1 . یک مجموعه داده شامل اشیاءA ، B، C، D، E، F با ماتریس فاصله زیر داده شده است:

distance	A	B	C	D	E	F
A	0	1	2	4	6	7
B		0	3	8	9	10
C			0	11	12	13
D				0	14	15
E					0	16
F						0

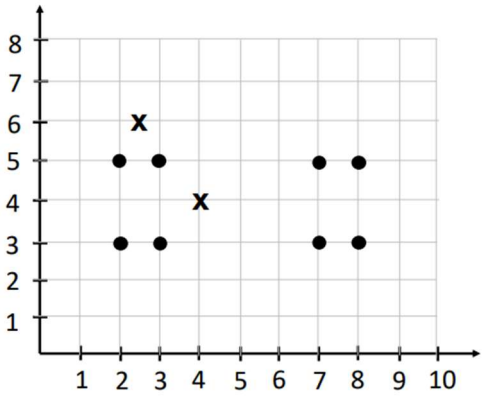
فرض کنید الگوریتم DBSCAN برای این مجموعه داده با MINPOINTS = 3 و epsilon = ε = 5 اجرا شود. الگوریتم DBSCANچند خوشه باز می‌گرداند و چه داده هایی در هر خوشه هستند؟ کدام اشیاء در نتیجه خوشه‌بندی قبلی به عنوان نقاط برون‌زای (outliers) و نقاط مرزی (borderpoints) شناخته می‌شوند؟

	Neighbors (witin epsilon)	Initial status (# of neighbors less than minpoints?)	Status (final status for border points)
A	A,B,C,D	Core	Core
B	A,B,C	Core	Core
C	A,B,C	Core	Core
D	A,D	Noise	Border (because A is Core)
E	E	Noise	Noise
F	F	Noise	Noise

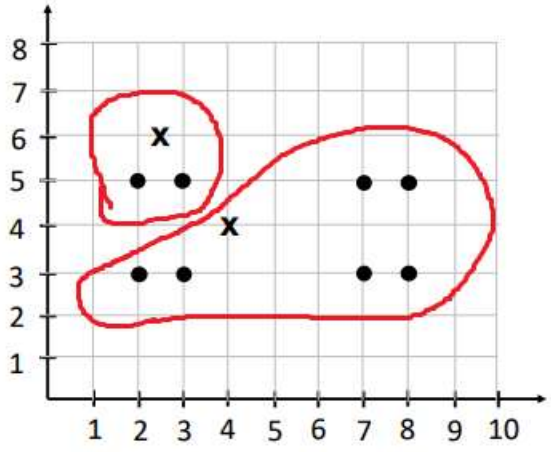
تنها یک خوشه خواهیم داشت: {A,B,C,D}

2 . فرض کنید تعداد خوشه‌ها k=2 برای تمام قسمت های زیر باشد.

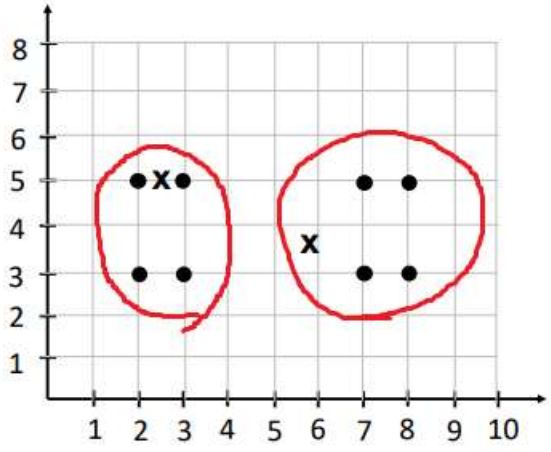
(الف) مراحل الگوریتم k-means را گام به گام طی کنید، از مقداردهی اولیه‌ای که در نمودار سمت چپ بالا در جعبه زیر نشان داده شده است شروع کنید. نقاط، داده‌های مشاهده شده را نشان می‌دهند. با دو علامت 'x' مکان مراکز خوشه‌ها و با رسم دوایر داده های هر خوشه رادر هر تکرار الگوریتم k-means مشخص کنید. مراکز خوشه های اولیه در حالت اولیه از پیش نشان داده شده‌اند.



