داده کاوی

فصل دهم

تهیه شده توسط
محمد پورزعفرانی
Pourzaferani.ir
دانشگاه اصفهان

دانشگاه اصفهان

تشکر

- آقایون حاضر در کلاس

فصل ۱۰ : خوشه‌بندی

- تحلیل خوشه‌ای: مفاهیم اصلی
- روش‌های بخش‌بندی (Partitioning Methods)
- روش‌های سلسله مراتبی (Hierarchical Methods)
- روش مبتنی بر چگالی (density-based)
- روش‌های مبتنی بر گرید (grid-based)
- ارزیابی روش‌های خوشه‌بندی
- خلاصه

تحلیل خوشه ای چیست؟

- **تعریف خوشه: مجموعه‌ای از اشیای داده‌ای**
  - مشابه (یا مربوط) همدیگر در یک گروه واحد
  - متفاوت (یا نامربوط) به اشیاء موجود در گروه‌های دیگر
- **تحلیل خوشه (یا خوشه بندی، بخش بندی داده ها، ...)**
  - پیدا کردن شباهت بین داده‌ها با توجه به ویژگی‌ها موجود و گروه‌بندی اشیاء داده‌ای مشابه در یک خوشه
- **آموزش بدون نظارت: هیچ کلاسی از پیش تعریف شده نیست**
- **برنامه های کاربردی نمونه**
  - به عنوان یک ابزار مستقل برای به دست آوردن دیدی نسبت به توزیع داده‌ها
  - به عنوان یک گام پیش‌پردازش برای الگوریتم‌های دیگر

چه چیزی تحلیل خوشه‌ای نیست؟

- **طبقه بندی نظارت شده (Supervised classification)**
  - داشتن برچسب کلاس
- **بخش بندی ساده (Simple segmentation)**
  - تقسیم دانش آموزان به گروه‌های مجزا بر اساس حروف الفبای نام خانوادگی
- **نتایج یک پرس و جو (Results of a query)**
  - گروه بندی ناشی از خصوصیات خارجی

خوشه بندی برای فهم داده ها و برنامه های کاربردی

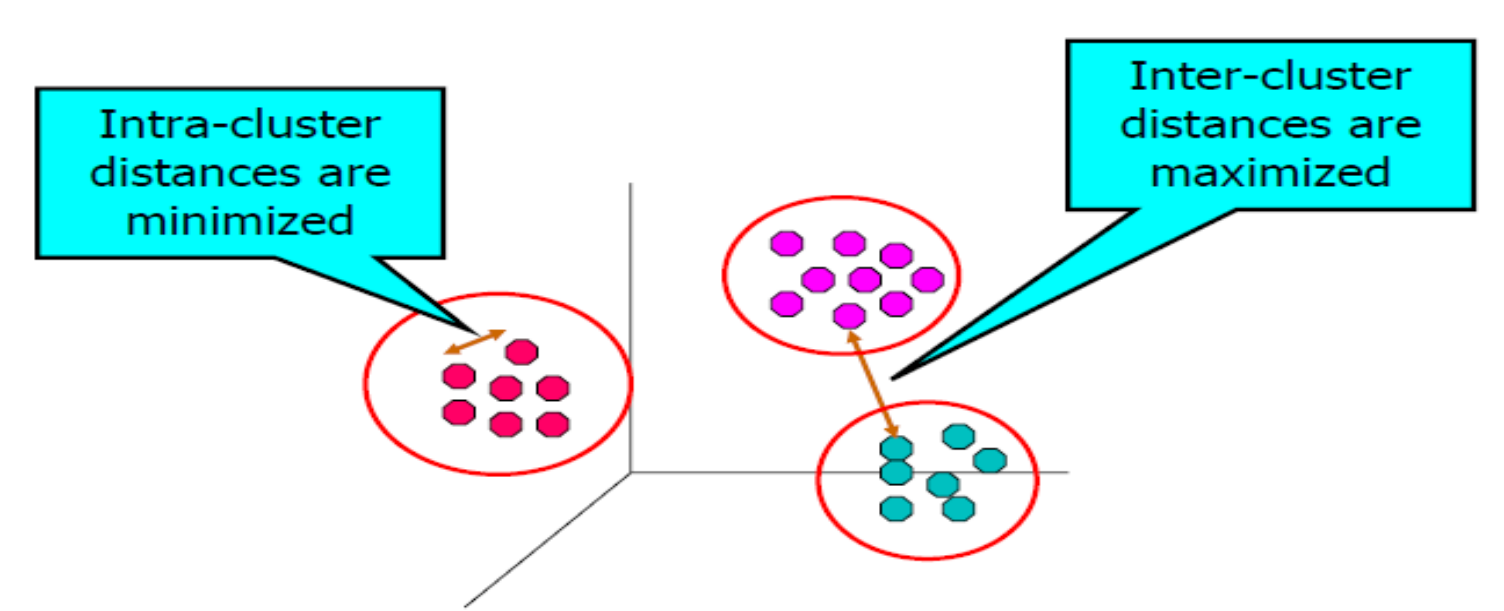
- زیست شناسی: طبقه بندی از موجودات زنده: قلمرو، شاخه، طبقه، نظم، خانواده، جنس و گونه
- بازیابی اطلاعات (Information retrieval): خوشه بندی اسناد
- استفاده از زمین: شناسایی مناطق با کاربری یکسان در پایگاه داده رصد زمین
- بازاریابی: کشف گروه‌های متمایز در پایگاه‌های مشتریان و سپس استفاده از این دانش برای توسعه برنامه های بازاریابی هدفدار
- برنامه ریزی شهری : شناسایی گروه‌های مسکن با توجه به نوع خانه، ارزش، و موقعیت جغرافیایی آن‌ها
- مطالعات زمین لرزه: زمین لرزه مشاهده شده در امتداد گسل‌های قاره‌ای هستند
- آب و هوا: بدست آوردن درک درستی از آب و هوا، برای پیدا کردن الگوهای جوی و اقیانوس‌ها
- علوم اقتصادی: مطالعات داد و ستد

خوشه بندی به عنوان یک ابزار پیش پردازش

- **خلاصه سازی:**
  - پردازش برای رگرسیون، PCA، دسته‌بندی، و ....
- **فشرده سازی:**
  - پردازش تصویر
- **پیدا کردن K-نزدیکترین همسایگان:**
  - محلی‌سازی (Localizing) جستجو به یک یا تعداد کمی از خوشه ها
- **آشکارسازی پرت:**
  - دادههای پرت اغلب به عنوان داده‌های «خیلی دور" از هر خوشه مشاهده می‌شوند

تحلیل خوشه چیست؟

- پیدا کردن گروه‌هایی از اشیاء به طوری که اشیاء در یک گروه مشابه (و یا مرتبط) به یک دیگر و متفاوت از (و یا نامربوط) اشیاء موجود در گروه‌های دیگر



چه خوشه‌بندی خوب است؟

- یک خوشه‌بندی خوب معیارهای زیر را دارد
  - بیشترین شباهت درون-کلاسی (intra-class) : چسباندی (cohesive) داخل خوشه‌ای
  - کمترین شباهت بین-کلاسی (inter-class) : تمایز (distinctive) بین خوشه‌ها
  - کیفیت یک روش خوشه‌بندی مبتنی است بر:
    - معیارهای اندازه‌گیری شباهت استفاده شده در روش
    - توانایی روش در کشف برخی یا همه الگوهای پنهان (hidden pattern)

اندازه گیری کیفیت خوشه بندی

- **عدم تشابه / شباهت متریک**
  - شباهت بر اساس تابع فاصله بیان می شود، معمولا متریک: $d(i, j)$
  - می‌توان به ویژگی‌های مختلف بر اساس برنامه کاربردی یا مفاهیم معنایی وزن داد
- **کیفیت خوشه بندی:**
- معمولا "کیفیت" توسط تابع جداگانه‌ای است که "خوبی" خوشه‌بندی را ارزیابی می‌کند، بیان می‌شود

ملاحظات برای تجزیه و تحلیل خوشه

- **معیارهای پارتیشن بندی**
- بخش‌بندی می‌تواند به صورت تک-سطحی یا سلسله مراتبی انجام شود (اغلب، چند سطح بخش‌بندی سلسله مراتبی مطلوب است)

