1/24/2025

Homework - 1

Topics in Data Science

VAISHAK BALACHANDRA
[0037831852]

1) Given two points $a(6,5,4)$ & $b(3,2,1)$. Calculate following measures b/w a & b:
   i) Euclidean distance   ii) Manhattan distance
   iii) Minkowski distance (order of norm, $p = 3$).

i) **Euclidean Distance:**

$$d(x,y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

$(x_1, x_2, x_3) = (6, 5, 4)$
$(y_1, y_2, y_3) = (3, 2, 1)$

$$d_{ab}(a,b) = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2 + (x_3-y_3)^2}$$

$$= \sqrt{(6-3)^2 + (5-2)^2 + (4-1)^2}$$

$$\therefore \boxed{d_{ab}(a,b) = \sqrt{3^2+3^2+3^2} = 3\sqrt{3} = 5.19615}$$ //

ii) **Manhattan Distance:**

$$d(x,y) = \sum_{k=1}^{n} |x_k - y_k|$$

$$d(a,b) = |x_1-y_1| + |x_2-y_2| + |x_3-y_3|$$

$$d(a,b) = |6-3| + |5-2| + |4-1|$$

$$= 3 + 3 + 3$$

$$\boxed{d(a,b) = 9}$$ //

iii) Minkowski Distance (order of norm, $p = 3$)

$$d(x, y) = \left(\sum_{k=1}^{n} |x_k - y_k|^3 \right)^{1/3}$$

$$d(a, b) = \left(|6-3|^3 + |5-2|^3 + |4-1|^3 \right)^{1/3}.$$

$$\boxed{d(a, b) \approx 4.3267}$$ //

---

2) Consider a dataset 2, 4, 6, 8, 8, 10, 12, 14. Express the data in standardized form using:

i) Min-max approach in $(0, 1)$ ii) Z-score

i) Min-max approach in $(0, 1)$:

Given :

$x = [2, 4, 6, 8, 8, 10, 12, 14]$

$\min(x) = 2$

$\max(x) = 14$

$$v' = \frac{v - \min(x)}{\max(x) - \min(x)}$$

ex :

$$v'_{(2)} = \frac{2 - 2}{14 - 2} = 0 \; //$$

$$v'_{(4)} = \frac{4 - 2}{14 - 2} = \frac{2}{12} = \frac{1}{6}$$

$$v'(6) = \frac{6 - 2}{14 - 2} = \frac{4}{12} = \frac{1}{3}$$

$$v'(8) = \frac{8 - 2}{14 - 2} = \frac{6}{12} = \frac{1}{2}$$

$$v'(10) = \frac{10 - 2}{14 - 2} = \frac{8}{12} = \frac{2}{3}$$

$$v'(12) = \frac{12 - 2}{14 - 2} = \frac{10}{12} = \frac{5}{6}$$

$$v'(14) = \frac{14 - 2}{14 - 2} = 1$$

$$\therefore \boxed{x'_{(0,1)} = \left[0, \frac{1}{6}, \frac{1}{3}, \frac{1}{2}, \frac{1}{2}, \frac{2}{3}, \frac{5}{6}, 1 \right]} \; //$$

ii) **Z-Score** : ; $n = 8$

$$v' = \frac{v - \text{mean}(x)}{\text{stand\_dev}(x)} = \frac{v - \mu}{\sigma}$$

Here : $\mu = \frac{\text{Sum of all points}}{\text{Count}} = \frac{2+4+6+8+8+10+12+14}{8}$

$$\boxed{\mu = 8}$$

Here : $\sigma = \sqrt{\frac{1}{n}\sum(x_i - \mu)^2}$

i.e $\sigma = \sqrt{\frac{1}{8}\left[\begin{array}{l}(2-8)^2 + (4-8)^2 + (6-8)^2 + (8-8)^2 + \\ (8-8)^2 + (10-8)^2 + (12-8)^2 + (14-8)^2\end{array}\right]}$

$$\boxed{\sigma = \sqrt{14} = 3.741657}$$ //

$\therefore$ $x = [2, 4, 6, 8, 8, 10, 12, 14]$

$v'(2) = \frac{2-8}{\sqrt{14}} = -1.6035$

$v'(4) = \frac{4-8}{\sqrt{14}} = -1.0690$

$v'(6) = \frac{6-8}{\sqrt{14}} = -0.5345$

$v'(8) = \frac{8-8}{\sqrt{14}} = 0$

$v'(10) = \frac{10-8}{\sqrt{14}} = 0.5345$

$v'(12) = \frac{12-8}{\sqrt{14}} = 1.0690$

$v'(14) = \frac{14-8}{\sqrt{14}} = 1.6035$

$$\boxed{\therefore\ x' = [-1.6035, -1.0690, -0.5345, 0, 0, 0.5345, 1.0690, 1.6035]}$$ //

3) Calculate the entropy of weighted 6-sided dice such that 3 sides of dice have 1/6 chances of facing, two sides of dice have 1/12 chance of facing up & one side has a 1/3 chance of facing up.

WKT: Entropy $= H(x) = -\sum_{i=1}^{n} p_i \log_2 p_i$

Given:
$p_1 = 1/6$; $p_4 = 1/12$
$p_2 = 1/6$; $p_5 = 1/12$
$p_3 = 1/6$; $p_6 = 1/3$

i.e $p = \left[\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{3}\right]$; Here $\sum p = 1$ ✓

$$H(x) = -\left[\left(\frac{1}{6}\log_2\left(\frac{1}{6}\right) \times 3\right) + \left(\frac{1}{12}\log_2\left(\frac{1}{12}\right) \times 2\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right]$$

$$H(x) = -\left[-1.29248 - 0.5974 - 0.5283\right]$$

$$\boxed{H(x) = \text{Entropy} = 2.41818}$$

4) For the dataset given below, find the approx entropy H (passed). This data describes whether students pass or not (Y for yes / N for no), based on their past GPA scores (H for high, A for average, L for low) & whether they prepared or not (Y/N).

CGPA	Prepared	passed
H	N	Y
H	Y	Y
A	N	N
A	Y	Y
L	N	N
L	Y	Y

Success : Passed(Y)

Failure : Failed(N)

$$P(Y) = \frac{4}{6} = \frac{2}{3}$$

$$P(N) = \frac{2}{6} = \frac{1}{3}$$

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

$$H(X) = -\left[\frac{4}{6} \log_2\left(\frac{4}{6}\right) + \frac{2}{6} \log_2\left(\frac{2}{6}\right) \right]$$

$$\boxed{H(X) = 0.91829} \;//$$

$$\therefore \boxed{\text{Entropy of passed variable} = H(X)_{\text{passed}} = 0.91829} \;//$$