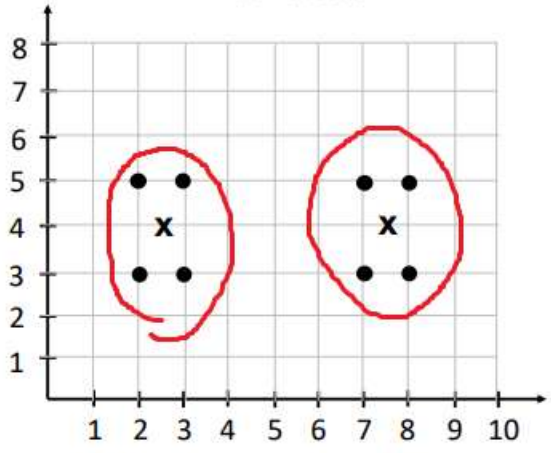
Initial state



Initial state



After iteration 1

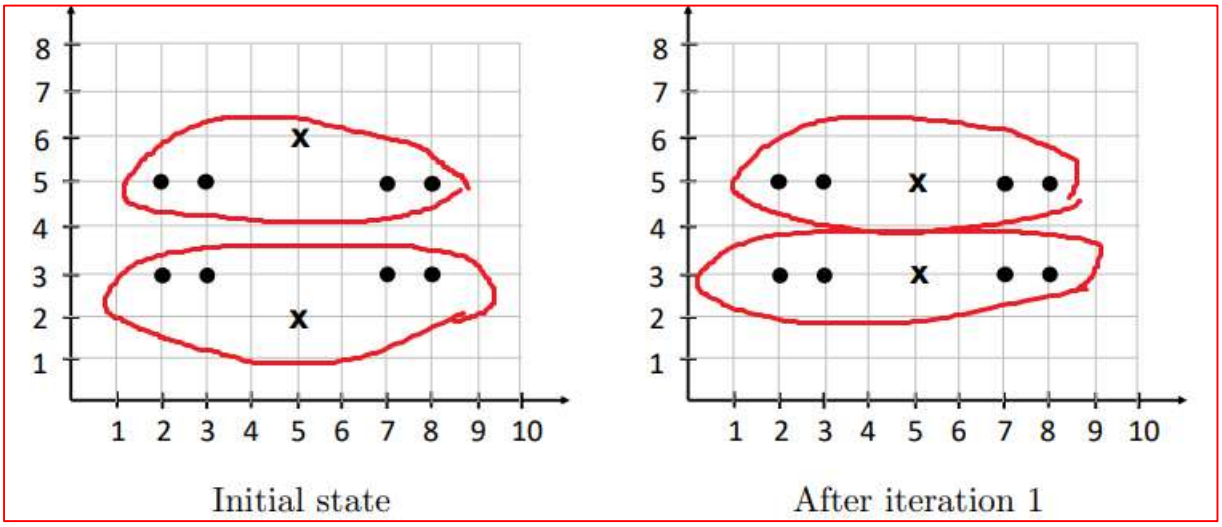


After iteration 2

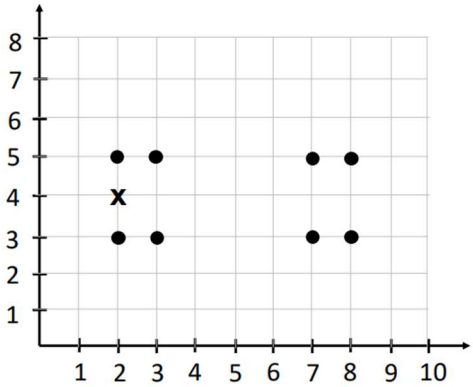
ب) مقدار عددی inertia برای خوشه‌بندی به‌دست‌آمده در قسمت (الف) پس از اتمام اجرای الگوریتم چیست؟

$$inertia = \sum_{j=1}^{k}\sum_{i=1}^{n_j} \left||x_i - C_j|\right|^2 = 8 \times ((2.5-2)^2 + (4-3)^2) = 10$$