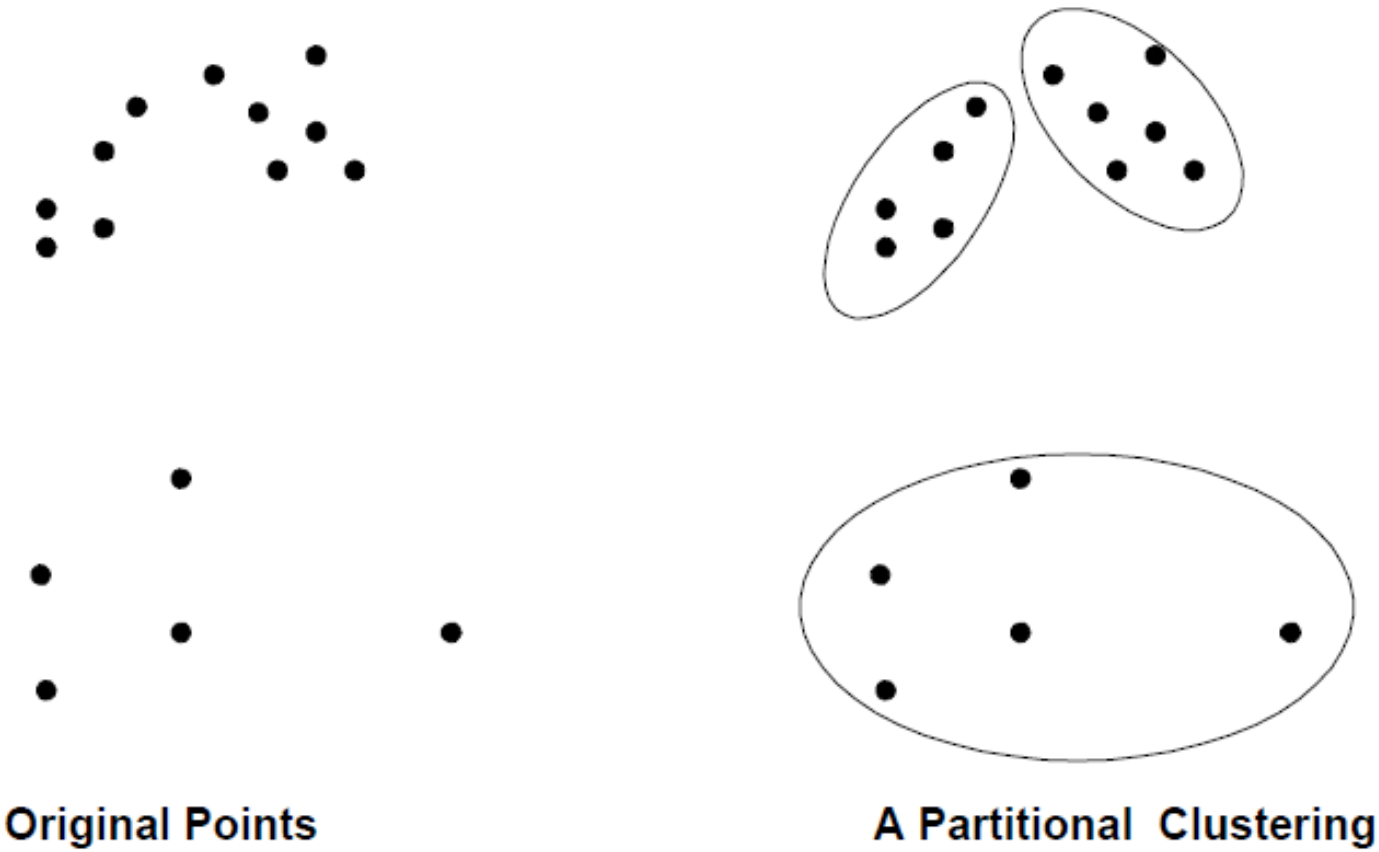
۱. **خوشه‌بندی به صورت Partitional**

- اطلاعات بخش اشیاء به غیر هم تداخل دارند زیر مجموعه (خوشه ها) به طوری که هر شی داده است در دقیقا یک زیر مجموعه

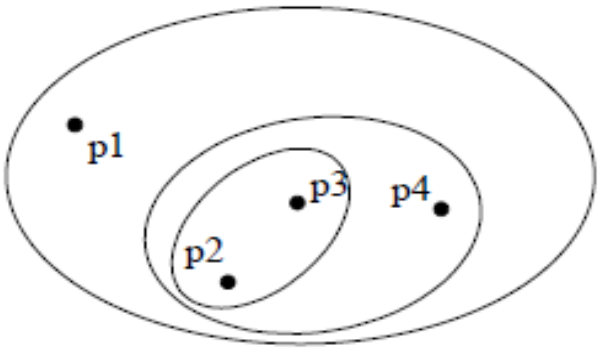
۱. **خوشه‌بندی سلسله مراتبی**

- مجموعه ای از خوشه ها که به صورت یک درخت سلسله مراتبی سازماندهی شده‌اند

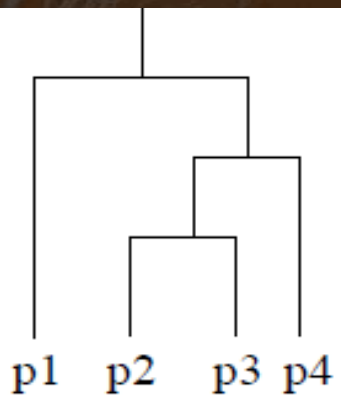
پارتیشن خوشه بندی



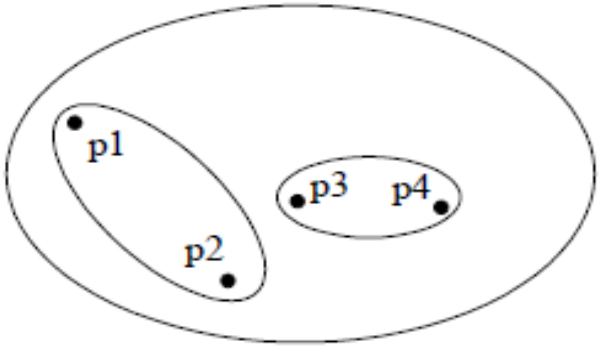
خوشهبندی سلسله مراتبی



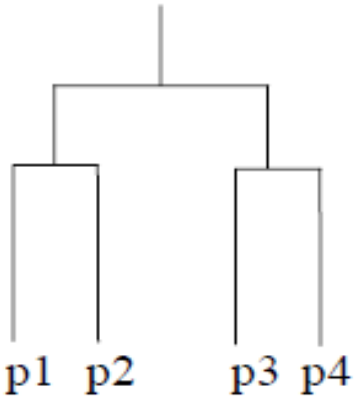
Traditional Hierarchical Clustering



Traditional Dendrogram



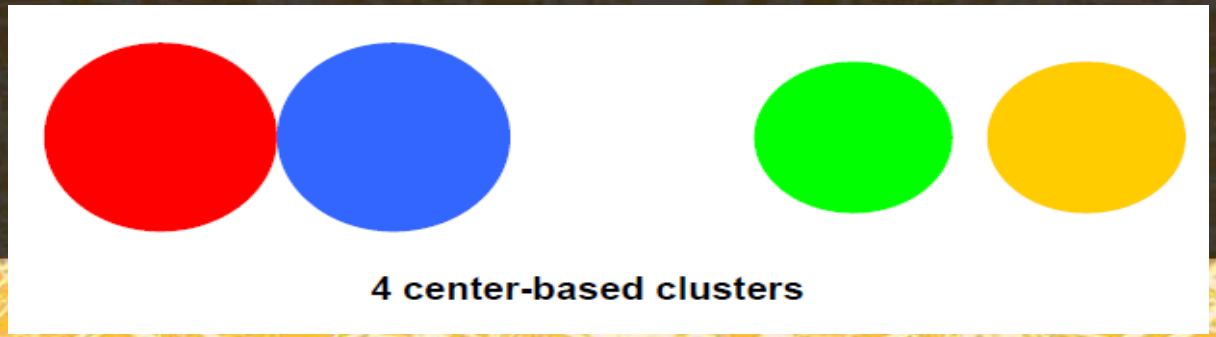
Non-traditional Hierarchical Clustering



Non-traditional Dendrogram

ملاحظات برای تجزیه و تحلیل خوشه

- **اندازه‌گیری شباهت**
- مبتنی بر فاصله (**Distance-based**) (به عنوان مثال ، اقلیدسی، فاصیه برداری) در مقابل
- مبتنی بر اتصال (**connectivity-based**) (به عنوان مثال ، چگالی و یا مجاورت)
- **مرکز خوشه‌ها**
- خوشه مجموعه ای از اشیاء به طوری که اشیا آن خوشه نزدیک (مشابه) به مرکز خوشه هستند
- مرکز یک خوشه که اغلب به صورت centroid بیان می‌شود که برابر است با میانگین نقاط داخل خوشه یا
- به صورت medoid بیان می‌شود که برابر است با میانه نقاط خوشه



ملاحظات تجزیه و تحلیل خوشه

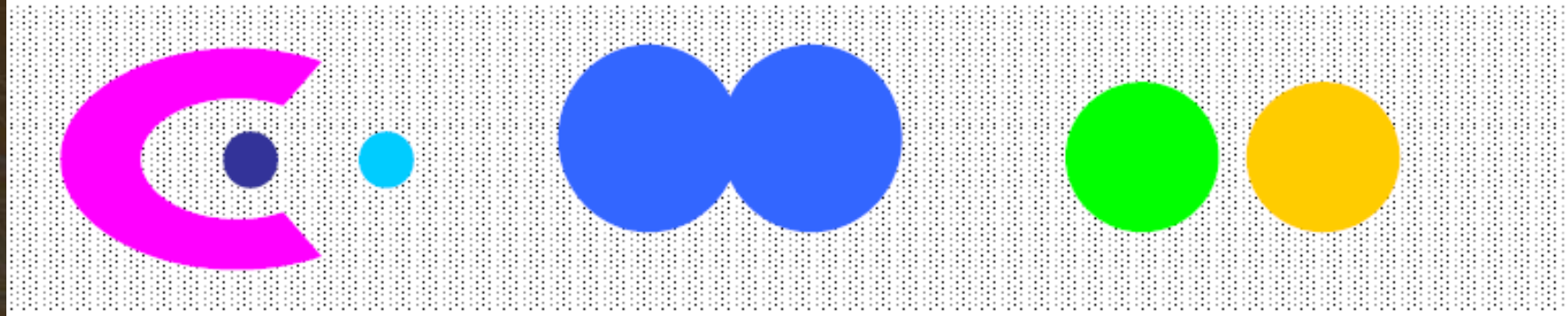
- **خوشه های همجوار(نزدیکترین همسایه یا متعدی)**

یک خوشه مجموعه ای از اشیاء است به طوری که هر شیء درخوشه به نقطه مرکز خوشه نزدیکتر(یا شبیه تر)از مرکز سایر خوشه ها باشد.

