نام و نام خانوادگی: حسنا اویارحسینی	تمرین سوم داده کاوی
شماره دانشجویی: ۹۸۲۳۰۱۰	تاریخ: خرداد - ۱۴۰۲

سوال ۱)

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 متغلق 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).

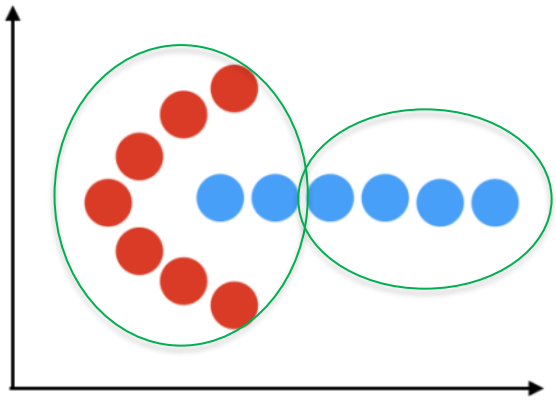
سوال ۲)

- نرم ۱: میانه. هنگام مینیمم کردن عبارت داده شده با فاصله نرم ۱ ، میخواهیم مجموع تعدادی قدر مطلق را مینیمم کنیم که این قدر مطلق ها همان تعریف میانه هستند (نقطه ای که مجموع فاصله آن از بقیه نقاط کمتر است)
- نرم ۲: میانگین. ابتدا مشتق این تابع را محاسبه میکنیم و آن را صفر قرار میدهیم تا نقطه مینیمم را پیدا کنیم، خواهیم داشت:

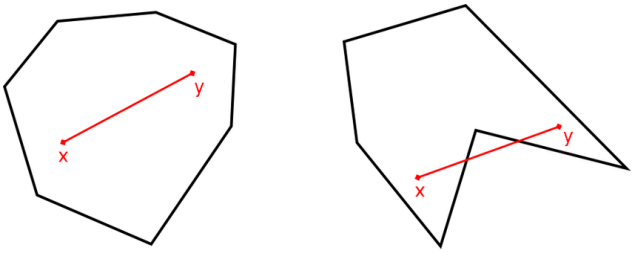
$$\sum_{d}\sum_{c} ۲||d-c|| = ۰ \rightarrow \text{به ازای یک دسته خواهیم داشت} \rightarrow c = \frac{\sum_d d}{۲}$$

سوال ۳)

الف) 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) 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.



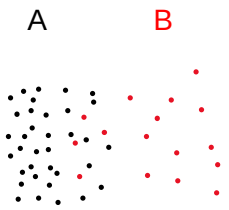
Convex region Non-convex region

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)

۱. 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:



۲. 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.

۳. 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.

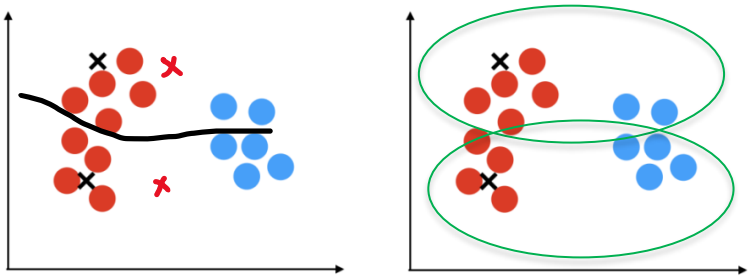
٤. شکل‌ها و اندازه‌های خوشه‌ای متفاوت: خوشه‌بندی مبتنی بر چگالی فرض می‌کند که خوشه‌ها نواحی متراکمی هستند که توسط مناطق با چگالی کمتر از هم جدا شده‌اند. با این حال، اگر خوشه ها دارای اشکال نامنظم یا اندازه های قابل توجهی متفاوت باشند، برای الگوریتم چالش برانگیز است که آستانه های چگالی مناسب برای جداسازی 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.

۵. 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 ۴)

الف)

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)

۱- 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.

۲- انتخاب نقاط اولیه با بیشترین فاصله:

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.

۳- انتخاب نقاط اولیه بر اساس توزیع داده ها:

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:

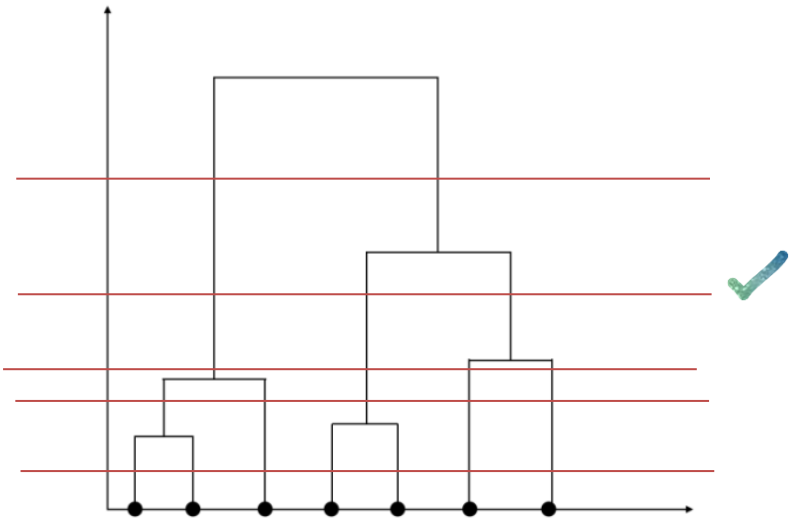
- شناسایی روش مبتنی بر توزیع مناسب برای انتخاب نقاط اولیه می تواند چالش برانگیز باشد و ممکن است به دانش حوزه یا تجزیه و تحلیل آماری نیاز داشته باشد.
- اثربخشی این روش به تناسب مدل توزیع انتخاب شده برای داده ها بستگی دارد.
- اگر توزیع داده‌ها نامنظم باشد یا خوشه‌ها اشکال یا اندازه‌های متفاوتی داشته باشند،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

