بخش تئوری

سوال اول

یک مجموعه داده از حیوانات مختلف به همراه ویژگی‌هایشان را در اختیار داریم، می‌خواهیم با استفاده از روش‌های خوشه‌بندی میزان شباهت هر دو حیوان به هم را از ۱ (کمترین) تا ۳ (بیشترین) مشخص نماییم. برای مثال میزان شباهت شیر و پلنگ ۳ و میزان شباهت شیر و گوسفند ۱ می‌تواند باشد. الگوریتمی ارائه دهید که این امر را به صورت غیرنظارت‌شده[1] ممکن سازد.

از اینجایی که این سوال، دارای جواب‌های خلاقانه‌ای می‌باشد، روش‌های خلاقانه‌ای که به درستی عمل کنند نیز، نمره خواهند گرفت. یکی از این روش‌ها، می‌تواند به شکل زیر باشد.

از روش دسته بندی partitioning استفاده میکنیم، یعنی ابتدا فرض میکنیم کل حیوانات یک دسته باشند، سپس این دسته را به نحوی که مجموع فاصله هر داده از مرکز دسته متعلق به آن کمترین شود به دو دسته تقسیم میکنیم. و در نهایت نیز یکی از دو دسته ایجاد شده را به همان نحوه قبلی به دو دسته تقسیم میکنیم. به این صورت حیوانات دسته بندی می شوند و فاصله پارتیشن ها از هم نشان دهنده میزان شباهت می باشد یعنی برای مثال حیوانات متعلق به یک دسته فاصله ۱ دارند، و حیوانات متعلق به دو دسته متفاوت فاصله ۲ یا ۳ میتوانند داشته باشند (با توجه به دسته ها).

سوال دوم

می‌دانیم که در الگوریتم خوشه‌بندی برای تابع مجاورت[2] موارد مختلفی را می‌توان استفاده کرد، در موارد زیر اثبات نمایید که نقطه نهایی که به عنوان مرکز انتخاب می‌شود چه نقطه‌ای است. (در رابطه زیر D مجموعه تمامی نقاط داده و C مجموعه تمامی مراکز خوشه ها می‌باشد.)

$$\sum_{d\in D}\sum_{c\in C} f(d,c)$$

- نرم۱ $(f(d,c)=|d-c|)$

برای حل این روش داریم:

$$SSE=\sum_{d\in D}|d-c_i|\,,\qquad d\in c_i, c_i\in C$$

بنابراین، برای محاسبه مقدار مینیمم این تابع، می‌توانیم با برابر صفر قرار دادن مشتق جزئی این تابع نسبت به مرکز خوشه آن را به دست آوریم.

$$\frac{\partial SSE}{\partial c_i}=\frac{\partial}{\partial c_i}\sum_{d\in D}|d-c_i|=\sum_{d\in D}\frac{\partial}{\partial_{c_i}}|d-c_i|=\sum_{d\in D}\frac{|d-c_i|}{d-c_i}=0$$

[1] Unsupervised
[2] Proximity Function

همانطور که می‌دانیم، عبارت $\frac{|d-c_i|}{d-c_i}$ بسته به علامت $d-c_i$ دارای مقدار $+1$ و یا -1 می‌باشد. بنابراین، این عبارت زمانی برابر صفر خواهد شد که تعداد عبارت‌های برابر با $+1$ با تعداد عبارت‌های -1 برابر باشد. بنابراین، c_i می‌تواند برابر با میانه نقاط در درون خوشه c_i باشد.