8 contiguous clusters

ملاحظات تجزیه و تحلیل خوشه

- **بر اساس چگالی (Density-based)**
- یک خوشه مجموعه‌ای از نقاط متراکم است که از طریق جداسازی نواحی با چگالی کم از نواحی با چگالی بالا بدست می‌آیند
- زمانی که شکل خوشه‌ها نامنظم و یا در هم تنیده شده باشد یا در زمانی که داده های نویز یا دور افتاده وجود دارد مورد استفاده قرار می‌گیرد



6 density-based clusters

الزامات و چالش‌های خوشه‌بندی

- **مقیاس پذیری**

  خوشه بندی همه داده ها به جای نمونه ای از داده.

- **توانایی در خوشه بندی انواع مختلف داده ها**

  عددی،باینری،ترتیبی، مرتبط، و مخلوطی از اینها

- **خوشه بندی مبتنی بر محدودیت**

  کاربر ممکن است در ورود اطلاعات محدودیت داشته باشد.

  استفاده از دامنه دانش برای تعیین پارامترهای ورودی

- **قابلیت تفسیر واستفاده**
  - **سایر ملزومات**
    - کشف شکل دلخواه برای خوشه
    - توانایی مقابله با داده های نویز
    - خوشه بندی افزایشی و حساس نبودن به ورودی ترتیبی
    - ابعاد بالا

روش‌های عمده خوشه‌بندی

- روش پارتیشن بندی(Partitioning approach):

ساختن پارتیشن های مختلف و ارزیابی آنها توسط برخی از معیارها به عنوان مثال:حداقل کردن مجموع مربع‌های خطا.

روشهای معمول:k-means, k-medoids, CLARANS

- روش سلسله مراتبی(Hierarchical approach):

ایجاد تجزیه سلسله مراتبی از مجموعه ای از داده (و یا اشیاء) با استفاده از برخی از معیارها

روشهای معمول:Diana, Agnes, BIRCH, CAMELEON

- روش مبتنی بر چگالی(Density-based approach):

براساس اتصال و تراکم نقاط

روشهای معمول:DBSACN, OPTICS, DenClue

- روش مبتنی بر گرید(Grid-based approach):

براساس ساختار چند سطحی دانه‌بندی شده (granularity

روشهای معمول:STING, WaveCluster, CLIQUE

فصل ۱۰. تجزیه خوشه : مفاهیم پایه و روش ها

- تجزیه و تحلیل خوشه: مفاهیم پایه
- روش بخش‌بندی (Partitioning)
- روش سلسله مراتبی
- روش بر اساس تراکم
- روشهای مبتنی بر شبکه
- ارزیابی خوشه بندی
- خلاصه

الگوریتم‌های بخش‌بندی: مفهوم پایه

- روش‌های بخش‌بندی:پارتیشن بندی پایگاه داده D از nشیء به صورت k خوشه، به طوری که مجموع مربعات خطا حداقل شود (که در آن ci مرکز و یا medoid خوشه C است)

$$E=\Sigma_{i=1}^{k}\Sigma_{p\in C_i}(p-c_i)^2$$

- هدف این است که با توجه به k داده شده، k خوشه‌ای را پیدا کنیم که معیار بخش‌بندی را بهینه نمایند
- بهینه سراسری :در نظر گرفتن تمام بخش‌بندی‌ها
- روشهای اکتشافی: الگوریتم های *k-means* و *k-medoids*

k-means:هر خوشه با استفاده از مرکز خوشه معرفی می‌شود.

k-medoids **or PAM (Partition around medoids)**: هر خوشه با یک نماینده معرفی می‌شود

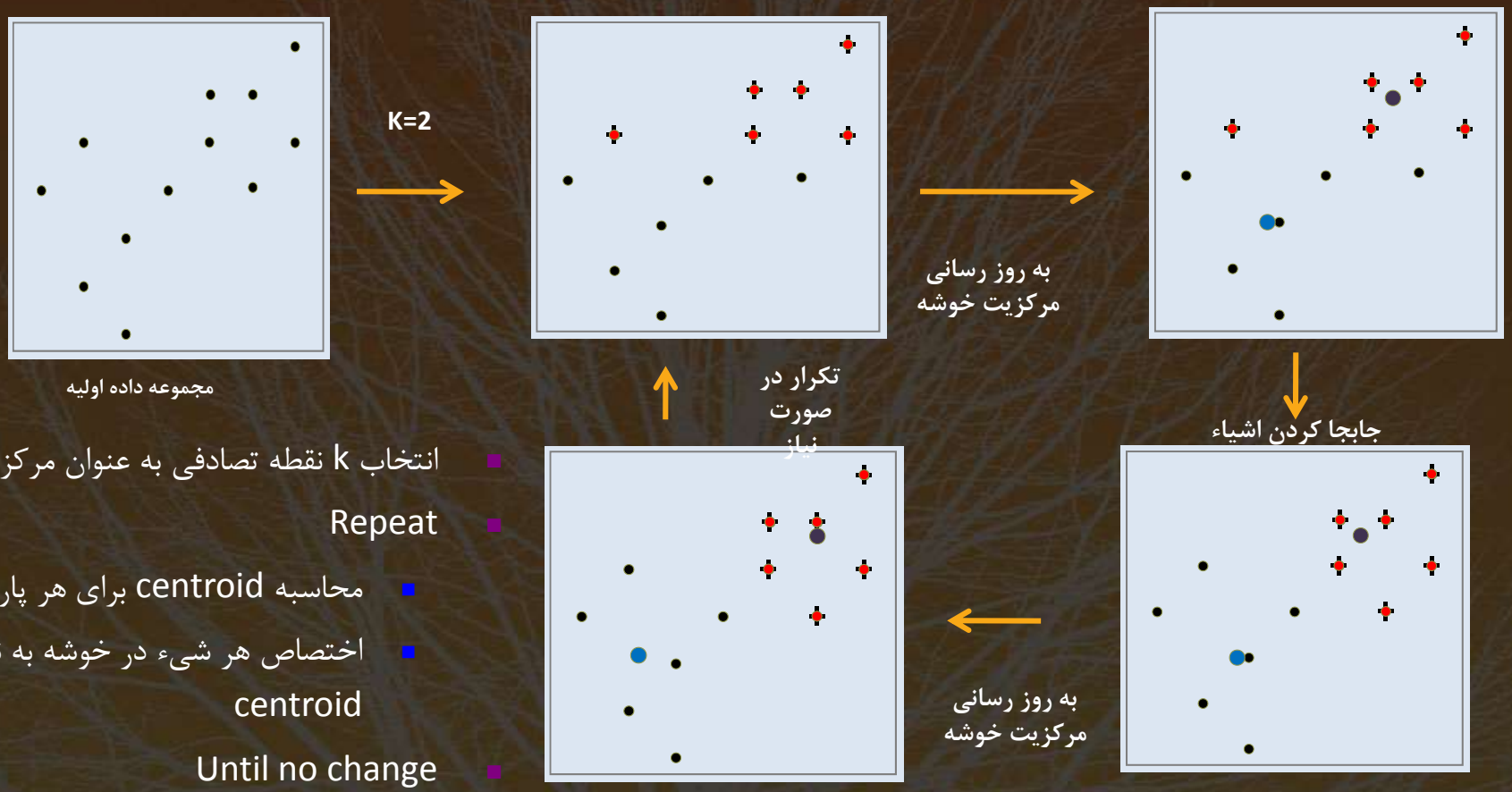
خوشه بندی K-Means

- روش خوشه بندی Partitional.
- هر خوشه با مرکز آن در ارتباط است (Centroid).
- هر نقطه در خوشه به نزدیکترین مرکز از خوشه‌ها اختصاص داده می‌شود.
- تعداد خوشه، K، باید مشخص باشد.
- الگوریتم اصلی بسیار ساده است.

