML, we need to choose manually and ~~for~~ need feature engineering for performance enhancement.

(v) Interpretability:

- Entire architecture of DL works as a black box.
  - interpretability is hard not impossible.

(vi) Linear Perceptron:

$\hat{y}(x) = h(x, \theta)$

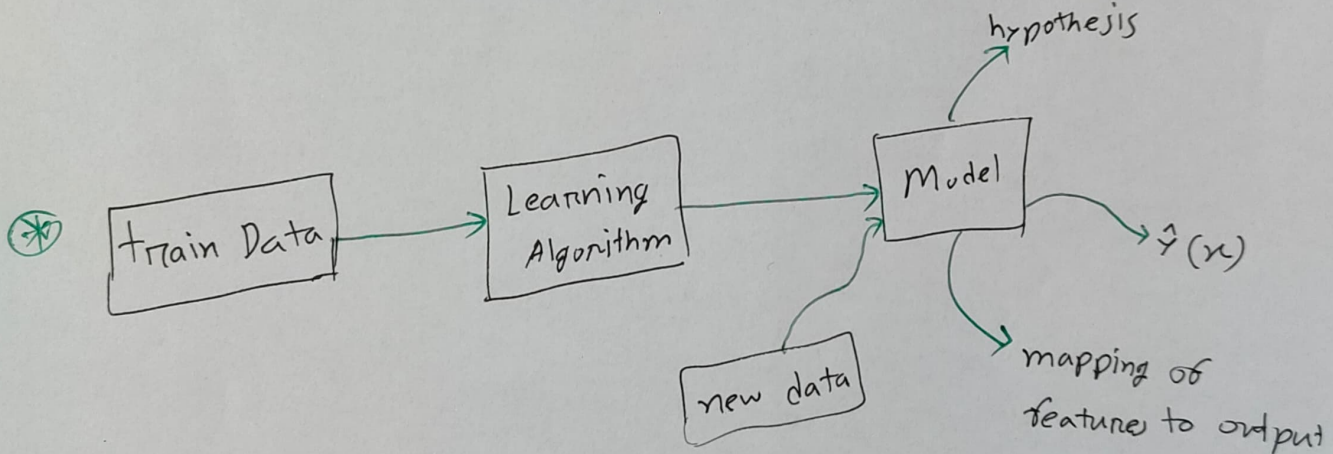
- $\hat{y}$: for predicted value
- h: model
- θ: parameters of the model
- x: inputs, list of features.

$$h(x,\theta) = \begin{cases} -1 & \text{if } x^T \cdot \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \theta_0 < 0 \\ 1 & \text{if } x^T \cdot \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \theta_0 \geq 0 \end{cases}$$

$\theta = [-24, 3, 4]$

$\theta_0 \quad \theta_1 \quad \theta_2$

$$h(x,\theta) = \begin{cases} -1 & \text{if } [x_1, x_2] \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \theta_0 < 0 \\ & \Rightarrow [x_1, x_2] \begin{bmatrix} 3 \\ 4 \end{bmatrix} - 24 < 0 \\ & \Rightarrow 3x_1 + 4x_2 - 24 < 0 \\ & \Rightarrow 3x_1 + 4x_2 < 24 \quad \text{Threshold value} \\ 1 & \text{if } 3x_1 + 4x_2 - 24 \geq 0 \end{cases}$$



Slide - 13

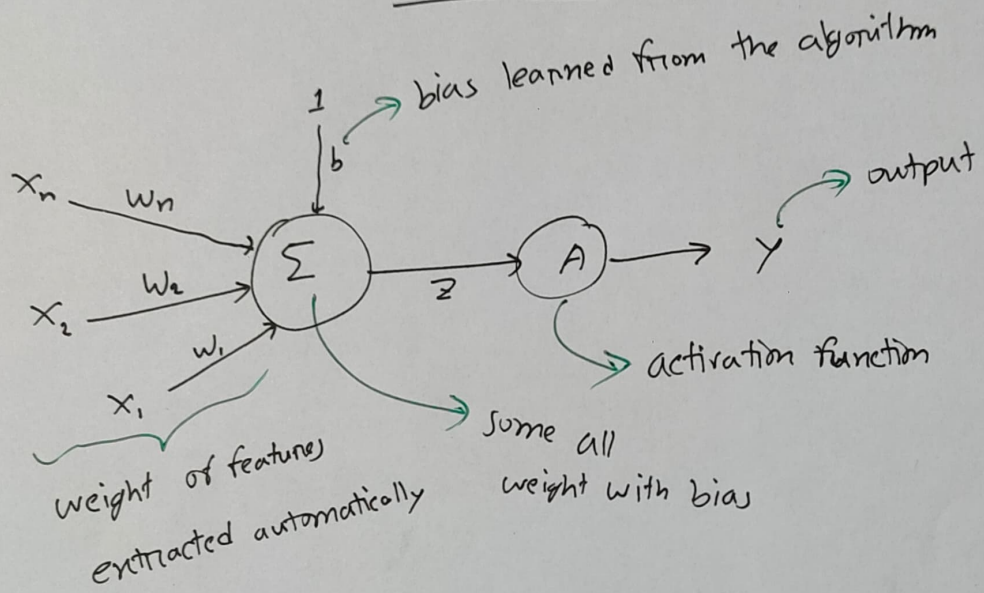
- 2 Features $\Rightarrow$ 2D $\Rightarrow$ Line
- 3 Features $\Rightarrow$ 3D $\Rightarrow$ plane
- 4+ Features $\Rightarrow$ $\geq$ 4D $\Rightarrow$ hyper plane

- Hyper parameter Vs parameter ?
  - parameter → learned value from data
  - hyper parameter of models, needs to set before training.

Neuron:

- ~~fund~~ foundational unit of a human brain
- one neuron make connection with other neurons.

L-04 / 28.01.2025



$$z = \sum_{i=1}^{n} w_i x_i + b$$

Feed forward neural networks:

- neurons in our brain are organized in layers.
- informations flows from one layer to another layer
- There are three type of layer in neural networks.
  - Input layer
    - just pass the raw input data
  - Hidden layers.
    - main black box, magic happens here
    - learn ~~th~~ features and patterns
    - there are multiple hidden layers with

different number of neuron

- Output layer
  - generate prediction.

- number of neuron in input layer and output layer may not be same.

Slide - 18

⊗ Activation functions: introduce non-linearities in the networks

(i) Sigmoid function:

$$\sigma(z) = \frac{1}{1+e^{-z}} \quad ; \quad (0,1)$$

~~- classica~~

- classification problems.

(ii) Tanh Neurons:

$$\tanh(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}} \quad ; \quad (-1,1)$$

(iii) ReLU Neurons:

$$ReLU(z) = \max(0, z) \quad ; \quad [0, \infty)$$

(iv) Softmax Output layers:

$$y_i = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

Example - Slide - 25

Comparison - Slide - 24

* bias are required to move the line. Otherwise every function will go through the ~~cent~~ center.

Example - Slide - 27

* different type of neural networks, depends on activation function.

- MLP : Multy layer perceptron
- CNN
- RNN
- LSTM
- GAN

Quiz-1

upto this

04.02.2025