# NON METRIC SIMILARITY MEASURES

Dr. Umarani Jayaraman
Assistant Professor

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY,
DESIGN AND MANUFACTURING,
KANCHEEPURAM

# Today's Topic

- KL-distance
- Bhattacharya Distance

# Kullback Leibler Distance

- Kullback Leibler distance is a measure of how a probability distribution is different from reference probability distribution.
- It is the natural distance function from a “true” probability distribution $p$, to a target probability distribution $q$.
- Kullback Leibler distance is also called relative entropy.

# KL Distance

- For a discrete probability distribution$(P.D)$,

  if $p = \{p1, p2, \dots., pn\}$ and

  $q = \{q1, q2, \dots., qn\}$,

  then the KL distance is defined as:

  $$D_{KL}(p, q) = \sum p_i \log_2 \frac{p_i}{q_i}$$

- For continuous $P.D$, the sum is replaced by an integral.

# KL Distance: Example

| $X$ | $0$ | $1$ | $2$ |
|---|---|---|---|
| $Distribution\ P(X)$ | $0.36$ | $0.48$ | $0.16$ |
| $Distribution\ Q(X)$ | $0.333$ | $0.333$ | $0.333$ |

The distribution $P(X)$ is a binomial distribution and $Q(X)$ is a uniform distribution

# KL Distance: Is it Metric?

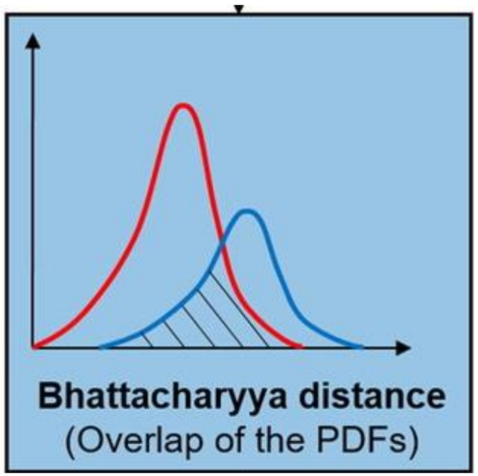
- $D(P,Q) = 0.36 \ln\left(\frac{0.36}{0.333}\right) + 0.48 \ln\left(\frac{0.48}{0.333}\right) + 0.16 \ln\left(\frac{0.16}{0.333}\right)$
  $= 0.0852$
- $D(Q,P) = 0.333 \ln\left(\frac{0.333}{0.36}\right) + 0.48 \ln\left(\frac{0.333}{0.48}\right) + 0.16 \ln\left(\frac{0.333}{0.48}\right)$
  $= 0.0974$
- $D(P,Q) \neq D(Q,P)$ hence KL distance is not a metric.
- It is not symmetric measure and
- does not qualify as a metric distance.

# Bhattacharya Distance

- Introduction
- Example
- Intuition

# Bhattacharya Distance

- It measures the similarity between two probability distributions.
- It is used to compare two normalized histograms.
- Bhattacharyya coefficient is an approximate measurement of two statistical distribution
- The coefficient can be used to determine the relative closeness of the two samples



# Bhattacharya Distance

- Bhattacharyya Coefficient and Distance
  - Both measures are named after Anil Kumar Bhattacharya (professor in ISI Kolkata, 1930)
- The **Mahalanobis distance**
  - It is a measure of the distance between a point P and a distribution D, introduced by PC Mahalanobis (Professor in ISI Kolkata, 1936)

PC Mahalanobis

# Bhattacharya Distance

- It measures the similarity between two probability distributions.
- It is used to compare two normalized histograms.
- Let the two normalized histograms be:

$$x = (x_1, x_2, x_3 \ldots \ldots x_n)$$
$$y = (y_1, y_2, y_3 \ldots \ldots y_n)$$