Algorithm 1 Basic K-means Algorithm.

```
Select K points as the initial centroids.
repeat
    Form K clusters by assigning all points to the closest centroid.
    Recompute the centroid of each cluster.
until The centroids don't change
```

یک مثال از خوشه بندی K-Means



- انتخاب k نقطه تصادفی به عنوان مرکز خوشه
- Repeat
  - محاسبه centroid برای هر پارتیشن
  - اختصاص هر شیء در خوشه به نزدیکترین centroid
- Until no change

جزئیات خوشه بندی K-Means

- نقطه مرکز اولیه اغلب به طور تصادفی انتخاب می شود.
  در هر اجرا خوشه های تولید شده با هم متفاوت است.
- نقطه مرکز(به طور معمول) میانگین نقاط در خوشه است.
- "نزدیکی" با فاصله اقلیدسی، شباهت کسینوسی، همبستگی و غیره اندازه گیری می شود.
- K-Means بر اساس یک مقدار مشخص از معیارهای تعیین شده در مرحله قبل همگرا می‌شود
- بسیاری از همگرایی در چند تکرار اول اتفاق می افتد.
- اغلب اوقات معیار توقف این‌گونه تعیین می‌شود. « تا زمانی که نقاط تغییر نکنند» ادامه بده
- پیچیدگی برابر است با:

O(n * K * I * d)