ج) مانند قسمت (الف)، مراحل الگوریتم k-means را گام به گام طی کنید، از مقداردهی اولیه‌ای که در نمودار سمت چپ بالا نشان داده شده است شروع کنید. مقدار inertia در این حالت چند است؟ کدام مقداردهی اولیه بهتر است؟



$$inertia = 4 \times ((7-5)^2 + (8-5)^2) = 52$$

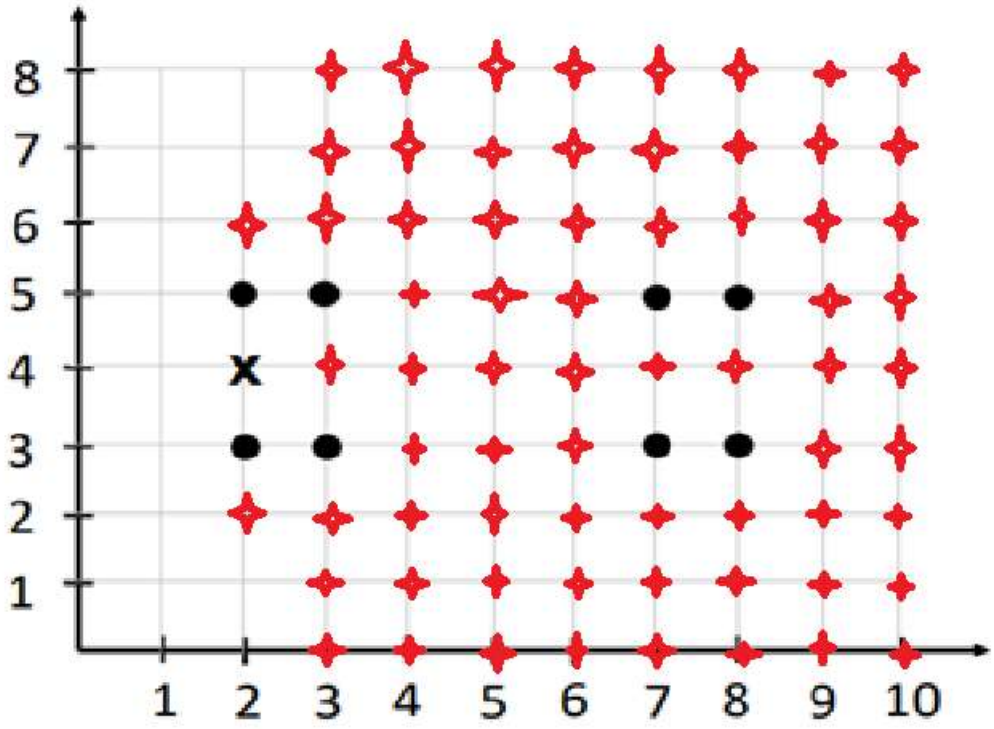


د) نقاط داده‌ای را که با دایره‌های سیاه در نمودار زیر نشان داده شده‌اند در نظر بگیرید. ما یک مرکز خوشه را با علامت x در مختصات (2,4) رسم کرده‌ایم. مرکز خوشه دوم را به گونه‌ای رسم کنید که شرط زیر را برآورده کند:

وقتی مراکز خوشه‌ها را در دو نقطه مقداردهی اولیه کنیم و الگوریتم k-means را تا همگرایی اجرا کنیم، وضعیت نهایی به گونه‌ای خواهد بود که یک خوشه شامل تمام نقاط داده باشد و خوشه دیگر هیچ نقطه‌ای نداشته باشد.

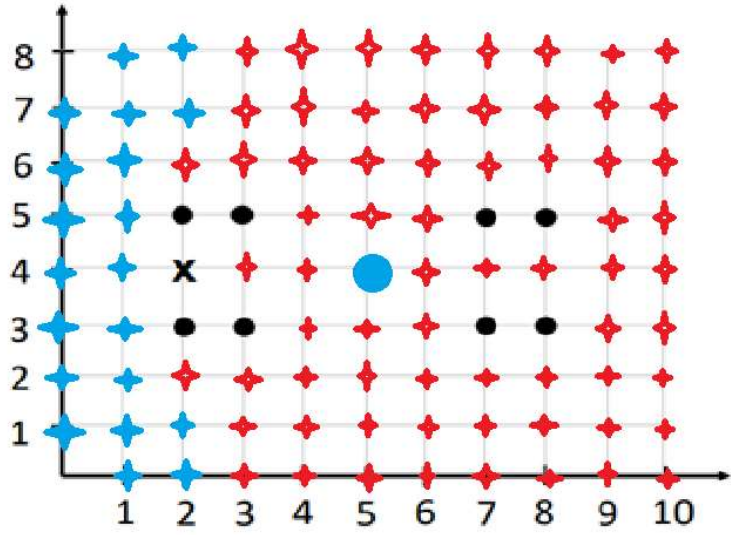
هر کدام از نقاط (0,0) و (0,8) جواب درست هستند.

