داده کاوی
استاد ناظرزاده
#HW1

تینا غلامی ۹۵۳۱۳۰۷

[redacted]

سؤال ۱)

A) supervised:

- نام روشی در یادگیری ماشین است که mapping ورودی به خروجی را برحسب دانش پیشین یا مثال‌هایی از جفت ورودی - خروجی های از قبل تعیین شده، به دست می آورد. داده های نهایی هم برچسب (label) زده می‌شوند. یعنی تابع این mapping را از روی داده‌های آموزشی (training data) به دست می آورد (یاد می‌گیرد).

B) Unsupervised:

- در این روش یادگیری، به دنبال پترن‌های ناشناخته در دیتاها هستیم به طوری که هیچ داده‌ی label زده شده‌ای نداریم. و نیز مینیمم حضور انسان در این روش موجود است. (برخلاف روش supervised که ما خودمان به کامپیوتر، داده‌ی label زده شده می‌دهیم.) پس در اینجا با داده‌های غیر برچسب زده شده (unlabeled) سروکار داریم.

C) Semi Supervised:

- در این روش یادگیری، مقدار کمی از labeled data را به همراه مقدار زیادی از unlabeled data در پروسه‌ی یادگیری و آموزش (training) استفاده می‌کند. و در واقع روشی است که بین دو روش بالایی قرار می‌گیرد.

D) Outlier:

Outlier در واقع دیتایی است که در یک فاصله‌ی دوری از بقیه‌ی دیتاها قرار دارد. و در واقع یک جور abnormality است. (در واقع همان دیتاهای پرت هستند که یا فاصله (distance) و یا کم شباهتی (similarity) قابل توجهی با بقیه‌ی داده‌ها دارند.) (همان بحث Anomaly detection در دیتا پیش می‌آید.)

E) Missing Value, Also Missing Value V.S. Noise:

Missing value یا Missing data وقتی پیش می‌آید که برای یک متغیر در یک مشاهده، مقداری ذخیره نشده باشد. در واقع آن داده را نداریم که

ولی Noise یا Noisy data به داده‌ای گویند که حضور دارد ولی درست نیست و در آن تغییراتی به وجود آمده است. مثلاً خراب شده است (corrupted data) و یا کیفیتش پایین آمده است (distorted data). و یا در بحث سیستم‌های مخابراتی و سیگنال‌ها snr پایین است (signal to noise ratio). چون در واقع: Data = Signell + Noise

سؤال ۲)

به دلیل حجیم بودن اطلاعات در دنیای امروز و نیز عدم دسترسی به تمام مقادیر متغیرها، Missing data در داده‌ها زیاد اتفاق می‌افتد. در نتیجه برای حل این مشکل چندین کار را می‌توان انجام داد:

① در نظر نگرفتن آن داده‌ها (ignore) و یا (unimportant): بیشتر زمانی این کار را می‌کنیم که یک کلاس کامل را نداریم و یا feature های زیادی miss شده‌اند.

② خودمان آن داده‌ها را پر می‌کنیم ← manualy

③ با استفاده از میانگین یکی از دیگر داده‌ها، آن داده‌ها را پر می‌کنیم. (attribute mean): که این میانگین یا متعلق به همان کلاس است و یا میانگین کل داده‌هاست

④ با استفاده از الگوریتم‌های پیشرفته‌تر در یادگیری ماشین و یا در داده کاوی، آن داده‌ها را پر می‌کنیم و یا آن‌ها را پیش‌بینی (predict) می‌کنیم.

سؤال ۳)

min sup = ۶۰٪ $\xRightarrow{\times ۵}$ minsup = ۳

min conf = ۸۰٪

Ⅰ : Apriori Algorithm

C_1

Items	support
{M}	۳
{O}	۳
~~{N}~~	~~۲~~
{K}	۵
{E}	۴
{Y}	۳
~~{D}~~	~~۱~~
~~{A}~~	~~۱~~
~~{U}~~	~~۱~~
~~{C}~~	~~۲~~
~~{I}~~	~~۱~~

L_1

Items	support
{M}	۳
{O}	۳
{K}	۵
{E}	۴
{Y}	۳

C_2

Items	Support
~~{M,O}~~	~~۱~~
{M,K}	۳
~~{M,E}~~	~~۲~~
~~{M,Y}~~	~~۲~~
{O,K}	۳
{O,E}	۳
~~{O,Y}~~	~~۲~~
{K,E}	۴
{K,Y}	۳
~~{E,Y}~~	~~۲~~