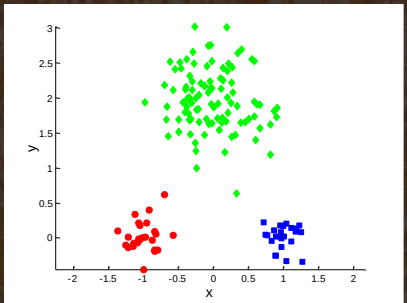
n=تعداد نقاط

k=تعداد خوشه

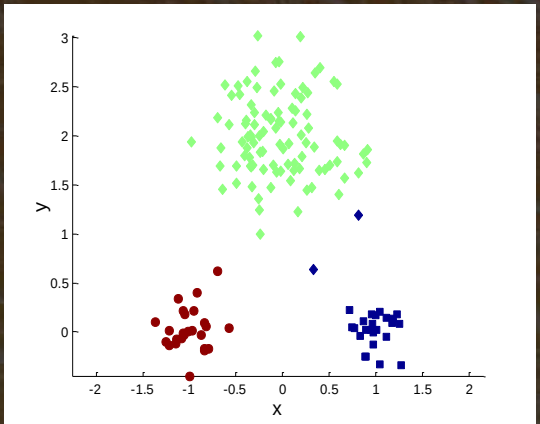
I=تعداد تکرار

d=تعداد ویژگی‌ها

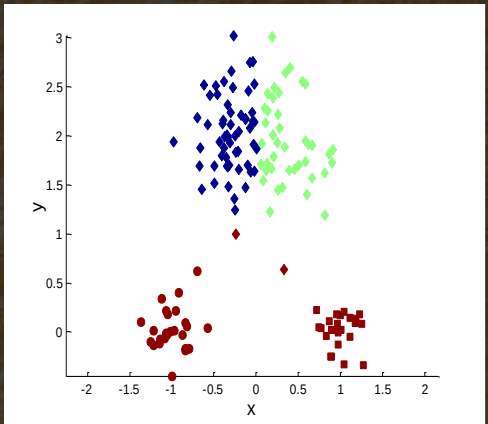
دو نوع مختلف خوشه بندی K-Means



Original Points

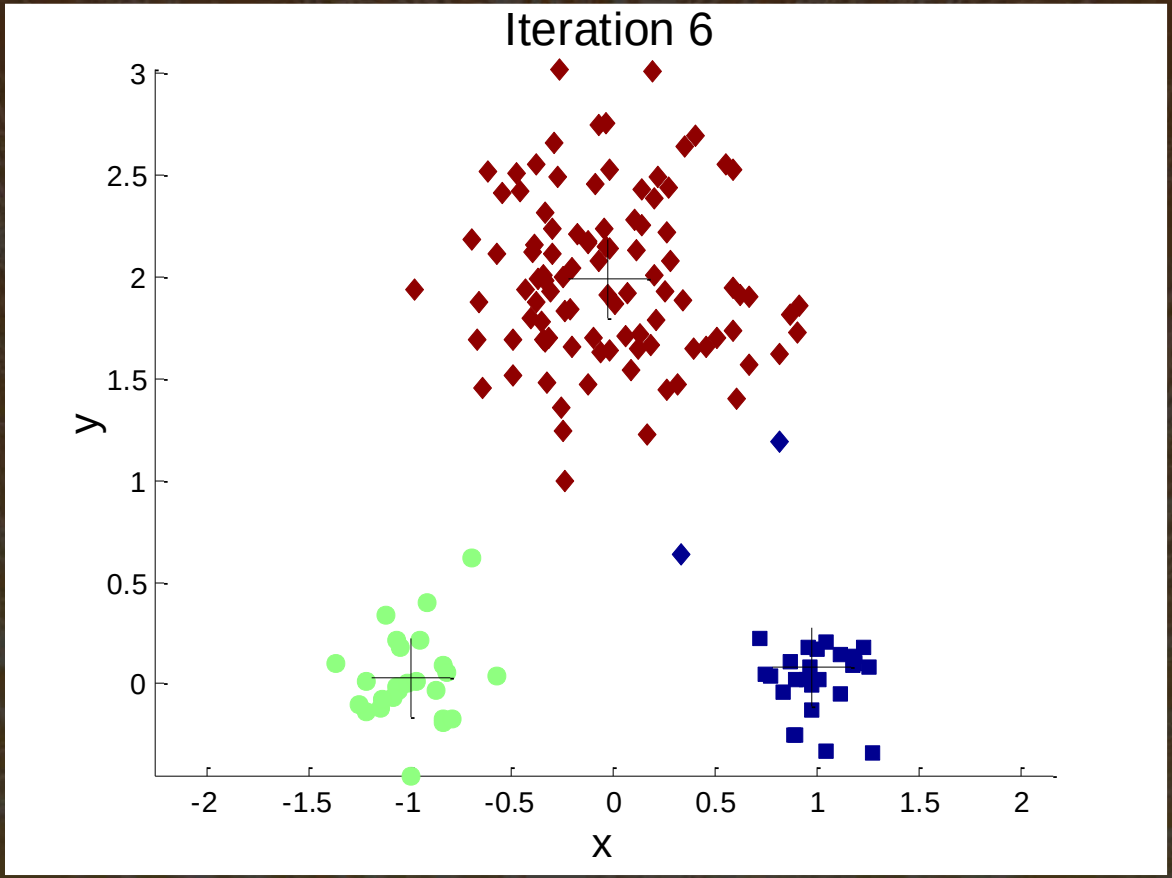


Optimal Clustering

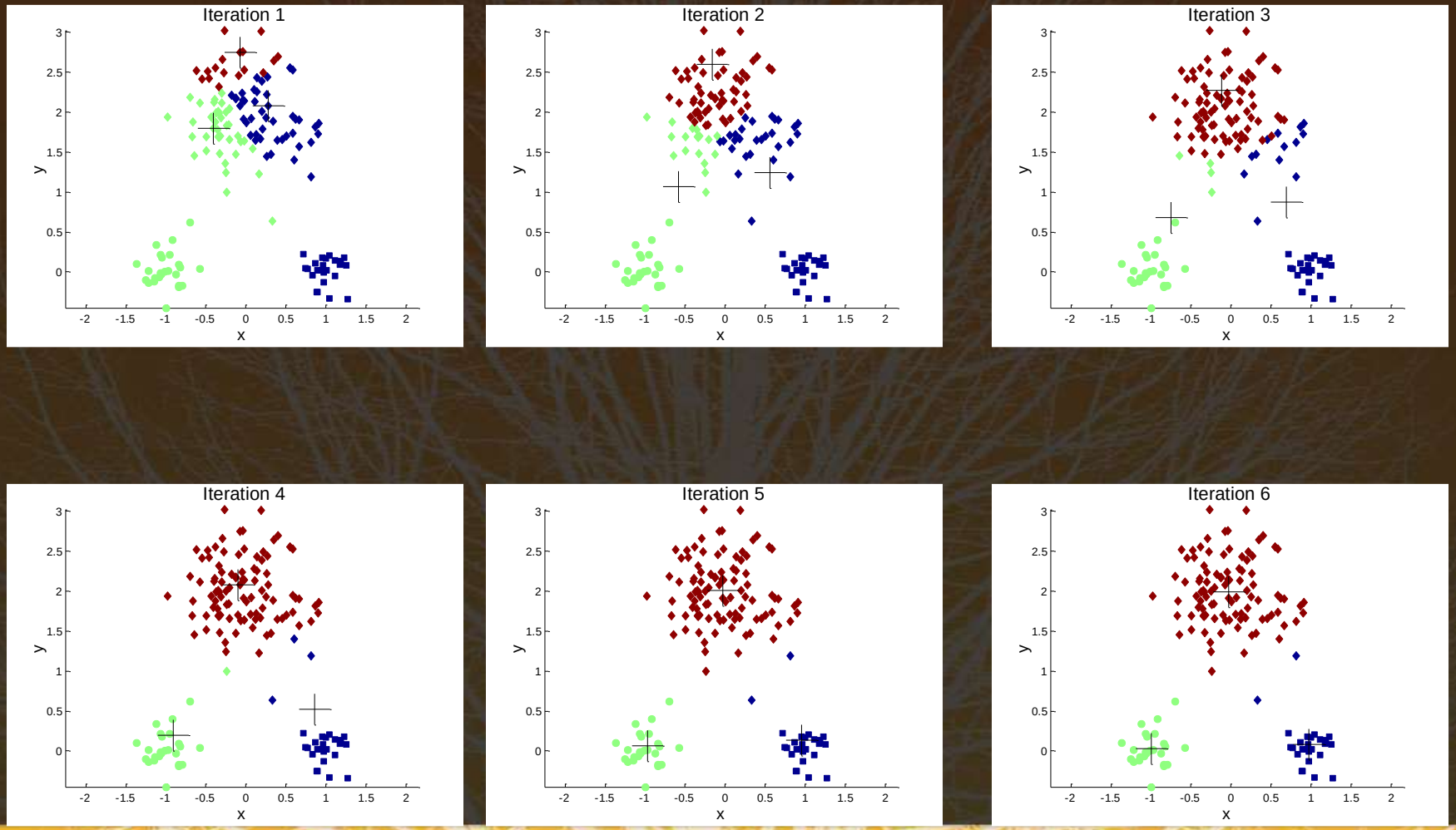


Sub-optimal Clustering

اهمیت انتخاب نقاط مرکز (Centroids) اولیه



اهمیت انتخاب نقاط مرکز (Centroids) اولیه



ارزیابی خوشه‌های k-means

- معمول‌ترین ارزیابی،مجموع مربعات خطا SSE (Sum of Squared Error) است.
  - برای هر نقطه، خطا، فاصله تا نزدیکترین خوشه است.
  - برای رسیدن به sse ما این خطاها را به توان دو می رسانیم و آنها را جمع می‌بندیم .

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

  - X یک نقطه داده ای در دسته ci است و mi نقطه مرکز برای خوشه ci است

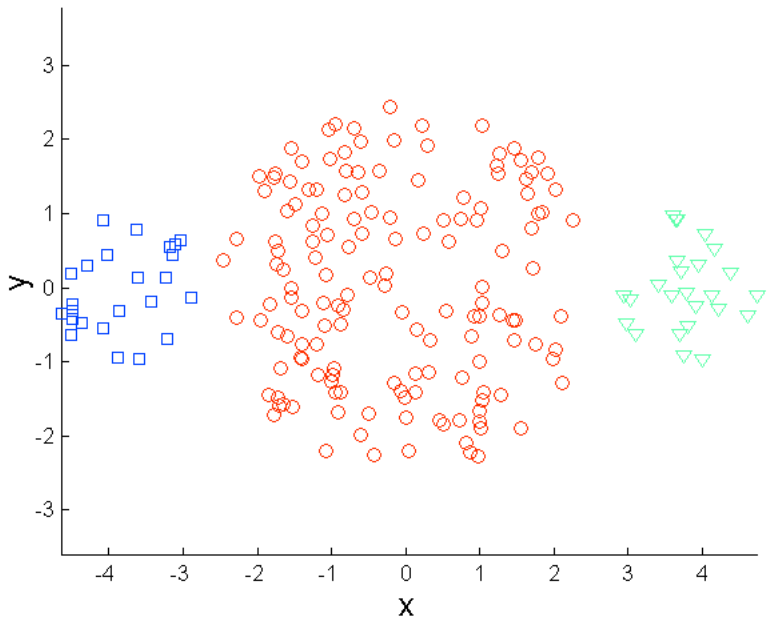
- می‌توان نشان داد که mi متناظر با مرکز (میانگین) دسته است .
- با داشتن دو خوشهه، می توانیم خوشه‌ای را که کوچکترین خطا را دارد انتخاب کنیم
- یک راه ساده برای کاهش SSE ، افزایش K یعنی تعداد خوشه‌ها است

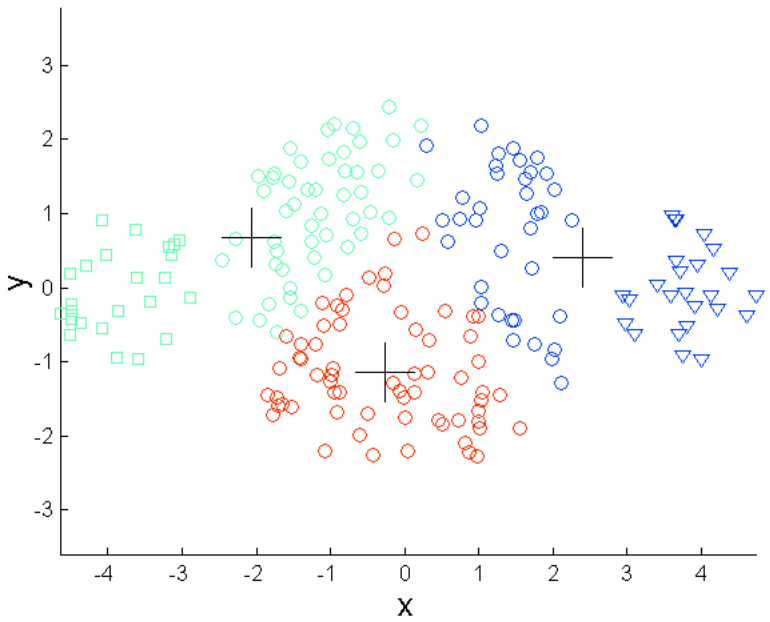
محدودیت های k-means

- K-means زمانی که خوشه‌ها در موارد زیر متفاوت باشند مشکل دارد
  - اندازه ها
  - تراکم ها (چگالی‌ها)
  - و شکل‌های غیر کروی