- Consider two vectors:

$$\left.\begin{array}{l} x' = (\sqrt{x_1}, \sqrt{x_2}, \sqrt{x_3} \ldots \ldots \sqrt{x_n}) \\ y' = (\sqrt{y_1}, \sqrt{y_2}, \sqrt{y_3} \ldots \ldots \sqrt{y_n}) \end{array}\right\} \quad ---- Eqn\ 1$$

# Bhattacharya Distance

- Now, find the dot product of $x'$ and $y'$

  $$x'.\,y' = |x'||y'|\cos\theta \;-----Eqn\;2$$

- Substituting the values from $Eqn\;1$ to $Eqn\;2$,

  $$\sqrt{x_1 y_1} + \sqrt{x_2 y_2} + \ldots\ldots. + \sqrt{x_n y_n} =$$

  $$\sqrt{x_1 + x_2 + \cdots x_n}\;\sqrt{y_1 + y_2 + \cdots y_n} cos\theta$$
  $$-----Eqn\;3$$

- As $x$ and $y$ denotes probability distributions;

  $$x_1 + x_2 + \cdots x_n = 1\; and$$
  $$y_1 + y_2 + \cdots y_n = 1$$

# Bhattacharya Distance

- From the above condition,

  $Eqn$ 3 becomes,

  $$\sqrt{x_1 y_1} + \sqrt{x_2 y_2} + \ldots\ldots. \sqrt{x_n y_n} = 1 \cos\theta$$

  $$\therefore \cos\theta = \sum_{i=1}^{n} \sqrt{x_i y_i}$$

- Bhattacharya Coefficient measures cosine similarity

  $$B(x, y) = \sum_{i=1}^{n} \sqrt{x_i y_i}$$

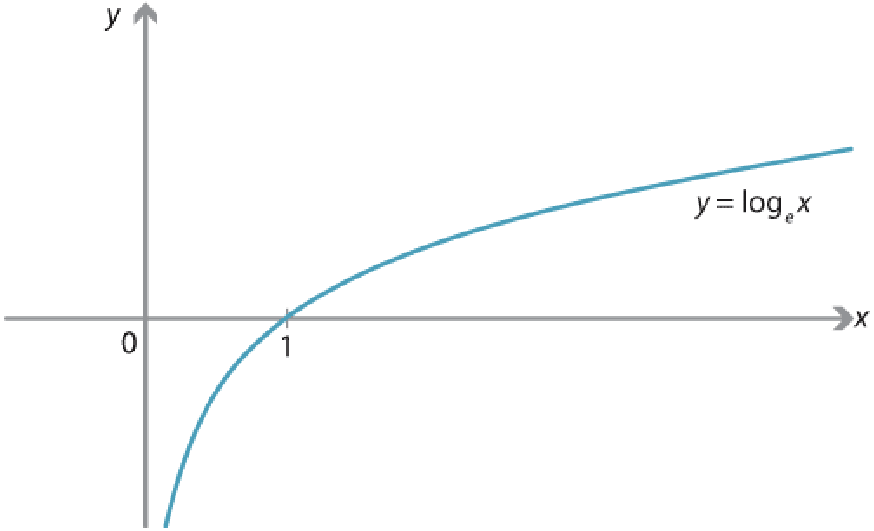
- Bhattacharya Distance

  $$D(x, y) = -\ln B(x, y)$$

# Bhattacharya Distance

- Bhattacharyya coefficient $B(x, y) = \sum_{i=1}^{n} \sqrt{x_i y_i}$
- $0 \leq B(x, y) \geq 1$, because it measures the cosine angle between two vectors that lie in first quadrant
- Bhattacharya Distance