- نرم۲ $(f(d,c) = \|d-c\|_2^2)$

برای حل این روش داریم:

$$SSE = \sum_{d \in D} \|d - c_i\|^2, \qquad d \in c_i, c_i \in C$$

بنابراین، برای محاسبه مقدار مینیمم این تابع، می‌توانیم با برابر صفر قرار دادن مشتق جزئی این تابع نسبت به مرکز خوشه آن را به دست آوریم.

$$\frac{\partial SSE}{\partial c_i} = \frac{\partial}{\partial c_i} \sum_{d \in D} \|d - c_i\|^2 = \sum_{d \in D} \frac{\partial}{\partial_{c_i}} \|d - c_i\|^2 = \sum_{d \in D} -2(d - c_i) = 0$$

$$\sum_{d \in D} d - c_i = 0$$

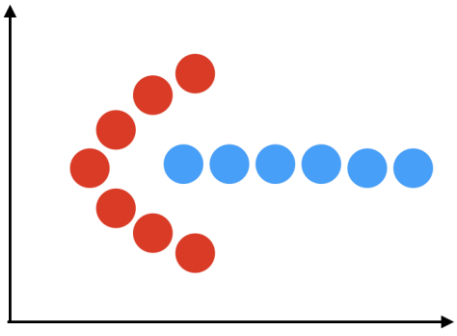
$$c_i N_{c_i} = \sum_{d \in D} d$$

$$c_i = \frac{\sum_{d \in D} d}{N_{c_i}}$$

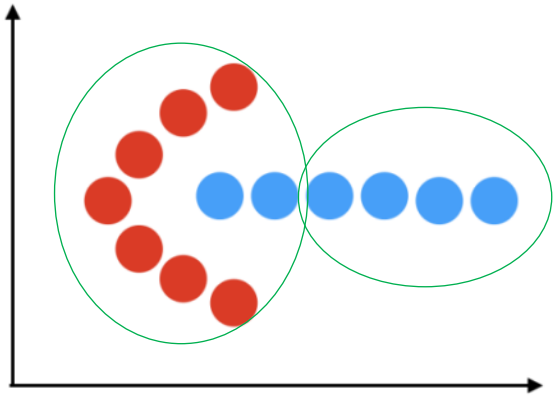
در اینجا، مرکز خوشه برابر میانگین موقعیت داده‌های موجود در کلاستر c_i می‌باشد. (در اینجا N_{c_i} معادل تعداد نقاط درون خوشه c_i می‌باشد.)

سوال سوم

الف) فرض کنید داده‌های زیر را می‌خواهیم به ۲دسته مختلف دسته‌بندی کنیم، پیش‌بینی شما از اجرا الگوریتم k-means را از داده‌های زیر بیان کنید و علت این پیش‌بینی را هم ذکر نمایید.

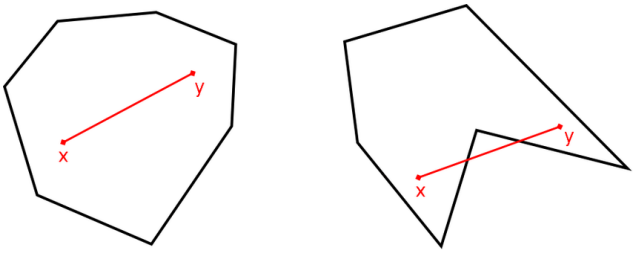


با توجه به اینکه شکل توزیع داده ها non-convex می باشد الگوریتم k-means نمیتواند به درستی خوشه بندی را انجام دهد زیرا در این الگوریتم هدف آپدیت کردن مراکز خوشه به نحوی است که میانگین فاصله داده ها از مرکز خوشه مینییمم شود و در نتیجه خروجی الگوریتم چیزی همانند شکل زیر خواهد بود که بخشی از داده ها با لیبل آبی را به اشتباه با داده های قرمز در یک خوشه قرار میدهد.



ب) آیا استفاده از روش DBSCAN میتواند برای داده‌های بالا عملکرد بهتری داشته باشد؟ علت را توضیح دهید.

بله، به علت اینکه شکل توزیع داده ها non-convex می باشد و الگوریتم DBSCAN خوشه ها را بر اساس تراکم نقاط داده تعریف می کند و به آن اجازه می دهد خوشه هایی با مرزهای نامنظم را شناسایی کند. در مقابل، k-means خوشه‌ها را محدب فرض می‌کند و داده‌ها را بر اساس فواصل اقلیدسی جدا می‌کند، که باعث می‌شود برای خوشه‌های غیر محدب کارایی کمتری داشته باشد.



