M1522.001400 Introduction to Data Mining

Solution of Final

Question 1.

A: 23

B: 1

C: 17

D: 17

Question 2.

(a) {butter} {diaper} {butter,diaper}

(b)

{butter}: No

{diaper}: No

{butter, diaper} Yes

(c)

- butter -> diaper

Confidence: 1, interest: 1/4

- diaper -> butter:

Confidence: 1, interest: 1/4

(d) {k, m, c}

(e) Yes

Question 3.

Consider a 2-D dataset with m points at (-1,0), m points at (0,0), and one point at (0,d) for some d >> 1 (e.g. d = 100) and m >> n * d^2. The optimal clustering with centers (-1,0) and (0,0) has cost d^2, but this initialization method picks (0,d) as one of the centers, and hence, the cost is at least n * d^2.

Question 4.

(a) Yes: perfect matching = (1, c), (2, d), (3, b), (4 ,a)

(b) Answer: 4

Reason: After round 3, A3 and A4 served 11 queries, respectively. At round 4, A3 and A4 receive additional 6 queries, and thus some of the queries are not served at that point.

Question 5.

(a) Yes

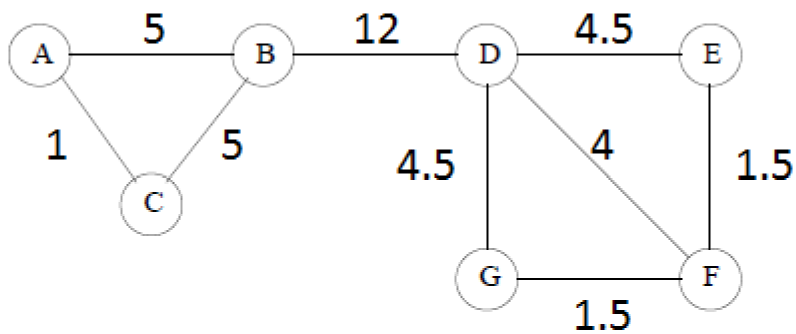
(b) No

(c) 1.1 * 0.3 + 2.1 * 0.5 + 0.3 * 0.6 = 1.56

(d) The user is drawn toward the (0,0) point.

Question 6.

(a)



(b) -1/2

(c)

Step 1) We first show that $cut(\boldsymbol{P_1}, \boldsymbol{P_2})$=$\frac{1}{(1+b)^2}\boldsymbol{z^T L z}$**.** Since $\boldsymbol{y} = \frac{1}{(1+b)}(\boldsymbol{z} - (1-b)\mathbf{1})$**,** and $cut(\boldsymbol{P_1}, \boldsymbol{P_2}) = \frac{1}{4}\boldsymbol{y}^T\boldsymbol{L}\boldsymbol{y}$, it is represented as follows: $cut(\boldsymbol{P_1}, \boldsymbol{P_2}) = \frac{1}{4}\boldsymbol{y}^T\boldsymbol{L}\boldsymbol{y} = \frac{1}{4(1+b)^2}(\boldsymbol{z} - (1-b)\mathbf{1})^{\boldsymbol{T}}\boldsymbol{L}(\boldsymbol{z} - (1-b)\mathbf{1})$.

Since $(\boldsymbol{x} + c\mathbf{1})^{\boldsymbol{T}}\boldsymbol{L}(\boldsymbol{x} + c\mathbf{1}) = \boldsymbol{x^T L x}$ for any $\boldsymbol{x}$ and c by the property of $\boldsymbol{L}$**,** the above is rewritten as follows:

$$\frac{1}{4(1+b)^2}(\boldsymbol{z} - (1-b)\mathbf{1})^{\boldsymbol{T}}\boldsymbol{L}(\boldsymbol{z} - (1-b)\mathbf{1}) = \frac{1}{4(1+b)^2}\boldsymbol{z^T L z}.$$

Step 2) It is straightforward to show that $\frac{1}{assoc(\boldsymbol{P_1},\boldsymbol{V})} + \frac{1}{assoc(\boldsymbol{P_2},\boldsymbol{V})} = \frac{1+b}{b}\frac{1}{\sum_{i\in \boldsymbol{P_2}}\boldsymbol{D_i}}$.

Step 3) We combine the results from step 1) and step 2). Then, $ncut(P_1, P_2)$ is represented as follows:

$$ncut(P_1, P_2) = \frac{1}{4(1+b)^2}\boldsymbol{z^T L z} \times \frac{1+b}{b}\frac{1}{\sum_{i\in \boldsymbol{P_2}}\boldsymbol{D_i}} = \frac{1}{4b}\frac{1}{(1+b)\sum_{i\in \boldsymbol{P_2}}\boldsymbol{D_i}}\boldsymbol{z^T L z}\,.$$

Let's focus on the term $(1+b)\sum_{i\in \boldsymbol{P_2}}\boldsymbol{D_i}$**.** By the definition of b**,** it is represented as follows**:**

$$(1+b)\sum_{i\in \boldsymbol{P_2}}\boldsymbol{D_i} = \sum_{i\in \boldsymbol{P_2}}\boldsymbol{D_i} + b\sum_{i\in \boldsymbol{P_2}}\boldsymbol{D_i} = \sum_{i\in \boldsymbol{P_2}}\boldsymbol{D_i} + \sum_{i\in \boldsymbol{P_1}}\boldsymbol{D_i} = \mathbf{1}^{\boldsymbol{T}}\boldsymbol{D}\mathbf{1}.$$

Note that $\mathbf{1}^{\boldsymbol{T}}\boldsymbol{D}\mathbf{1}$ is the sum of degrees of all nodes. Hence, according to the hint, the last result of step 3) is as follows:

$$\frac{1}{4b}\frac{1}{(1+b)\sum_{i\in \boldsymbol{P_2}}\boldsymbol{D_i}}\boldsymbol{z^T L z} = \frac{\boldsymbol{z^T L z}}{4b\mathbf{1}^{\boldsymbol{T}}\boldsymbol{D}\mathbf{1}} = \frac{\boldsymbol{z^T L z}}{\boldsymbol{z^T D z}}.$$

Therefore, the objective function for Normalized-cut problem based on only $\boldsymbol{D}, \boldsymbol{L}$ and $\boldsymbol{z}$ is as follows**:**

$$\underset{(P_1,P_2)}{\operatorname{argmin}}\, ncut(P_1, P_2) = \underset{\boldsymbol{z}}{\operatorname{argmin}}\frac{\boldsymbol{z^T L z}}{\boldsymbol{z^T D z}} \quad .$$

Question 7.

(a) 12,012,012 bytes

(Note that for $\mathrm{A} = \mathrm{U}\Sigma\mathrm{V}^{\mathrm{T}}$, the second matrix requires only 3 numbers (=12 bytes)).

(b) 4,004,400 bytes

(C requires 4 Mega bytes; R requires 4k bytes; U requires 400 bytes)