چرا؟ در گام اول باید تمام نقاط فاصله کمتری با مرکز (2,4) داشته باشند. من جمله (8,5) و (8,3) که فاصله $\sqrt{37}$ با این نقطه دارند. کدام مراکز ممکن دوم فاصله کمتری با (8,5) و (8,3) دارند؟ این نقاط در زیر علامت زده شده اند. این نقاط نمی توانند به عنوان مرکز دوم انتخاب شوند.



تمام این نقاط نمی توانند به عنوان مرکز دوم انتخاب شوند چون فاصله آنها با (8,3) و (8,5) از رادیکال ۳۷ کمتر یا مساوی است.

حالا فرض کنید در گام دوم، تمام نقاط در یک خوشه جمع شده باشند و مرکز جدید برابر با (5,4) خواهد بود که در شکل زیر با دایره بزرگ نشان داده شده است. حالا باید نقاط (2,3) و (2,5) فاصله بیشتری با مرکز دوم داشته باشند تا مرکز جدید (5,4). یعنی فاصله مرکز دوم با این دو نقطه باید از $\sqrt{16+1}=\sqrt{17}$ بیشتر باشد. نقاطی که نمی توانند این شرط را برآورده کنند، در زیر با آبی مشخص شده اند. این نقاط نیز نمی توانند مرکز دوم باشند.



لذا تنها نقاط (0,0) و یا (0,8) می توانند به عنوان مرکز دوم انتخاب شوند.

---

۳. کدام‌یک از روش‌های خوشه بندی داده شده داده‌های موجود شکل زیر را به دو خوشه دایره قرمز و خط افقی آبی تقسیم می‌کند؟ هر نقطه در دایره و خط یک نقطه داده است. در تمام گزینه‌هایی که شامل خوشه‌بندی سلسله‌مراتبی هستند، الگوریتم تا زمانی اجرا می‌شود که دو خوشه به دست آید.

الف)خوشه بندی سلسله مراتبی complete-link

ب) خوشه بندی سلسله مراتبی single-link

ج) خوشه بندی kmeans

فقط single linkage.

خوشه بندی k-means علاقمند به تشکیل خوشه های دایروی است و خط مستقیم رو به دو قسمت خواهد شکست (خوشه ها یک قسمت از خط و قسمت باقی مانده از خط به همراه دایره در خوشه دیگر). خوشه بندی complete-linkage نیز به همین ترتیب اشتباه خواهد کرد (بخاطر وجود max در رابطه محاسبه فاصله دو خوشه در حین ادغام).

---

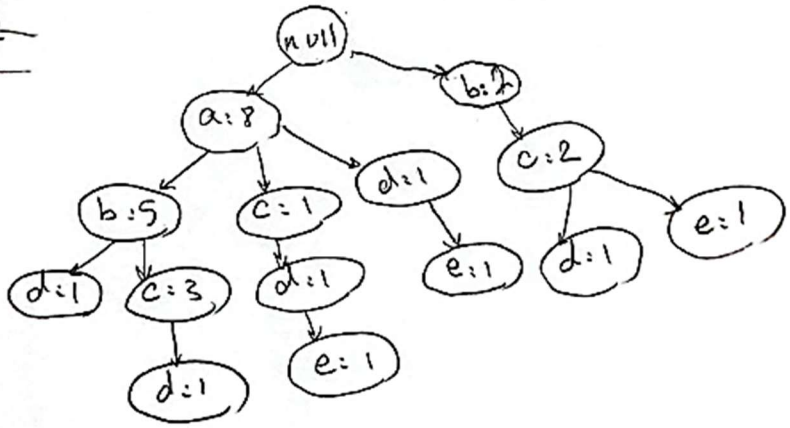
۴. با استفاده از روش Fp-growth آیتم های پرتکرار با min-support count=3 را بدست آورید:

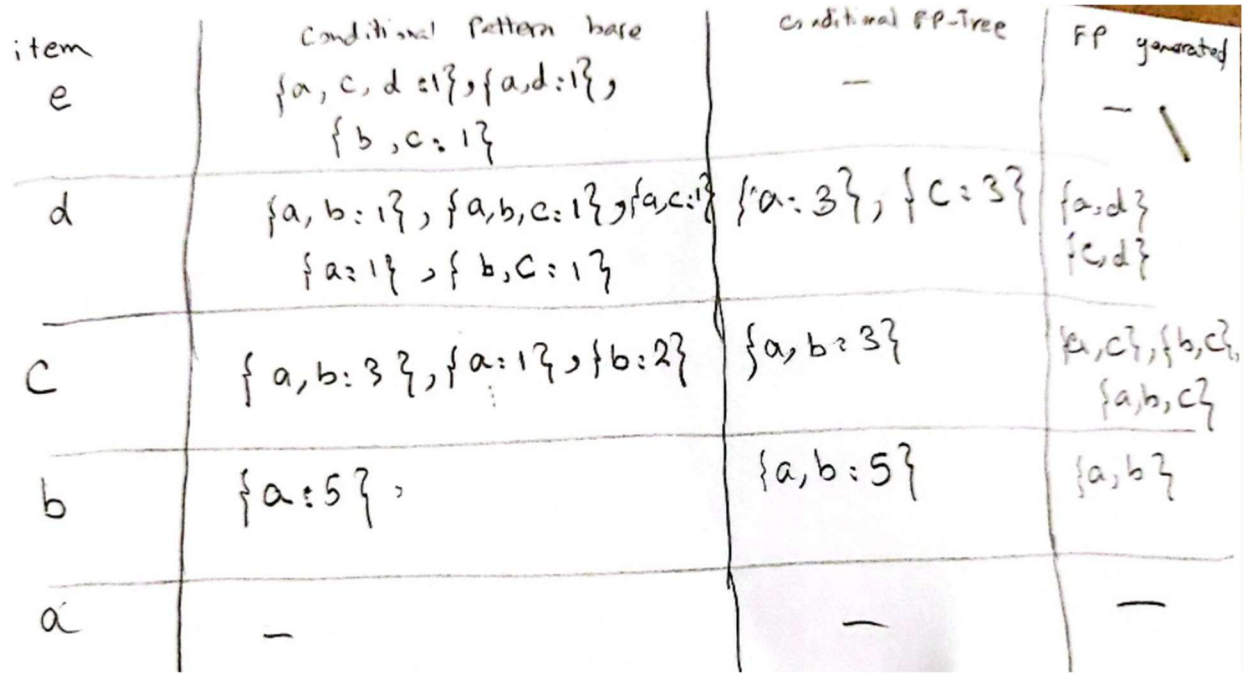
TID	Items Bought
1	a, b, f
2	b, g, c, d
3	h, a, c, d, e
4	a, d, p, e
5	a, b, c
6	a, b, q, c, d
7	a
8	a, m, b, c
9	a, b, n, d
10	b, c, e, m

item	support
a	8
b	7
c	6
d	5
e	5
~~f~~	~~1~~
~~g~~	~~1~~
~~h~~	~~1~~
~~p~~	~~1~~
~~q~~	~~1~~
~~m~~	~~2~~
~~n~~	~~1~~

sort & filter transactions

Tid	items bought
1	a, b
2	b, c, d
3	a, c, d, e
4	a, d, e
5	a, b, c
6	a, b, c, d
7	a
8	a, b, c
9	a, b, d
10	b, c, e





item	Conditional Pattern base	Conditional FP-Tree	FP generated
e	{a, c, d :1}, {a,d:1}, {b ,c: 1}	—	—
d	{a, b: 1}, {a,b,c: 1}, {a,c:1} {a:1}, {b,c: 1}	{a: 3}, {c: 3}	{a,d} {c,d}
c	{a,b: 3}, {a: 1}, {b: 2}	{a, b: 3}	{a,c}, {b,c}, {a,b,c}
b	{a: 5},	{a,b: 5}	{a,b}
a	—	—	—