Convex region Non-convex region

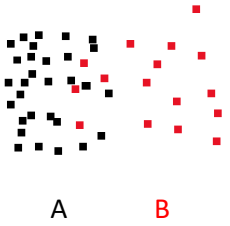
DBSCAN به طور خودکار آستانه چگالی مناسب را برای جداسازی خوشه ها از نویز تعیین می کند. برای تعیین مناطق متراکم، تعداد نقاط مجاور را در یک شعاع مشخص در نظر می گیرد. این ماهیت تطبیقی به DBSCAN اجازه می دهد تا خوشه هایی با چگالی های مختلف، از جمله خوشه های غیر محدب را کشف کند. در مقابل، K-means از کاربر می‌خواهد که تعداد خوشه‌ها را از قبل مشخص کند و تأثیر یکسانی از همه نقاط را در نظر می‌گیرد، که باعث می‌شود در مدیریت داده‌های غیر محدب با چگالی‌های متفاوت انعطاف‌پذیرتر نباشد.

ج) توضیح دهید در چه زمانی خوشه‌بندی بر مبنای چگالی عملکرد مناسبی نخواهد داشت؟ مثال بزنید.

- حساسیت به انتخاب پارامتر چگالی: الگوریتم‌های خوشه‌بندی مبتنی بر چگالی نیازمند تعیین پارامترهایی مانند حداقل تعداد نقاط و شعاع همسایگی هستند. انتخاب مقادیر پارامتر مناسب می تواند در سناریوهایی با خوشه های همپوشانی

چالش برانگیز باشد. اگر پارامتر چگالی خیلی زیاد تنظیم شود، منطقه همپوشانی ممکن است به عنوان یک خوشه جداگانه شناسایی نشود. برعکس، تنظیم پارامتر چگالی خیلی کم ممکن است باعث شود الگوریتم خوشه A و ناحیه همپوشانی را ادغام کند و منجر به نتایج خوشه‌بندی نادرست شود.

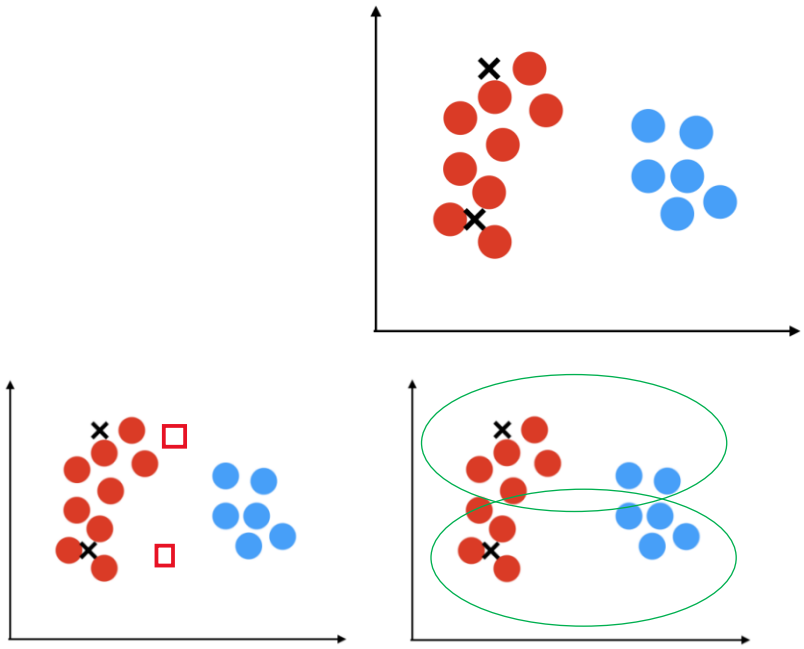
مثال:



- چگالی متغیر: خوشه‌بندی مبتنی بر چگالی فرض می‌کند که خوشه‌ها چگالی مشابهی دارند. هنگامی که چگالی خوشه ها به طور قابل توجهی تغییر می کند، تعیین آستانه چگالی مناسب برای الگوریتم چالش برانگیز می شود. در چنین مواردی، خوشه‌هایی با چگالی کمتر ممکن است با خوشه‌های نزدیک با چگالی بالاتر ادغام شوند یا به عنوان نویز در نظر گرفته شوند که منجر به نتایج نادرست می‌شود.
- داده‌های با ابعاد بالا: الگوریتم‌های خوشه‌بندی مبتنی بر چگالی در فضاهای با ابعاد بالا به دلیل «نفرین ابعاد» با مشکلاتی روبرو هستند. در داده‌های با ابعاد بالا، مفهوم فاصله کمتر معنادار می‌شود و چگالی نقاط تمایل به یکنواخت‌تر شدن دارد. در نتیجه، الگوریتم ممکن است برای شناسایی خوشه های معنی دار مشکل داشته باشد و ممکن است بیشتر نقاط را به عنوان نویز در نظر بگیرد.
- شکل‌ها و اندازه‌های خوشه‌ای متفاوت: خوشه‌بندی مبتنی بر چگالی فرض می‌کند که خوشه‌ها نواحی متراکمی هستند که توسط مناطق با چگالی کمتر از هم جدا شده‌اند. با این حال، اگر خوشه ها دارای اشکال نامنظم یا اندازه های قابل توجهی متفاوت باشند، برای الگوریتم چالش برانگیز است که آستانه های چگالی مناسب برای جداسازی خوشه ها را به طور موثر تعریف کند. ممکن است منجر به ادغام خوشه های کوچکتر به خوشه های بزرگتر یا تقسیم نادرست خوشه های بزرگ به خوشه های کوچکتر شود.
- نویز و نقاط پرت: الگوریتم‌های خوشه‌بندی مبتنی بر چگالی معمولاً نویز و نقاط پرت را با برچسب‌گذاری آنها به عنوان نقاط نویز به خوبی کنترل می‌کنند. با این حال، اگر مجموعه داده حاوی مقدار قابل توجهی نویز یا نقاط پرت باشد که به طور متراکم بسته بندی شده اند، ممکن است بر تعیین آستانه چگالی تأثیر بگذارد. الگوریتم ممکن است به اشتباه این نقاط نویز را به عنوان بخشی از یک خوشه در نظر بگیرد یا خوشه های واقعی را شناسایی نکند.

سوال چهارم

الف) نتیجه اعمال الگوریتم k-means را بر روی داده‌های زیر مشخص کنید. (ضرب در بیانگر مراکز اولیه است)



نتیجه الگوریتم به صورت رو به رو خواهد بود که نادرست است:

زیرا مراکز اولیه نزدیک هم انتخاب شده اند و مقدار دهی اولیه نامناسب باعث می شود حتی با پایان رساندن الگوریتم هنوز نتوان به خوشه بندی مناسب دست یافت.

ب) برای حل مشکل بالا از راهکارهای گوناگونی استفاده می‌شود در رابطه با هر یک از این راهکارها را تحقیق کرده و مزایا و معایب آن‌ها را توضیح دهید

- استفاده از medoid به جای median

در این روش به جای استفاده از میانگین (میانگین) نقاط داده به عنوان مرکز اولیه، از medoid استفاده می شود. Medoid نقطه داده ای در یک خوشه است که کمترین تفاوت میانگین را با سایر نقاط آن خوشه دارد.

مزایا:

- Medoids نقاط داده واقعی هستند که اطمینان حاصل می کنند که مرکزهای اولیه نمایندگان معتبر داده ها هستند.
- مدوئیدها در مقایسه با استفاده از میانگین به عنوان مرکز نسبت به نقاط پرت مقاوم تر هستند.