L_2

Items	support
{M,K}	۳
{O,K}	۳
{O,E}	۳
{K,E}	۴
{K,Y}	۳

(ادامه ی سوال ۳)

$C_۳$

Items	support
~~{M,K,O}~~	~~۱~~
~~{M,K,E}~~	~~۲~~
~~{M,K,Y}~~	~~۲~~
{O,K,E}	۳
~~{O,K,Y}~~	~~۲~~
~~{K,E,Y}~~	~~۲~~

$L_۳$

Items	support
{O,K,E}	۳

تنها یک frequent Itemset داریم: {O,K,E} ⇐ دیگر ⇐ 4_Items نمی‌توانیم بسازیم.

⇒ Association Rules:

$O \to \{K,E\} \Rightarrow C(O \to \{K,E\}) = \frac{۳}{۴} = ۷۵\%$

$K \to \{O,E\} \Rightarrow C(K \to \{O,E\}) = \frac{۳}{۵} = ۶۰\%$

$E \to \{O,K\} \Rightarrow C(E \to \{O,K\}) = \frac{۳}{۴} = ۷۵\%$

$\{O,K\} \to E \Rightarrow C(\{O,K\} \to E) = \frac{۳}{۳} = ۱۰۰\%$ ✓

$\{O,E\} \to K \Rightarrow C(\{O,E\} \to K) = \frac{۳}{۳} = ۱۰۰\%$ ✓

$\{K,E\} \to O \Rightarrow C(\{K,E\} \to O) = \frac{۳}{۴} = ۷۵\%$

⇒ Strong Association Rules:

$\{O,K\} \to E$

$\{O,E\} \to K$

FP-Growth Algorithm (۱۱)

Items	support
M	۳
O	۳
~~N~~	~~۲~~
K	۵
E	۴
Y	۳
~~D~~	~~۱~~
~~A~~	~~۱~~
~~U~~	~~۱~~
~~G~~	~~۲~~
~~I~~	~~۱~~

⇒

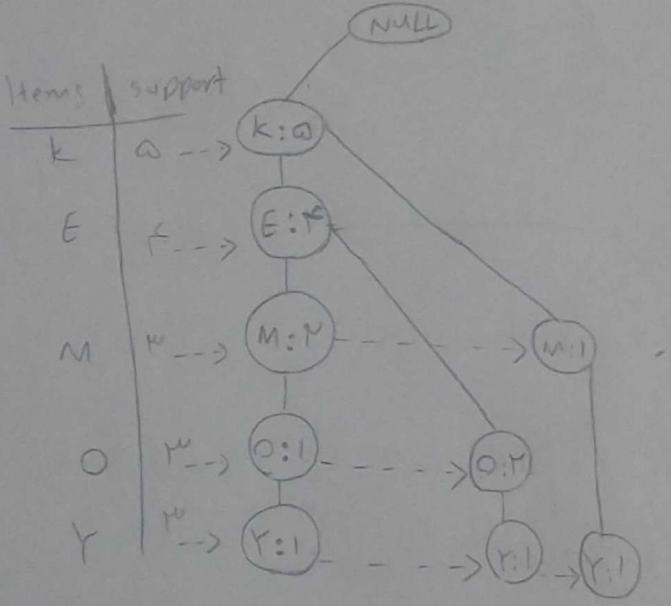
Items	support
K	۵
E	۴
M	۳
O	۳
Y	۳

↑ Frequent Pattern

⇒ ordered Items ها را می‌نویسیم. ⇐

TID	Items_bought	ordered Items	Items
T100	{M,O,N,K,E,Y}	K,E,M,O,Y	Y
T200	{D,O,N,K,E,Y}	K,E,O,Y	O
T300	{M,A,K,E}	K,E,M	M
T400	{M,U,C,K,Y}	K,M,Y	E
T500	{C,O,O,K,I,E}	K,E,O	K