٤- انتخاب چندگانه مراکز اولیه:

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



سوال ۵)

Standardization:

	Feature1	Feature2
	۱	۱
	۱	۲
	۲	۱

	۱-	۱-
	۱-	۲-
	۲-	۱-
Mean	۰	۰
variance	√۲	√۲

Covariance matrix:

$$\frac{\begin{bmatrix} ۱ & ۱ & ۲ & -۱ & -۱ & -۲ \\ ۱ & ۲ & ۱ & -۱ & -۲ & -۱ \end{bmatrix} \times \begin{bmatrix} ۱ & ۱ \\ ۱ & ۲ \\ ۲ & ۱ \\ -۱ & -۱ \\ -۱ & -۲ \\ -۲ & -۱ \end{bmatrix}}{(۶-۱)\sqrt{۲}\sqrt{۲}} = \frac{\begin{bmatrix} ۱۲ & ۱۰ \\ ۱۰ & ۱۲ \end{bmatrix}}{۱۰}$$

Eigen value:

$\det|A - \lambda I| = ۰ \rightarrow \lambda = ۲.۲, ۰.۲$

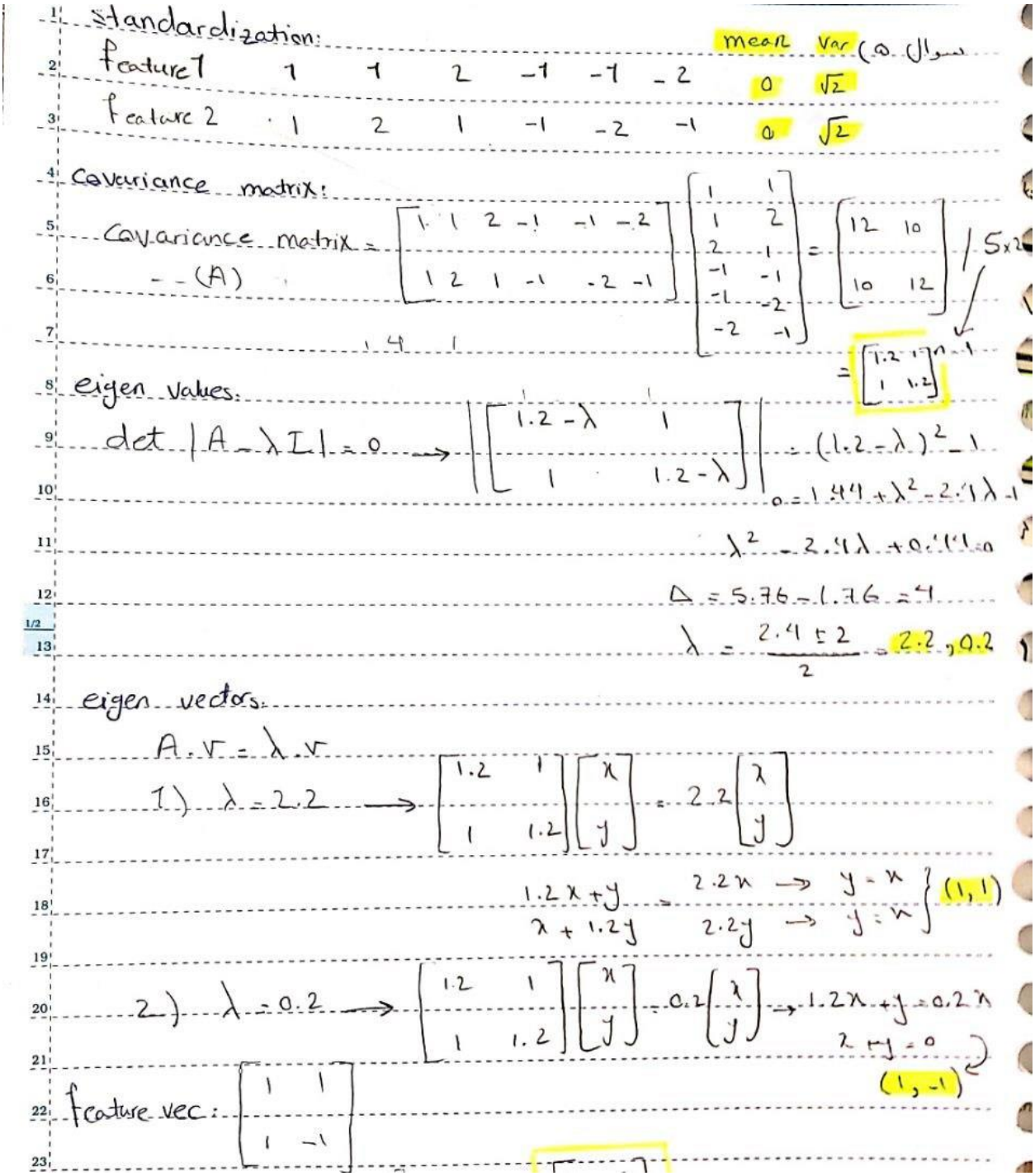
Eigen vectors:

