Example

	X_1	X_2
A	1	1
B	2	1
C	4	3
D	5	4
E	6	5

- Single linkage
- Euglidean distance

① Compute the distance matrix

	A	B	C	D	E
A	0	1	3.6	5	6.4
B	1	0	2.8	4.2	5.6
C	3.6	2.8	0	1.41	2.8
D	5	4.2	1.41	0	1.41
E	6.4	5.6	2.8	1.41	0

$A \to B = \sqrt{(1-2)^2 + (1-1)^2} = \sqrt{1^2+0} = 1$

$A \to C = \sqrt{(1-4)^2 + (1-3)^2} = \sqrt{3^2+2^2} = 3.6$

$A \to D = \sqrt{(1-5)^2 + (1-4)^2} = \sqrt{16+9} = \sqrt{25} = 5$

$A \to E = \sqrt{(1-6)^2 + (1-5)^2} = \sqrt{5^2+4^2} = \sqrt{39} = 6.4$

$B \to C = \sqrt{(2-4)^2 + (1-3)^2} = \sqrt{4+4} = \sqrt{8} = 2.8$

$B \to D = \sqrt{(2-5)^2 + (1-4)^2} = \sqrt{9+9} = \sqrt{18} = 4.2$

$B \to E = \sqrt{(2-6)^2 + (1-5)^2} = \sqrt{16+16} = \sqrt{32} = 5.6$

$C \to D = \sqrt{(4-5)^2 + (3-4)^2} = \sqrt{1+1} = \sqrt{2} = 1.41$

$C \to E = \sqrt{(4-6)^2 + (3-5)^2} = \sqrt{4+4} = \sqrt{8} = 2.8$

$D \to E = \sqrt{(5-6)^2 + (4-5)^2} = \sqrt{1+1} = \sqrt{2} = 1.41$

② Merge A & B → # we select the columns with the least distance because they are the most related

③ New clusters are ; (AB) , C , D , E

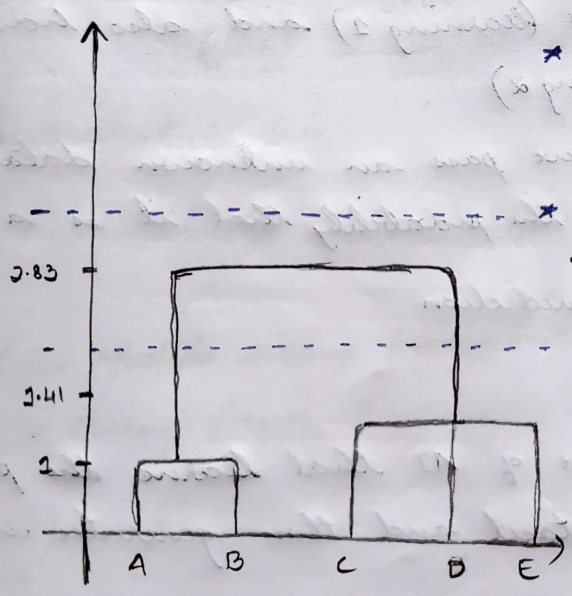
↳ # Repeat the same for new clusters

	AB	C	D	E
AB	0	2.8	4.2	5.6
C	2.8	0	(1.41)	2.2
D	4.2	1.41	0	(1.41) min (select DE)
E	5.6	2.8	1.41	0

$AB \rightarrow C = \min[A \rightarrow C , B \rightarrow C] = \min[3.6, 2.8] = 2.8$

$AB \rightarrow D = \min[A \rightarrow D , B \rightarrow D] = \min[5 , 4.2] = 4.2$

$AB \rightarrow E = \min[A \rightarrow E , B \rightarrow E] = \min[6.4 , 5.6] = 5.6$



* In 2nd iteration (DC) & E will be clustered with same score. (1.41)

* In 3rd iteration (CDE) & (AB) will be clustered.

* No. of clusters will be based on the score we select.