→ ادامه در پایین


NULL
Items
support
K
۵
E
۴
M
۳
O
۳
Y
۳
K:۵
E:۴
M:۲
M:۱
O:۱
O:۲
Y:۱
Y:۱
Y:۱

ادامه جدول

→

conditional Pattern	combined FP-tree	Frequent patterns
{KEMO:1}{KEO:1}{KM:1}	{K:۳}	{K,Y:۳}
{KEM:1}{KE:۲}	{KE:۳}	{O,K:۳}{O,E:۳}{O,K,E:۳}
{KE:۲}{K:1}	{K:۳}	{M,K:۳}
{K:۴}	{K:۴}	{E,K:۴}
—	—	—

سؤال ۴)

از راه Apriori Algorithm :

A) => $d = ۶$ => max Itemsets $= ۲^۶ = ۶۴$

total number of Association Rules $= R = \sum_{k=1}^{d-1}\left[\binom{d}{k} \times \sum_{j=1}^{d-k}\binom{d-k}{j}\right] =$

$۳^d - ۲^{d+1} + ۱ = ۳^۶ - ۲^۷ + ۱ = ۶۰۲$

B)

C_1

Items	support
Milk	۵
Beer	۴
Diapers	۷
Bread	۵
Butter	۵
Cookies	۴

=>

L_1

Items	support
Milk	۵
Beer	۴
Diapers	۷
Bread	۵
Butter	۵
Cookies	۴

=>

$C_۲$

Items	support
{Milk, Beer}	
{Milk, Diapers}	
{Milk, Bread}	
{Milk, Butter}	
{Milk, Cookies}	
{Beer, Diapers}	... ✓
{Beer, Bread}	
{Beer, Butter}	
{Beer, Cookies}	
{Diapers, Bread}	
{Diapers, Butter}	
{Diapers, Cookies}	
{Bread, Butter}	
{Bread, Cookies}	
{Butter, Cookies}	

که با انجام مراحل بالا خواهیم داشت :

بزرگ ترین transaction شامل ۴ تا item است پس ۴ .

C) $\binom{۳}{۶} = ۲۰$

D) {Bread, Butter}

E) (Beer, Cookies)

سؤال ۵)

data Reduction یکی از مراحل مهم در بحث data preprocessing است تا در آن، داده‌ها را حجم کمتری و با صورت ساده‌تر و فشرده‌تری نمایش بدهیم، در حالی که همچنان، همان نتایج قبلی از بررسی تولید شود.

دلیل این کار این است که جاهایی که داده‌ها را آن‌جا نگه داری می‌کنند، (همان database) (و یا datasets warehouse (و یا تقریبا همان)) مقادیر زیادی از داده را در دیتابیس نگه داری می‌کند (در حدود چندین ترابایت.) و عملیاتی که روی داده‌ها انجام می‌دهیم در این صورت بسیار طولانی و وقت‌گیر می‌شود تا یک الگوریتم مثلاً run شود. به خاطر همین از data reduction استفاده می‌کنیم.

استراتژی‌های مورد استفاده شامل:

① کاهش بُعد (dimensiality Reduction): حذف feature های غیرمهم و غیر ضروری

مانند روش‌های:

wavelet Transforms
PCA
Feature Subset Selection
feature Creation (یا feature Generation)

② کاهش مقدار (Numerosity Reduction): و یا به طور ساده‌تر، همان data Reduction در معنای عام‌تر

مانند:

Regression & log-linear models
Histograms, clustering, sampling
Data cube aggregation

③ فشرده سازی (Data compression)

سؤال ۶)

اگر correlation (کوواریانس) دو داده صفر شود، یعنی اگر $\rho(X,Y)=0$، می‌گوییم X، Y uncorrelated اند. و این لزوماً به این معنا نیست که غیروابسته (independant) اند.

ولی اگر دو داده از هم independant باشند، پس حتماً correlation آن دو برابر صفر می‌شود.