$\lambda = ۲.۲ \rightarrow A.v = \lambda.v \rightarrow x = y \rightarrow v = \begin{bmatrix} ۱ \\ ۱ \end{bmatrix} / \sqrt{۲}$

Feature vector:

$$\begin{bmatrix} ۱ & ۱ \\ ۱ & ۲ \\ ۲ & ۱ \\ -۱ & -۱ \\ -۱ & -۲ \\ -۲ & -۱ \end{bmatrix} \times \begin{bmatrix} ۱ \\ ۱ \end{bmatrix} = \begin{bmatrix} ۲ \\ ۳ \\ ۳ \\ -۲ \\ -۳ \\ -۳ \end{bmatrix} / \sqrt{۲}$$

جزییات محاسبات در ادامه امده است:



سوال ۶)

Frequent itemset:

Itemset - 1	freq	support	OK?
Br	۵	۰.۷۲	✓
M	۵	۰.۷۲	✓

D	۶	۰.۸۵	✓
Be	۴	۰.۵۷	✓
E	۱	۰.۱۴	X
C	۳	۰.۴۳	✓

Itemset - 2	freq	support	OK?
Br, M	۳	۰.۴۳	✓
Br, D	۴	۰.۵۷	✓
Br, Be	۳	۰.۴۳	✓
Br, C	۱	۰.۱۴	X
M, D	۴	۰.۵۷	✓
M, Be	۲	۰.۲۸	X
M, C	۳	۰.۴۲	✓
D, Be	۴	۰.۵۷	✓
D, C	۳	۰.۴۳	✓
Be, C	۱	۰.۱۴	X

Itemset - 3	freq	support	OK?
Br, M, D	۲	۰.۲۸	X
Br, D, Be	۳	۰.۴۳	✓
Br, D, C	۱	۰.۱۴	X
M, D, Be	۲	۰.۲۸	X
M, D, C	۲	۰.۲۸	X
Be, D, C	۱	۰.۱۴	X
Br, Be, M	۱	۰.۱۴	X
M, C, Br	۱	۰.۱۴	X

Rules:

Rule	confidence	OK?
Br → Be, D	۳/۵ = ۰.۶	✓
Be → Br, D	۳/۴ = ۰.۷۵	✓

D → Br, Be	۳/۵ = ۰.۶	✓
D, Be → Br	۳/۴ = ۰.۷۵	✓
Br, D → Be	۳/۴ = ۰.۷۵	✓
Be, Br → D	۳/۳ = ۱	✓
Br → M	۳/۵	✓
M → Br	۳/۵	✓
Br → D	۴/۵	✓
D → Br	۴/۶	✓
Br → Be	۳/۵	✓
Be → Br	۳/۴	✓
M → D	۴/۵	✓
D → M	۴/۶	✓
M → C	۳/۵	✓
C → M	۳/۳	✓
D → Be	۴/۶	✓
Be → D	۴/۴	✓
Be → C	۳/۴	✓
C → Be	۳/۳	✓

Final rules:

همه قوانینی که در بالا امده

سوال ۷)

Frequent itemset:

Itemset - 1	freq	support	OK?
A	۵	۰.۶۲۵	✓
B	۵	۰.۶۲۵	✓
C	۵	۰.۶۲۵	✓
D	۴	۰.۵	✓
E	۲	۰.۲۵	X

Itemset - 2	freq	support	OK?

A, B	۳	۰.۳۷۵	✓
A, C	۳	۰.۳۷۵	✓
A, D	۲	۰.۲۵	X
B, C	۴	۰.۵	✓
B, D	۲	۰.۲۵	X
C, D	۲	۰.۲۵	X

Itemset - 3	freq	support	OK?
A, B, C	۳	۰.۳۷۵	✓

Rules:

Rule	confidence	OK?
A → B, C	۳/۵ = ۰.۶	X
B → A, C	۳/۵ = ۰.۶	X
C → A, B	۳/۵ = ۰.۶	X
A, B → C	۳/۳ = ۱	✓
A, C → B	۳/۳ = ۱	✓
B, C → A	۳/۴ = ۰.۷۵	✓
A → B	۳/۵	X
B → A	۳/۵	X
A → C	۳/۵	X
C → A	۳/۵	X
B → C	۴/۵	✓
C → B	۴/۵	✓

Final rules:

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