24-S1-Q4

Q (a) A 0 2 6

B 3 9

C 11

HAC (i) single (ii) complete

(b) K-Means

$k = 2$

5 7 10 12

$C_1 = 3$ $C_2 = 13$

(i) assign

(ii) centroids.

(c) (i) PCA

(ii) first principal

(iii) PCA LR.

(d) Naive

(i) $P(k=1 \mid a=1, b=1, c=0)$

(ii) $P(k=0 \mid a=1, b=1)$

Solution (a) (i) Single Linkage

① we denote that $d(i,j)$ is the distant between i and j.

② $d(A,B) = \min\{d_{0,3}, d_{0,9}, d_{2,3}, d_{2,9}, d_{6,3}, d_{6,9}\}$

$= \min\{3, 9, 1, 7, 3, 3\}$

$= 1$

$d(A,C) = \min\{d_{0,11}, d_{2,11}, d_{6,11}\}$

$= \min\{11, 9, 5\}$

$= 5$

$d(B,C) = \min\{d_{3,11}, d_{9,11}\}$

$= \min\{8, 2\}$

$= 2$

	A	B	C
A	0		
B	1	0	
C	5	2	0

③ $1 < 2 < 5$

So, the cluster A and cluster B will be merged at the next iteration.

(2) Complete Linkage

① we denote that $d(i,j)$ is the distant between i and j.

② $d(A,B) = \max\{d_{0,3}, d_{0,9}, d_{2,3}, d_{2,9}, d_{6,3}, d_{6,9}\}$

$= \max\{3, 9, 1, 7, 3, 3\}$

$= 9$

$d(A,C) = \max\{d_{0,11}, d_{2,11}, d_{6,11}\}$

$= \max\{11, 9, 5\}$

$= 11$

$d(B,C) = \max\{d_{3,11}, d_{9,11}\}$

$= \max\{8, 2\}$

$= 8$

	A	B	C
A	0		
B	9	0	
C	11	8	0

③ $8 < 9 < 11$

So, the cluster B and cluster C will be merged at the next iteration.

(b) (i) we denote distant between samples and clusters centroids as $d(i,j)$, $i \in \{5,7,10,12\}$, $j \in (1,2)$

samples / C	5	7	10	12
$C_1 = 3$	2	4	7	9
$C_2 = 13$	8	6	3	1
Assign	C_1	C_1	C_2	C_2

So we assign $\{5, 7\}$ to C_1

and assign $\{10, 12\}$ to C_2

(2) new centroids

① C_1 : $\frac{5+7}{2} = 6$

② C_2 : $\frac{10+12}{2} = 11$

(c)(i) ① object: The mean squre error (MSE) of the projected data is minimized and the variance of the projected data is maximized.

② To learn the best low-dimensional subspace for data projection

③ achieve: Apply SVD to the data matrix. Select the unit weight vector V_k from the V matrix and calculate $(V_k^T x_i) V_k$

(ii) PC1 is the eigenvector with the largest eigon value; it defines the direction along which the data show maximum variance. All subsequent PCs are orthogonal to PC1 and to each other, each explaing the greatest remaining variance

(iii) ① Learning target: PCA is unsupervised Linear regression is supervised.

② Optimisation criterion: PCA max variance and min MSE; regression min residual variance of the responce

(d) (i)

① $N_{k=1} = 4$, $N_{k=0} = 4$

$p(k=1) = p(k=0) = 0.5$

$p(a=1 \mid k=1) = 0.5$

$p(a=1 \mid k=0) = 0.5$

$p(b=1 \mid k=1) = 0.25$

$p(b=1 \mid k=0) = 0.5$

$p(c=0 \mid k=1) = 0.5$

$p(c=0 \mid k=0) = 0.25$

② $p(a=1, b=1, c=0)$

$= p(a=1, b=1, c=0 \mid k=0)\, p(k=0) + p(a=1, b=1, c=0 \mid k=1)\, p(k=1)$

$= p(a=1 \mid k=0)\, p(b=1 \mid k=0)\, p(c=0 \mid k=0)\, p(k=0)$

$+ p(a=1 \mid k=1)\, p(b=1 \mid k=1)\, p(c=0 \mid k=1)\, p(k=1)$

$= \frac{2}{4} \times \frac{2}{4} \times \frac{1}{4} \times \frac{4}{8} + \frac{2}{4} \times \frac{1}{4} \times \frac{2}{4} \times \frac{4}{8}$

$= \frac{1}{32} + \frac{1}{32}$

$= 0.0625$

③ $p(k=1 \mid a=1, b=1, c=0)$

$$= \frac{p(a=1, b=1, c=0 \mid k=1)\, p(k=1)}{p(a=1, b=1, c=0)}$$

$$= \frac{p(a=1 \mid k=1)\, p(b=1 \mid k=1)\, p(c=0 \mid k=1)\, p(k=1)}{p(a=1, b=1, c=0)}$$

$$= \frac{0.5 \times 0.25 \times 0.5 \times 0.5}{0.0625}$$

$$= 0.5$$

(ii) ① $p(a=1, b=1)$

$= p(a=1, b=1 \mid k=0)\, p(k=0) + p(a=1, b=1 \mid k=1)\, p(k=1)$

$= p(a=1 \mid k=0)\, p(b=1 \mid k=0)\, p(k=0) + p(a=1 \mid k=1)\, p(b=1 \mid k=1)\, p(k=1)$

$= \frac{2}{4} \times \frac{2}{4} \times \frac{4}{8} + \frac{2}{4} \times \frac{1}{4} \times \frac{4}{8}$

$= \frac{1}{8} + \frac{1}{16}$

$= \frac{3}{16}$

$= 0.1875$

② $p(k=0 \mid a=1, b=1)$

$$= \frac{p(a=1, b=1 \mid k=0)\, p(k=0)}{p(a=1, b=1)}$$

$$= \frac{p(a=1 \mid k=0) \times p(b=1 \mid k=0)\, p(k=0)}{p(a=1, b=1)}$$

$$= \frac{0.5 \times 0.5 \times 0.5}{0.1875}$$

$$= \frac{2}{3}$$

$$= 0.6667$$