⇒ After iteration - 2

	AB	C	DE
AB	0	2.8	4.2
C	2.8	0	(1.41)
DE	4.2	(1.41)	0

9) Accuracy of test = 99%. Tested (+) rare disease = $\frac{1}{10,000}$

$$P(D=1 \mid T=1) = \frac{P(T=1 \mid D=1)\, P(D=1)}{P(T=1 \mid D=0)\, P(D=0) + P(T=1 \mid D=1)\, P(D=2)}$$

$$= \frac{(0.99)\,(10^{-4})}{0.01 \times \left(1 - \frac{1}{10,000}\right) + (0.99)(10^{-4})}$$

$$= \frac{0.000099}{0.009999 + 0.000099}$$

$$= 0.0098$$

$$P(D \mid T) = 0.0098$$

⇒

- Discriminative Models

$$P(y \mid x)$$

- generative Models

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

⇒ Prediction Models using generative Models

* Using Bayes Theorem

$$\underset{\text{posterior}}{P(y \mid x)} = \frac{\overset{\text{likelihood}}{P(x \mid y)}\ \overset{\text{class prior}}{P(y)}}{P(x \mid y=0)\, P(y=0) + P(x \mid y=1)\, P(y=1)}$$

$$\arg\max_{y} \; P(x \mid y)\, P(y)$$

x_1	x_2	x_3	y
0	0	0	1
0	0	1	1
			0
			0
1	1	1	0

* We will consider all possible combinations of the feature values [Eg; 000 → 1 , 001 → 1] and assign a label to each of the combinations

x_1	x_2	x_3	y
1	0	1	Spam
1	1	0	S
1	1	0	S
0	0	1	NS
0	0	0	NS
1	1	1	NS
0	0	0	S
1	0	0	S
0	1	0	S
1	0	0	S
1	0	0	NS

⟹

x_1	x_2	x_3	$P(y=s)$	$P(y=ns)$
0	0	0	0	1/4
0	1	0	0	1/4
0	0	1	2/3	0
0	1	1	2/3	0
1	0	0	2/3	0
1	0	1	2/3	0
1	1	0	2/3	1/4
1	1	1	0	1/4

argmax P

⟹

x_1	x_2	x_3	Prob

eg:-

x	y
1) Free win Now	S
2) Win a prize	S
3) Hello, How are you	NS
4) Lets win it	NS
5) Free lunch today.	NS

S:- spam
NS:- Not-spam.

Test msg = "Free win"

	x_1 = Free	x_2 = win	y
1	yes	yes	spam
2	No	yes	spam
3	No	No	Not-spam
4	No	yes	Not-spam
5	yes	No	Not-spam

① Calculate prior-probability

$P_y = P(y=1) = 2/5$ $\quad P(y=0) = 3/5$

② Compute the likelyhood :-

for spam class :- Free appears 1 of 2 spam class

$P(\text{free} = \text{yes} \mid \text{spam} = 1) = 1/2 = 0.5$

$P(\text{win} = \text{yes} \mid \text{spam} = 1) = 2/2 = 1$

for not-spam class :- Free appears 1 of 3 non-spam

$P(\text{free} = \text{yes} \mid \text{spam} = 0) = 1/3$

$P(\text{win} = \text{yes} \mid \text{spam} = 0) = 1/3$

③ Inference phase :-

$P(\text{spam} \mid x) = P(x \mid \text{spam})\ P(\text{spam})$ # assume denominal is almost 1

$P(x \mid \text{spam}) = P(x_1, x_2 \mid \text{spam})$

$= P(x_1 \mid \text{spam}) \cdot P(x_2 \mid \text{spam})$

$\therefore P(\text{spam}/X) = P(X_1/\text{spam})\, P(X_2/\text{spam})\, P(\text{spam})$

$= P(\text{free}=\text{yes}/\text{spam})\, P(\text{win}=\text{yes}/\text{spam})\, P(\text{spam})$

$= \frac{1}{2} * 1 \times \frac{2}{5}$

$= \frac{1}{5} = 0.2$

$P(\text{not spam}/X) = P(X/\text{not spam})\, P(\text{not spam})$

$= \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{3}{5} = \frac{1}{5} = 0.066$

$\therefore$ This implies test msg is spam.