در واقع در این جا منظور، مانند correlation، نوع خاصی از آن به نام "Pearson" است و میزان وجود یک رابطه‌ی خطی (a linear relationship) و یا میزان وابستگی خطی آن دو متغیر را می‌سنجد. مثلاً:

$$\rho(X,Y)=1 \Rightarrow Y=aX+b \quad , \quad a=\frac{\sigma_y}{\sigma_x}$$

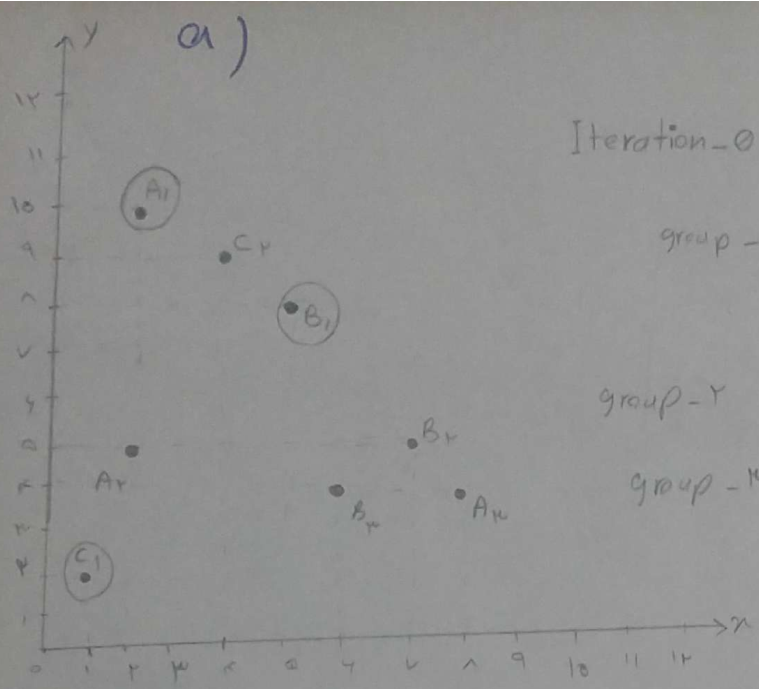
$$\rho(X,Y)=-1 \Rightarrow Y=aX+b \quad , \quad a=-\frac{\sigma_y}{\sigma_x}$$

$\rho(X,Y)=0 \Rightarrow$ X، Y هیچگونه رابطه‌ی خطی ندارند.

$\Longrightarrow$ ممکن است correlation دو داده صفر باشد، این یعنی که رابطه‌ی خطی بین آن دو داده وجود ندارد ولی ممکن است رابطه‌ی غیرخطی (non linear) بین‌شان وجود داشته باشد. پس لزوماً غیروابسته یا همان Independant ممکن است نباشند.

سوال ۷)

a)



Iteration_0 :   centroids = A۱ , B۱ , C۱

$$\begin{array}{ll} \text{group-۱} \leftarrow & \\ & D^{۰} = \begin{bmatrix} \boxed{۰} & ۵ & ۸٫۴۸ & ۳٫۶۰ & ۷٫۰۷ & ۷٫۲۱ & ۸٫۰۶ & ۲٫۲۳ \\ ۳٫۶۰ & ۴٫۲۴ & \boxed{۵} & \boxed{۰} & \boxed{۳٫۶۰} & \boxed{۴٫۱۲} & ۷٫۲۱ & \boxed{۱٫۴۱} \\ ۸٫۰۶ & \boxed{۳٫۱۶} & ۷٫۲۸ & ۷٫۲۱ & ۶٫۷۰ & ۵٫۳۸ & \boxed{۰} & ۷٫۶۱ \end{bmatrix} \\ \text{group-۲} \leftarrow & \\ \text{group-۳} \leftarrow & \end{array}$$

$$\begin{array}{cccccccc} A_۱ & A_۲ & A_۳ & B_۱ & B_۲ & B_۳ & C_۱ & C_۲ \end{array}$$
$$\begin{bmatrix} ۲ & ۲ & ۸ & ۵ & ۷ & ۶ & ۱ & ۴ \\ ۱۰ & ۵ & ۴ & ۸ & ۵ & ۴ & ۲ & ۹ \end{bmatrix}$$

$d(A_۱, A_۱) = ۰$
$d(A_۱, A_۲) = ۵$
$d(A_۱, A_۳) = ۸٫۴۸$
$d(A_۱, B_۱) = ۳٫۶۰$
$d(A_۱, B_۲) = ۷٫۰۷$
$d(A_۱, B_۳) = ۷٫۲۱$
$d(A_۱, C_۱) = ۸٫۰۶$
$d(A_۱, C_۲) = ۲٫۲۳$