- روش k-means زمانی‌که داده‌ها حاوی داده‌های پرت باشند با مشکل مواجه می‌شود

اندازه‌های متفاوت

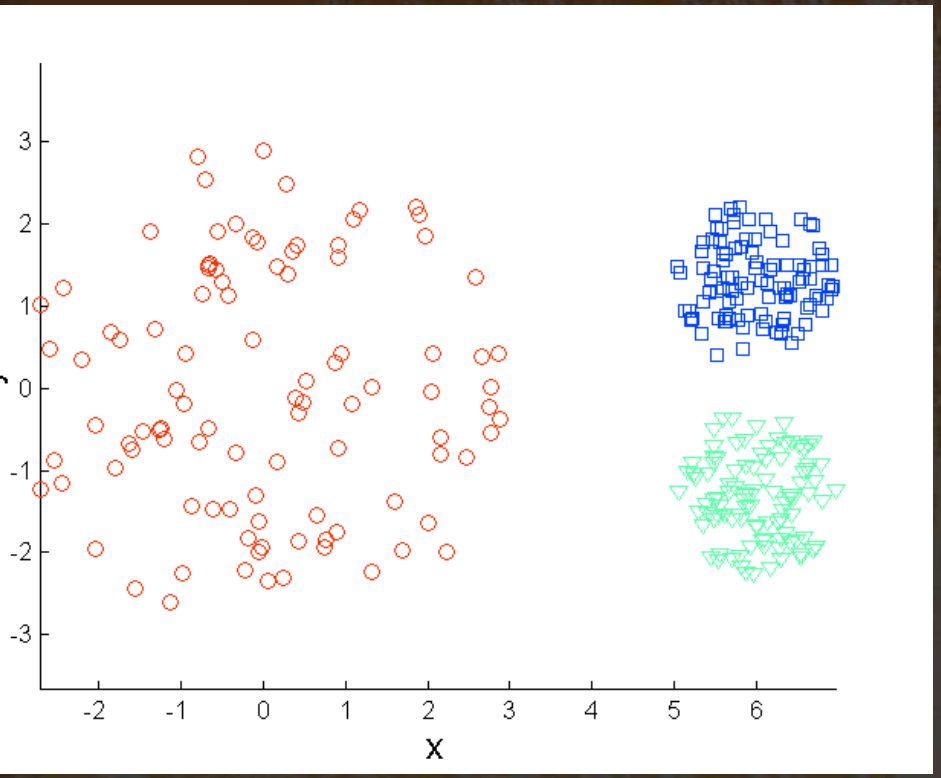




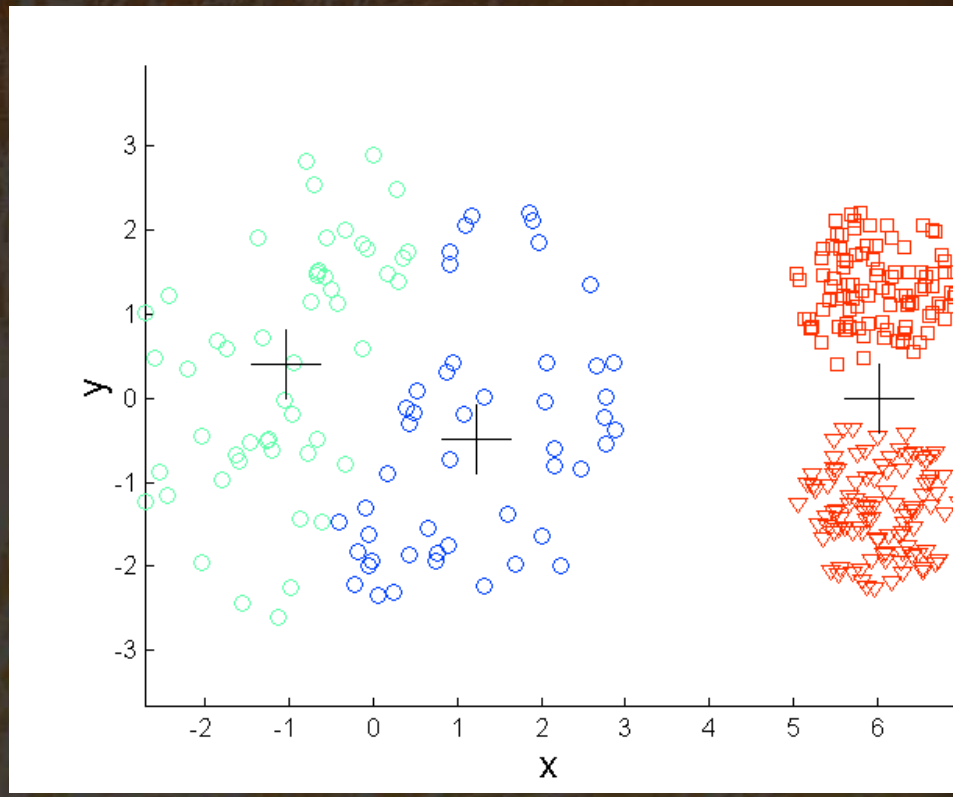
نقاط اصلی

K-means (3 clusters)

چگالی‌های متفاوت

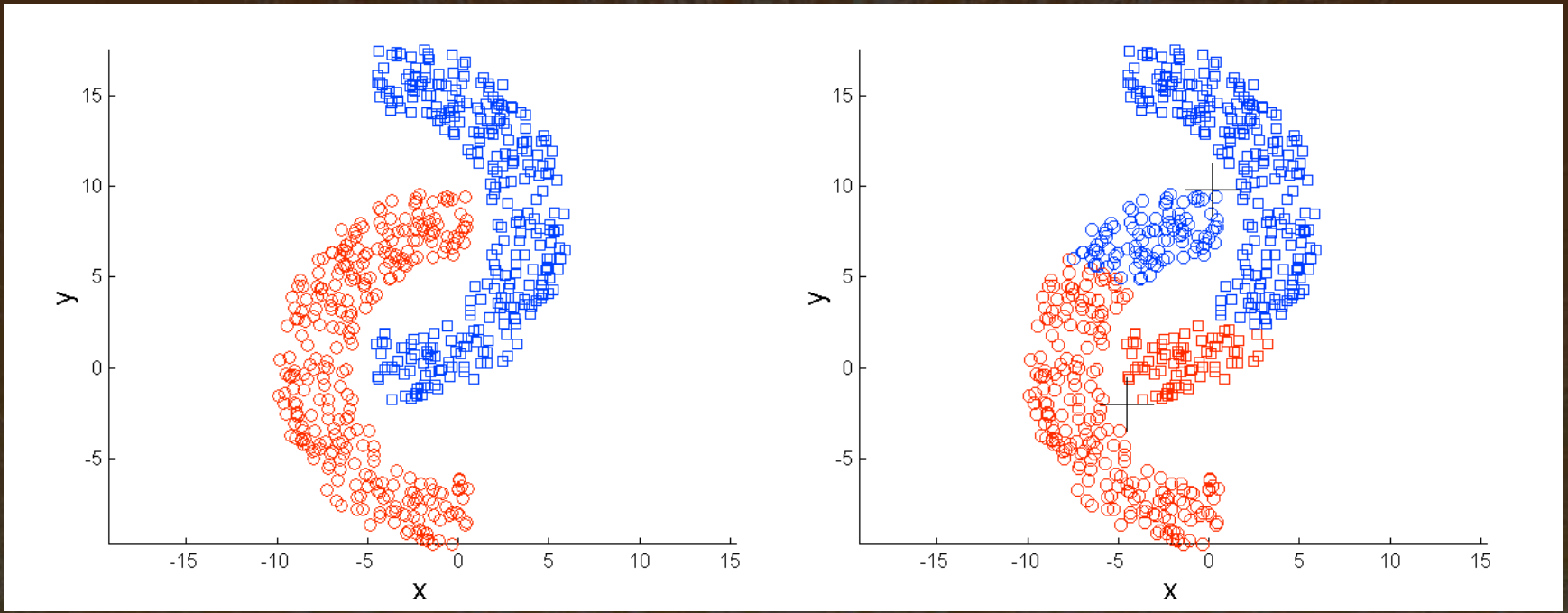


نقاط اصلی



K-means (3 clusters)

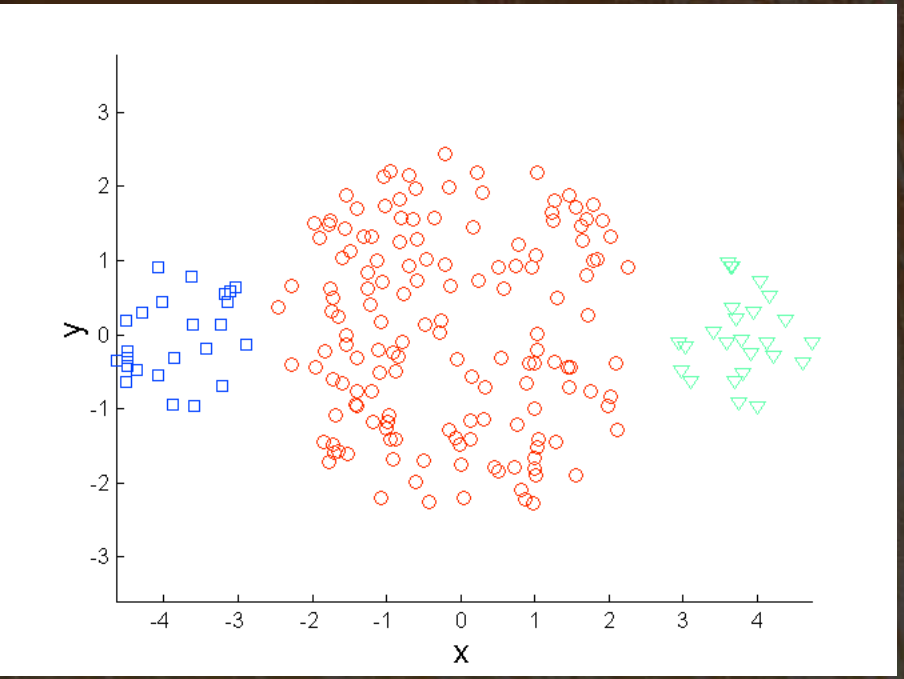
شکل‌های غیر کروکی

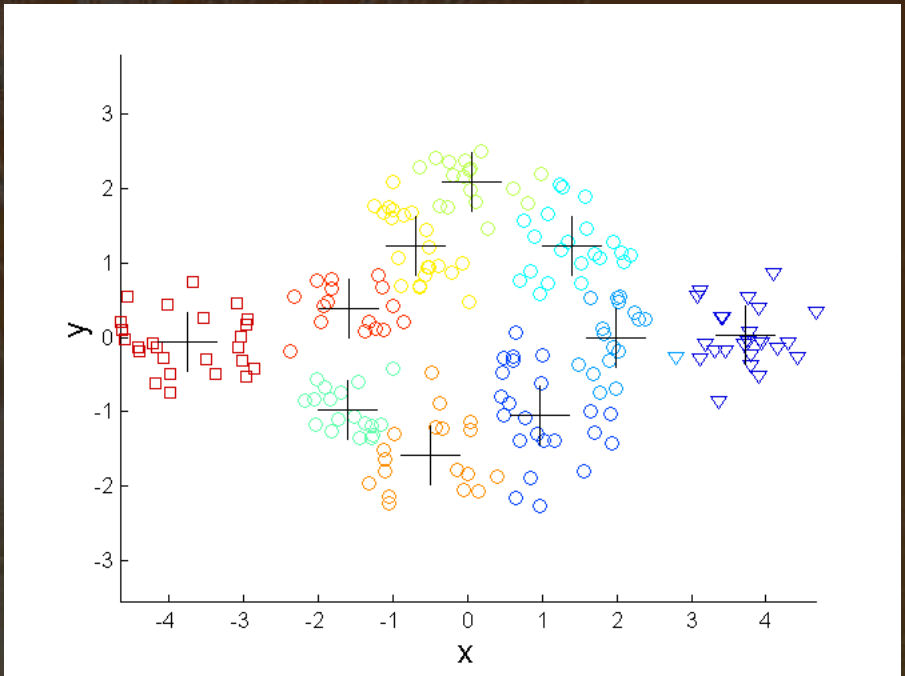


نقاط اصلی

K-means (2 clusters)