معایب:

- محاسبه medoidها به محاسبات عدم تشابه زوجی بین تمام نقاط داده نیاز دارد که می تواند از نظر محاسباتی گران باشد.
- شناسایی مناسب‌ترین مدویدها ممکن است چالش‌برانگیز باشد، به‌ویژه زمانی که با داده‌های با ابعاد بالا یا اشکال خوشه‌ای پیچیده سروکار داریم.

- انتخاب نقاط اولیه به شکلی که بیشترین فاصله را از هم داشته باشند

در این روش مرکزهای اولیه به گونه ای انتخاب می شوند که حداکثر فاصله زوجی را از یکدیگر داشته باشند. این تضمین می کند که centroid ها به خوبی در سراسر مجموعه داده توزیع شده اند.

مزایا:

- انتخاب نقاط اولیه با حداکثر فاصله زوجی به جلوگیری از قرار دادن مرکزها در مجاورت نزدیک کمک می کند.
- احتمال همگرایی به یک راه حل خوب را افزایش می دهد.

معایب:

- یافتن نقاطی با بیشترین فاصله نیاز به محاسبات فاصله زوجی بین تمام نقاط داده دارد که می تواند از نظر محاسباتی گران باشد.
- اگر مجموعه داده شامل نویز باشد تاثیر نویز بر روی الگوریتم ما در این حالت زیاد میشود و دچار مشکل میشویم.

- انتخاب نقاط اولیه بر اساس توزیع داده‌ها

در این روش، مرکزهای اولیه بر اساس توزیع داده ها انتخاب می شوند. به عنوان مثال، می توانید مرکزهای اولیه را از مناطق با چگالی بالا یا بر اساس توزیع احتمال متناسب با داده ها انتخاب کنید.

مزایا:

- این روش توزیع داده ها را در نظر می گیرد و می تواند به گرفتن ساختار یا حالت های اساسی داده ها کمک کند.

معایب:

- شناسایی روش مبتنی بر توزیع مناسب برای انتخاب نقاط اولیه می تواند چالش برانگیز باشد و ممکن است به دانش حوزه یا تجزیه و تحلیل آماری نیاز داشته باشد.
- اثربخشی این روش به تناسب مدل توزیع انتخاب شده برای داده ها بستگی دارد.
- اگر توزیع داده‌ها نامنظم باشد یا خوشه‌ها اشکال یا اندازه‌های متفاوتی داشته باشند، این رویکرد ممکن است خوب کار نکند.

- انتخاب چندباره مراکز اولیه برای رسیدن به جواب مناسب

در این روش، مجموعه های متعددی از مرکزهای اولیه به طور تصادفی انتخاب می شوند و الگوریتم k-means چندین بار اجرا می شود. بهترین نتیجه خوشه بندی از اجرای چندگانه انتخاب می شود.

مزایا:

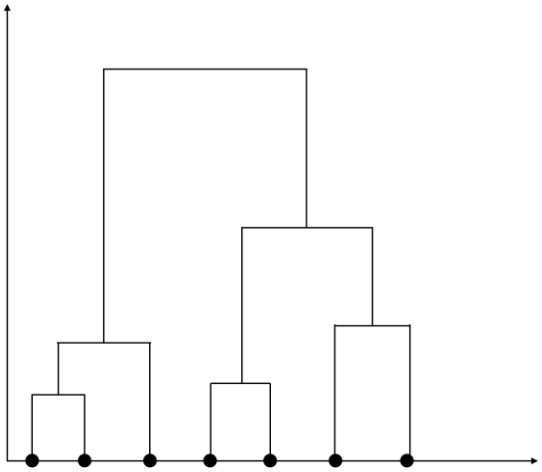
- انتخاب چندین مجموعه از مراکز به کاهش مشکل گیر کردن در بهینه محلی کمک می‌کند.

- اجرای الگوریتم با مقداردهی اولیه، شانس یافتن راه حل بهینه را افزایش می دهد.