$d(B_۱, A_۱) = ۳٫۶۰$
$d(B_۱, A_۲) = ۴٫۲۴$
$d(B_۱, A_۳) = ۵$
$d(B_۱, B_۱) = ۰$
$d(B_۱, B_۲) = ۳٫۶۰$
$d(B_۱, B_۳) = ۴٫۱۲$
$d(B_۱, C_۱) = ۷٫۲۱$
$d(B_۱, C_۲) = ۱٫۴۱$

$d(C_۱, A_۱) = ۸٫۰۶$
$d(C_۱, A_۲) = ۳٫۱۶$
$d(C_۱, A_۳) = ۷٫۲۸$
$d(C_۱, B_۱) = ۷٫۲۱$
$d(C_۱, B_۲) = ۶٫۷۰$
$d(C_۱, B_۳) = ۵٫۳۸$
$d(C_۱, C_۱) = ۰$
$d(C_۱, C_۲) = ۷٫۶۱$

حالا باید در ماتریس $D^{۰}$، در هر ستون، min ها را انتخاب کنیم. (با ✱)

=> object clustering:

$$C^{۰} = \begin{array}{l} \left[\begin{array}{cccccccc} ۱ & ۰ & ۰ & ۰ & ۰ & ۰ & ۰ & ۰ \\ ۰ & ۰ & ۱ & ۱ & ۱ & ۱ & ۰ & ۱ \\ ۰ & ۱ & ۰ & ۰ & ۰ & ۰ & ۱ & ۰ \end{array}\right] \begin{array}{l} \rightarrow \text{group-0} \\ \rightarrow \text{group-1} \\ \rightarrow \text{group-۲} \end{array} \\ \;\;\begin{array}{cccccccc} A_۱ & A_۲ & A_۳ & B_۱ & B_۲ & B_۳ & C_۱ & C_۲ \end{array} \end{array}$$

در ماتریس $C^{۰}$، حضور هر object
در هر cluster را با 1
و عدم حضورش را با 0 نشان می‌دهیم.

=> Iteration_1 :

حالا باید centroid های جدید را تعیین کنیم:

(بدین صورت که در هر group، از مختصات نقاطی که در آن group قرار می‌گیرند طبق ماتریس $C^۰$، میانگین بگیریم.)

$$\text{group-0}: A_۱ = (۲, ۱۰)$$

$$\text{group-1}: \left(\frac{x_{A_۳}+x_{B_۱}+x_{B_۲}+x_{B_۳}+x_{C_۲}}{۵}, \frac{y_{A_۳}+y_{B_۱}+y_{B_۲}+y_{B_۳}+y_{C_۲}}{۵}\right) = (۶, ۶)$$

$$\text{group-2}: \left(\frac{x_{A_۲}+x_{C_۱}}{۲}, \frac{y_{A_۲}+y_{C_۱}}{۲}\right) = \left(\frac{۲+۱}{۲}, \frac{۵+۲}{۲}\right) = (۱,۵ , ۳,۵)$$

پس بعد از اولین مرحله اجرای الگوریتم، مراکز cluster ها بصورت زیر اند:

$\text{group\_0}: (۲, ۱۰) = T_۱$

$\text{group\_1}: (۶, ۶) = T_۲$ ← New Centroids

$\text{group\_2}: (۱,۵ , ۳,۵) = T_۳$

b)

باید الگوریتم را آن قدر ادامه دهیم تا centroid ها دیگر عوض نشوند؛ یعنی ماتریس های C، تغییر نکنند.

پس دوباره مشابه (a) انجام می‌دهیم:

$$D^۱ = \begin{array}{c|cccccccc|l} & A_۱ & A_۲ & A_۳ & B_۱ & B_۲ & B_۳ & C_۱ & C_۲ & \\ \hline & \boxed{۰} & ۵ & ۸,۴۸ & ۳,۶۰ & ۷,۰۷ & ۷,۲۱ & ۸,۰۶ & \boxed{۲,۲۳} & \rightarrow \text{group-0} \\ & ۵,۶۵ & ۴,۱۲ & \boxed{۲,۸۳} & \boxed{۲,۲۳} & \boxed{۱,۴۱} & \boxed{۲} & ۶,۴۰ & ۳,۶۰ & \rightarrow \text{group-1} \\ & ۶,۵۲ & \boxed{۱,۵۸} & ۶,۵۲ & ۵,۷۰ & ۵,۷۰ & ۴,۵۳ & \boxed{۱,۵۸} & ۶,۰۴ & \rightarrow \text{group-2} \end{array}$$