غلبه بر محدودیت های k-means





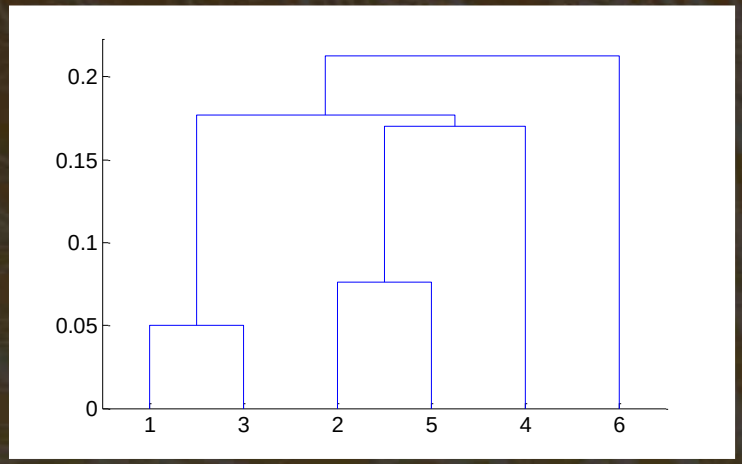
اصلی نقاط

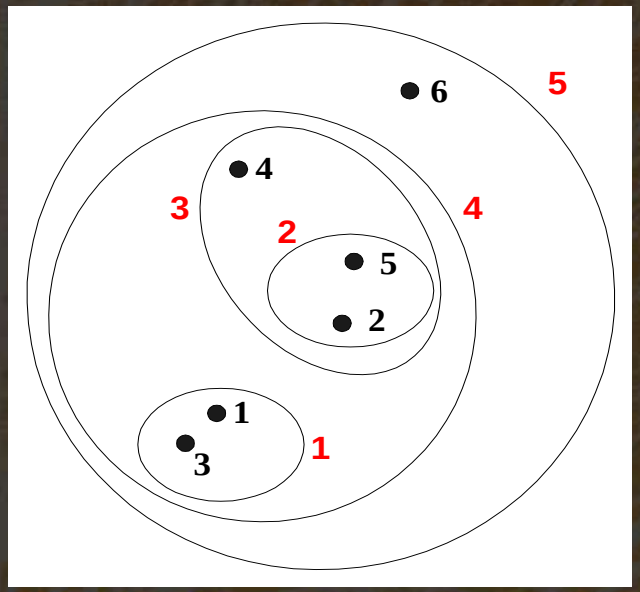
k-means Clusters

یک راه حل ، استفاده از دسته های زیاد است

خوشه‌بندی سلسله مراتبی

- تولید مجموعه‌ای از خوشه‌های تو در تو سازمان‌یافته به عنوان یک درخت سلسله مراتبی
- می توان بوسیله یک dendrogram مصورسازی نمود.
- dendogram، توالی ادغام و یا انشعابات را ثبت می‌کند.





نقاط قوت خوشه‌بندی سلسله مراتبی

- نباید هر تعداد خاصی از خوشه را فرض کنیم.
- هر تعداد مورد نظر از خوشه را می‌توان با برش dendogram در سطح مناسب به دست آورد.
- آنها ممکن است با طبقه‌بندی معنی‌داری با هم مرتبط باشند.
- به عنوان مثال در علوم زیستی (به عنوان مثال، قلمرو حیوانات، بازسازی نژادی جانور یا گیاه، ...)

خوشه‌بندی سلسله مراتبی

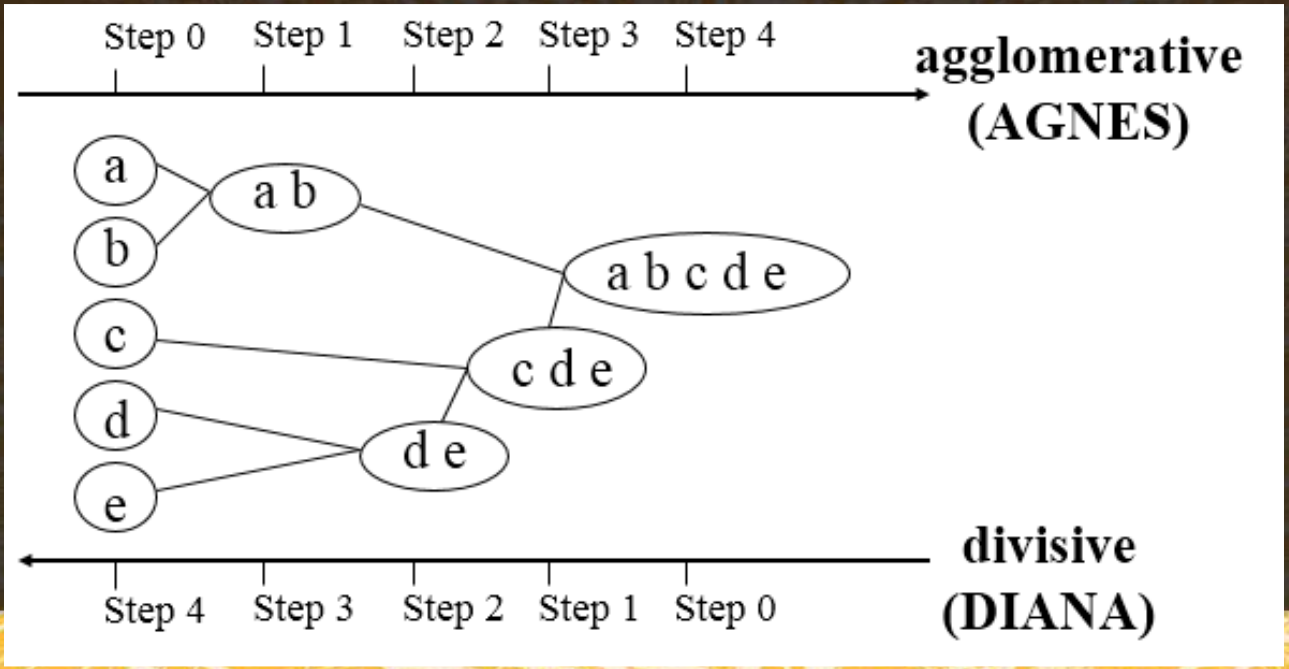
- دو نوع اصلی خوشه‌بندی سلسله مراتبی
- Agglomerative:
- شروع با نقاط خوشه‌های فردی
- در هر مرحله، ادغام نزدیکترین جفت خوشه تا تنها یک خوشه (یا K خوشه) باقی بماند.
- Divisive:
- شروع با یک خوشه all-inclusive
- در هر مرحله، تقسیم یک خوشه تا هر خوشه شامل یک نقطه (یا K خوشه) گردد.

خوشه‌بندی سلسله مراتبی

- الگوریتم‌های سلسله مراتبی سنتی از یک ماتریس شباهت یا فاصله استفاده می‌کنند.
- ادغام و یا تقسیم یک خوشه در یک زمان صورت می‌گیرد.

خوشه‌بندی سلسله مراتبی

- استفاده از ماتریس فاصله به عنوان معیار خوشه: این روش تعداد خوشه K به عنوان ورودی نیاز ندارد، اما نیاز به یک شرط خاتمه دارد.



الگوریتم خوشه بندی Agglomerative

- تکنیک محبوب خوشه‌بندی سلسله مراتبی
- الگوریتم پایه عبارتند از:

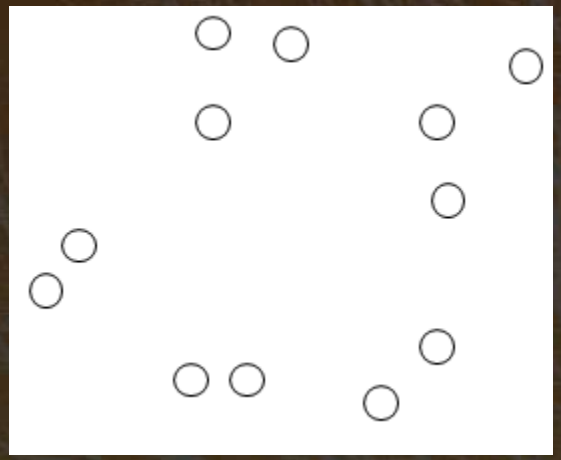
1. Compute the proximity matrix
2. Let each data point be a cluster
3. **Repeat**
4. Merge the two closest clusters
5. Update the proximity matrix
6. **Until** only a single cluster remains

الگوریتم خوشه بندی Agglomerative

- عملیات کلیدی، محاسبه تقارب دو خوشه است.
- روش‌های مختلفی برای تعریف فاصله بین خوشه تشخیص در الگوریتم‌های مختلف وجود دارد.