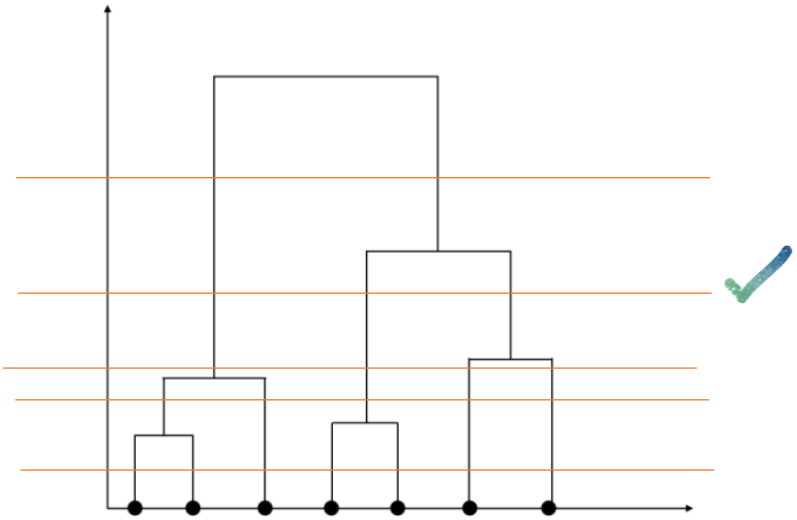
معایب:

- این رویکرد هزینه محاسباتی را افزایش می دهد زیرا به اجرای چندین بار الگوریتم k-means نیاز دارد.
- تعیین بهترین نتیجه خوشه‌بندی در بین چند مقدار اولیه ممکن است به معیارهای ارزیابی اضافی یا قضاوت ذهنی نیاز داشته باشد.

ج) دندروگرام زیر، انجام خوشه‌بندی سلسله مراتبی را بر روی یک مجموعه‌دادگان را نشان می‌دهد، با توجه به دندروگرام مشخص نمایید که اگر بخواهیم بر روی داده‌های زیر الگوریتم k-means را اجرا نماییم بهتر است که چه مقداری را به k دهیم.



با توجه به دندوگرم داده شده تعداد ۳ خوشه برای خوشه بندی مناسب است پی در k-means نیز k = 3 قرارمیدهیم.



سوال پنجم

ماتریس زیر را در نظر بگیرید با استفاده از روش PCA داده‌ها را به یک بعد انتقال داده و ماتریس داده حاصل را بدست آورید.

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 2 & 1 \\ -1 & -1 \\ -1 & -2 \\ -2 & -1 \end{bmatrix}$$

$$Cov(x,y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

$$Cov = \frac{\begin{bmatrix} 1 & 1 & 2 & -1 & -1 & -2 \\ 1 & 2 & 1 & -1 & -2 & -1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 2 & 1 \\ -1 & -1 \\ -1 & -2 \\ -2 & -1 \end{bmatrix}}{6-1} = \frac{\begin{bmatrix} 12 & 10 \\ 10 & 12 \end{bmatrix}}{5} = \begin{bmatrix} 2.4 & 2 \\ 2 & 2.4 \end{bmatrix}$$

حال می بایست، مقادیر ویژه این ماتریس را بررسی کنیم.

$$\det(Cov - \lambda I) = 0$$

$$\det\left(\begin{bmatrix} 2.4-\lambda & 2 \\ 2 & 2.4-\lambda \end{bmatrix}\right) = (2.4-\lambda)^2 - 4 = (2.4-\lambda-2)(2.4-\lambda+2) = 0$$

$$\lambda_1 = 0.4\,, v_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \qquad \lambda_2 = 4.4\,, v_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$dataset_{new} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 2 & 1 \\ -1 & -1 \\ -1 & -2 \\ -2 & -1 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 3 \\ -2 \\ -3 \\ -3 \end{bmatrix}$$

سوال ششم

با فرض آستانه پشتیبانی[3] برابر 0.3 و آستانه اطمینان[4] برابر 0.4 ، مجموعه تمام قوانین انجمنی ممکن را بنویسید

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke
6	Milk, Diaper, Coke
7	Bread, Diaper, Beer

[3] Support
[4] Confidence

Br = Bread , M= Milk , D = Diaper , E = Eggs , Be = Beer , C = Coke

Frequent itemset:

Itemset - 1	frequency	support	Is Frequent?
Br	5	0.72	✓
M	5	0.72	✓
D	6	0.85	✓
Be	4	0.57	✓
E	1	0.14	X
C	3	0.43	✓

در نتیجه هیچ قانونی شامل Eggs نداریم.