$d(T_۱, A_۱) = ۰$

$d(T_۱, A_۲) = ۵$

$d(T_۱, A_۳) = ۸,۴۸$

$d(T_۱, B_۱) = ۳,۶۰$

$d(T_۱, B_۲) = ۷,۰۷$

$d(T_۱, B_۳) = ۷,۲۱$

$d(T_۱, C_۱) = ۸,۰۶$

$d(T_۱, C_۲) = ۲,۲۳$

$d(T_۲, A_۱) = ۵,۶۵$

$d(T_۲, A_۲) = ۴,۱۲$

$d(T_۲, A_۳) = ۲,۸۳$

$d(T_۲, B_۱) = ۲,۲۳$

$d(T_۲, B_۲) = ۱,۴۱$

$d(T_۲, B_۳) = ۲$

$d(T_۲, C_۱) = ۶,۴۰$

$d(T_۲, C_۲) = ۳,۶۰$

$d(T_۳, A_۱) = ۶,۵۲$

$d(T_۳, A_۲) = ۱,۵۸$

$d(T_۳, A_۳) = ۶,۵۲$

$d(T_۳, B_۱) = ۵,۷۰$

$d(T_۳, B_۲) = ۵,۷۰$

$d(T_۳, B_۳) = ۴,۵۳$

$d(T_۳, C_۱) = ۱,۵۸$

$d(T_۳, C_۲) = ۶,۰۴$

$$\Rightarrow C^{1} = \begin{array}{c|cccccccc|l} & A_1 & A_2 & A_3 & B_1 & B_2 & B_3 & C_1 & C_2 & \\ \hline & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \rightarrow \text{group-0} \\ & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & \rightarrow \text{group-1} \\ & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & \rightarrow \text{group-2} \end{array}$$

<= مرکزهای جدید :

group-0 : $T_1 = (۲,۱۰) = T_۴$

group-1 : $\left(\frac{۸+۵+۷+۶}{۴} , \frac{۴+۸+۵+۴}{۴}\right) = (۶,۵ \;,\; ۵,۲۵) = T_۵$

group-2 : $T_۳ = (۱,۵ \;,\; ۳,۵) = T_۶$

=> Iteration_۲ :

$$D^{۲} = \begin{array}{c|cccccccc|l} & A_1 & A_2 & A_3 & B_1 & B_2 & B_3 & C_1 & C_2 & \\ \hline & \boxed{۰} & ۵ & ۸,۴۸ & ۳,۶۰ & ۷,۰۲ & ۷,۲۱ & ۸,۰۶ & \boxed{۲,۲۳} & \rightarrow \text{group-0} \\ & ۶,۵۴ & ۴,۵۱ & \boxed{۱,۹۵} & \boxed{۳,۱۳} & \boxed{۰,۵۶} & \boxed{۱,۳۵} & ۶,۳۹ & ۴,۰۴ & \rightarrow \text{group-1} \\ & ۶,۵۲ & \boxed{۱,۵۸} & ۶,۵۲ & ۵,۷۰ & ۵,۷۰ & ۴,۵۳ & \boxed{۱,۵۸} & ۶,۰۴ & \rightarrow \text{group-2} \end{array}$$

* چون $\left\{\begin{array}{l} T_1 = T_۴ \\ T_۳ = T_۶ \end{array}\right.$ پس ردیف اول و سوم یعنی $\left\{\begin{array}{l} \text{group-0} \\ \text{و} \\ \text{group-1} \end{array}\right.$ ماتریس $D^۲$ مشابه ماتریس $D^۱$ است

پس فقط کافی است سطر دوم یعنی group-۲ را پیدا کنیم :

$d(T_۵, A_۱) = ۶,۵۴$

$d(T_۵, A_۲) = ۴,۵۱$

$d(T_۵, A_۳) = ۱,۹۵$

$d(T_۵, B_۱) = ۳,۱۳$

$d(T_۵, B_۲) = ۰,۵۶$

$d(T_۵, B_۳) = ۱,۳۵$

$d(T_۵, C_۱) = ۶,۳۹$

$d(T_۵, C_۲) = ۴,۰۴$

$$C^{۲} = \begin{array}{c|cccccccc|l} & A_1 & A_2 & A_3 & B_1 & B_2 & B_3 & C_1 & C_2 & \\ \hline & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \rightarrow \text{group-0} \\ & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & \\ & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & \end{array}$$

=>

پس چون $C^۲ = C^۱$ شد پس یعنی مرکزهای $\left\{\begin{array}{l} T_۴ \\ T_۵ \\ T_۶ \end{array}\right.$ همگرا شدند. پس Iteration در اینجا تمام خواهد شد.