$$D(x, y) = -\ln B(x, y)$$



# Bhattacharya Distance

- Hellinger Distance

$$D(x,y) = 1 - B(x,y)$$

  - Metric? No (cosine distance, 1- $\cos\theta$ )
- Bhattacharya Distance

$$D(x,y) = -\ln B(x,y)$$

  - Metric? No

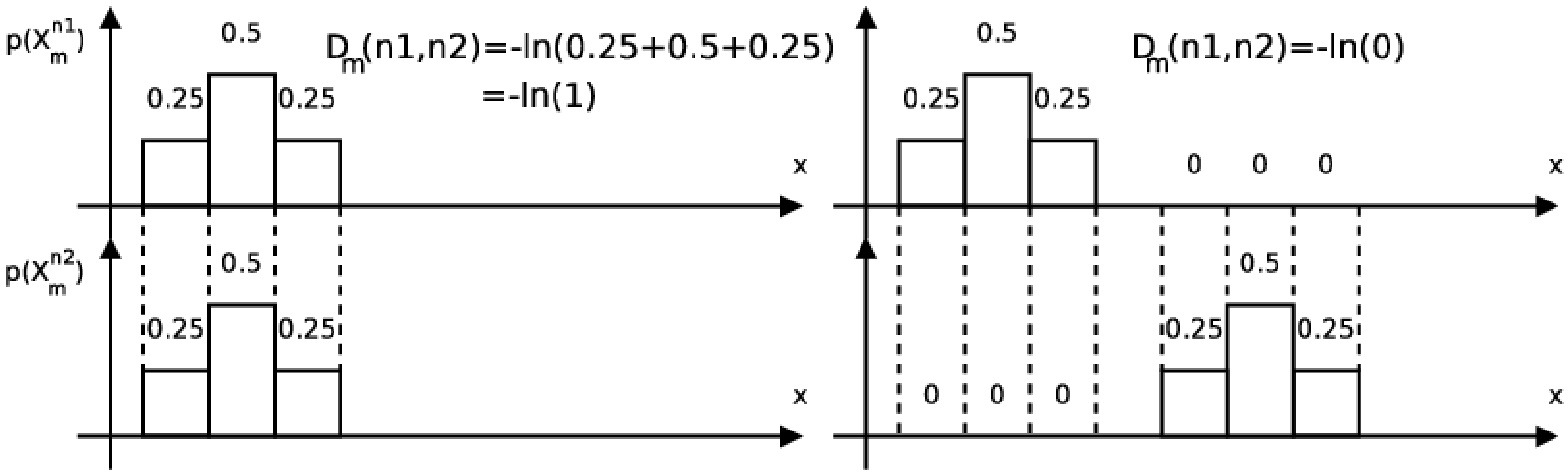
# Bhattacharya Distance: It is Metric?

- Let X = (1,0), Y = (2,2) and Z = (2,1) are three vectors
  - $D(X,Y) = -\ln B(X,Y) = -ln\sum_{i=1}^{n}\sqrt{x_i y_i} = -ln\sqrt{2} = -0.3465$
  - $D(X,Z) = -\ln B(X,Z) = -ln\sum_{i=1}^{n}\sqrt{x_i z_i} = -ln\sqrt{2} = -0.3465$
  - $D(Z,Y) = -\ln B(Z,Y) = -ln\sum_{i=1}^{n}\sqrt{z_i y_i} = -\ln(\sqrt{4} + \sqrt{2}) = -\ln(2 + \sqrt{2}) = -1.2279$

# Bhattacharya Distance: It is Metric?

- Let X = (1,0), Y = (2,2) and Z = (2,1) are three vectors
- $D(X,Y) \leq D(X,Z) + D(Z,Y)$
- $-0.3465 \leq -0.3465 + (-1.2279)$
- $-0.3465 \nleq -1.5744$
- it is not satisfied triangular inequality, hence it is not a metric.

# Intuition



p(X_m^n1)
0.5
0.25
0.25
D_m(n1,n2)=-ln(0.25+0.5+0.25)
=-ln(1)
x
p(X_m^n2)
0.5
0.25
0.25
x
0.5
0.25
0.25
D_m(n1,n2)=-ln(0)
0 0 0
x
0.5
0.25
0.25
0 0 0
x

# Summary

- KL distance with example
- Bhattacharya Distance

# THANK YOU