Itemset - 2	frequency	support	Is Frequent?
Br, M	3	0.43	✓
Br, D	4	0.57	✓
Br, Be	3	0.43	✓
Br, C	1	0.14	X
M, D	4	0.57	✓
M, Be	2	0.28	X
M, C	3	0.42	✓
D, Be	4	0.57	✓
D, C	3	0.43	✓
Be, C	1	0.14	X

استفاده از Apriori :

Itemset - 3	frequency	support	Is Frequent?
Br, M, D	2	0.28	X
Br, D, Be	3	0.43	✓
M, D, C	2	0.28	X

حال قوانین را براساس مجموعه آیتم ستهای پرتکرار به دست می‌آوریم

Rules:

Rule	confidence	OK?
Br → Be, D	3/5 = 0.6	✓
Be → Br, D	3/4 = 0.75	✓
D → Br, Be	3/5 = 0.6	✓
D, Be → Br	3/4 = 0.75	✓
Br, D → Be	3/4 = 0.75	✓
Be, Br → D	3/3 = 1	✓
Br → M	3/5	✓
M → Br	3/5	✓
Br → D	4/5	✓
D → Br	4/6	✓

Br → Be	3/5	✓
Be → Br	3/4	✓
M → D	4/5	✓
D → M	4/6	✓
M → C	3/5	✓
C → M	3/3	✓
D → Be	4/6	✓
Be → D	4/4	✓
Be → C	3/4	✓
C → Be	3/3	✓

سوال هفتم

با فرض آستانه پشتیبانی $\frac{1}{3}$ و آستانه اطمینان $\frac{2}{3}$، مجموعه آیتم‌های پرتکرار را به دست آورید. در مرحله بعد مجموعه‌ی تمام قوانین انجمنی ممکن را به دست آورید.

تراکنش	آیتم‌ها
T1	$\{a, b, c\}$
T2	$\{d, c\}$
T3	$\{a, b, c\}$
T4	$\{d, b\}$
T5	$\{a, e\}$
T6	$\{a, e, d\}$
T7	$\{b, c\}$
T8	$\{a, b, c, d\}$

Frequent itemset:

Itemset - 1	frequency	support	Is Frequent?
A	5	0.625	✓
B	5	0.625	✓
C	5	0.625	✓
D	4	0.5	✓
E	2	0.25	X

در نتیجه هیچ قانونی شامل E نداریم

Itemset - 2	frequency	support	Is Frequent?
A, B	3	0.375	✓
A, C	3	0.375	✓
A, D	2	0.25	X
B, C	4	0.5	✓

B, D	2	0.25	X
C, D	2	0.25	X

استفاده از Apriori :

Itemset - 3	frequency	support	Is Frequent?
A, B, C	3	0.375	✓

حال قوانین را براساس مجموعه آیتم ست‌های پرتکرار به دست می‌آوریم

Rules:

Rule	confidence	OK?
A → B, C	3/5 = 0.6	X
B → A, C	3/5 = 0.6	X
C → A, B	3/5 = 0.6	X
A, B → C	3/3 = 1	✓
A, C → B	3/3 = 1	✓
B, C → A	3/4 = 0.75	✓
A → B	3/5	X
B → A	3/5	X
A → C	3/5	X
C → A	3/5	X
B → C	4/5	✓
C → B	4/5	✓

Final rules:

A, B → C	**A, C → B**	**B, C → A**
B → C	C → B	