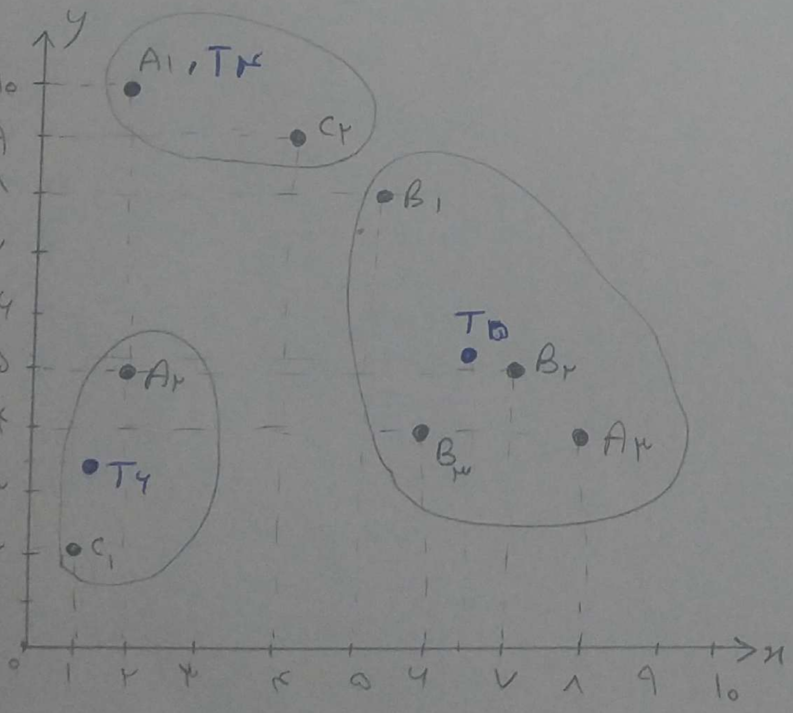
درنتیجه، ۳ cluster نهایی به صورت زیر است :

group_0 :
با مرکز = $T_۴$ = (۲,۱۰)
A۱=(۲,۱۰)
C۲=(۴,۹)

group_1 :
با مرکز = $T_۵$ = (۶,۵ , ۵,۲۵)
A۳=(۸,۴)
B۱=(۵,۸)
B۲=(۷,۵)
B۳=(۶,۴)

group_2 :
با مرکز = $T_۶$ = (۱,۵ , ۳,۵)
A۲=(۲,۵)
C۱=(۱,۲)

که در دستگاه مختصات به صورت زیر در می آید:



سؤال ۸)

برای مثال، A مجموعه‌ای از نقاط است که قرار است آن را خوشه‌بندی کنیم:

A = {۱, ۲, ۳, ۴}

و ما قرار است ۲ خوشه برای آن ارائه کنیم. (۲-means).

حداقل دو حالت برای این کار وجود دارد:

حالت ۱)

centroid1 = ۱ => cluster1 = {۱}

centroid۲ = ۳ => cluster۲ = {۲, ۳, ۴}

الگوریتم در همینجا به پایان می‌رسد. → پس objective = ۲

حالت ۲)

centroid1 = ۱٫۵ => cluster1 = {۱, ۲}

centroid۲ = ۳٫۵ => cluster۲ = {۳, ۴}

→ پس objective = ۱/۴

پس اگر الگوریتم k-means به فرم حالت ۱ مقداردهی اولیه می‌شد، پیش‌بینی کردیم و در همانجا پایان می‌یافت ولی این حالت، global minimum نیست! پس الگوریتم k-means لزوماً نمی‌تواند بهینه‌ حالت global minimum را ایجاد کند.