حالت شروع

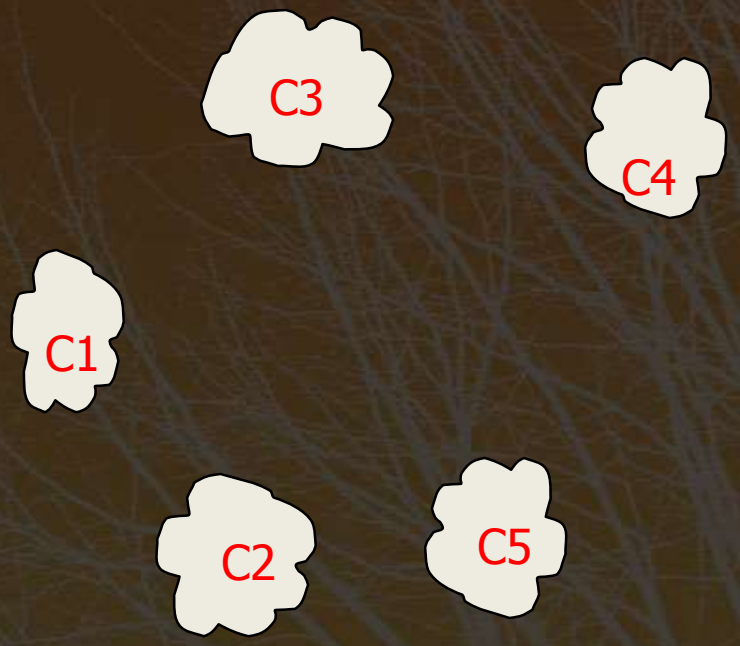
- شروع با خوشه‌ای از نقاط منحصر به فرد و یک ماتریس مجاورت



	p1	p2	p3	p4	p5	. .
p1						
p2						
p3						
p4						
p5						

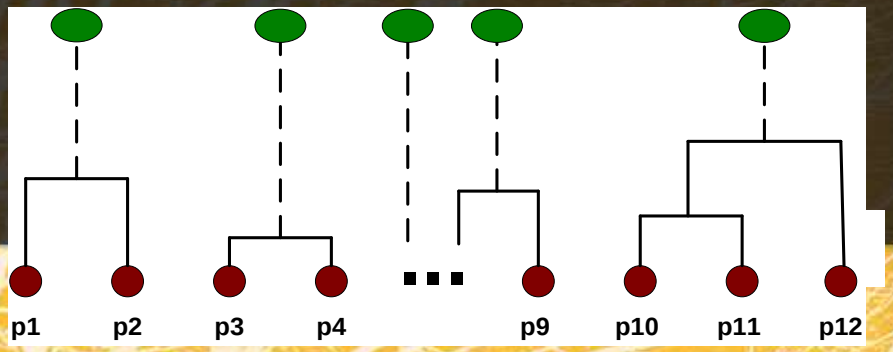
حالت متوسط

- پس از انجام چند عملیات ادغام ما چند خوشه خواهیم داشت.



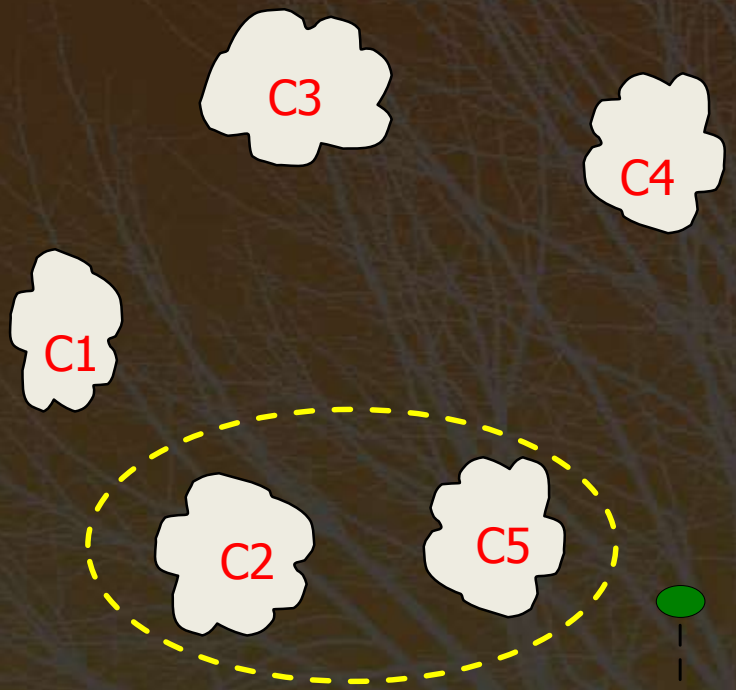
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

ماتریس مجاورت



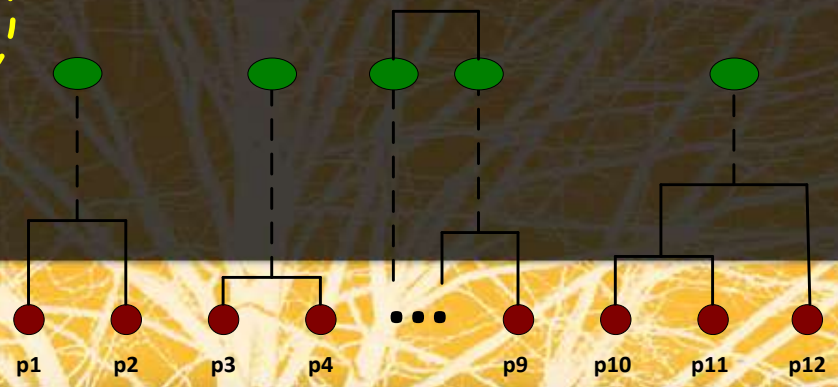
حال متوسط

- ما نیاز به ادغام دو خوشه نزدیک (C2 and C5) و بروزرسانی ماتریس مجاورت داریم.



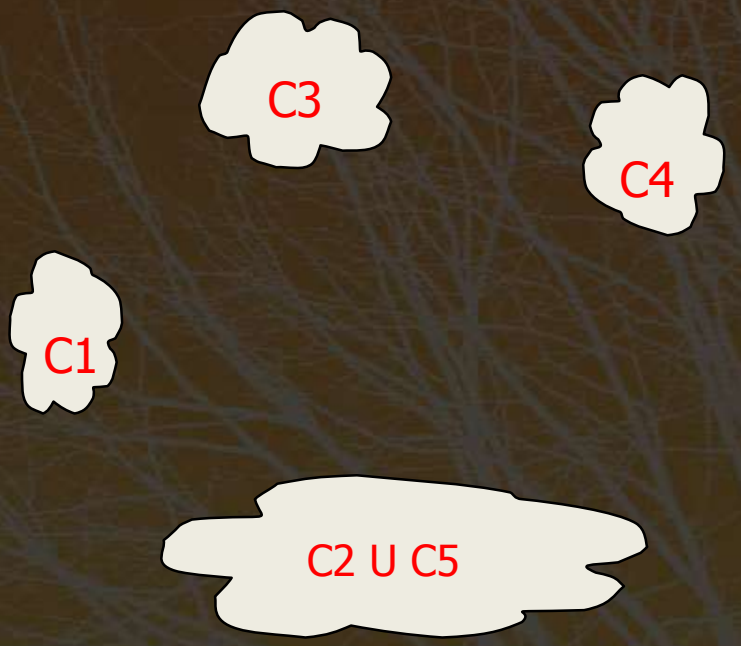
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

ماتریس مجاورت



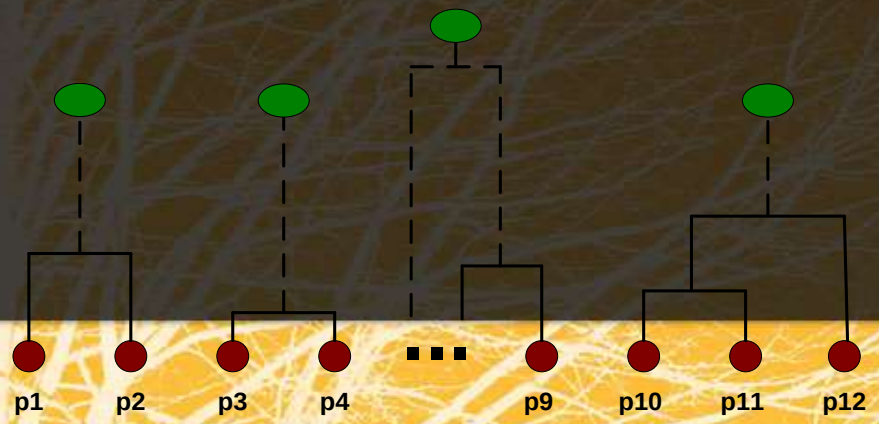
بعد از ادغام

- سوال این است که ماتریس مجاورت چگونه بروزرسانی می‌شود

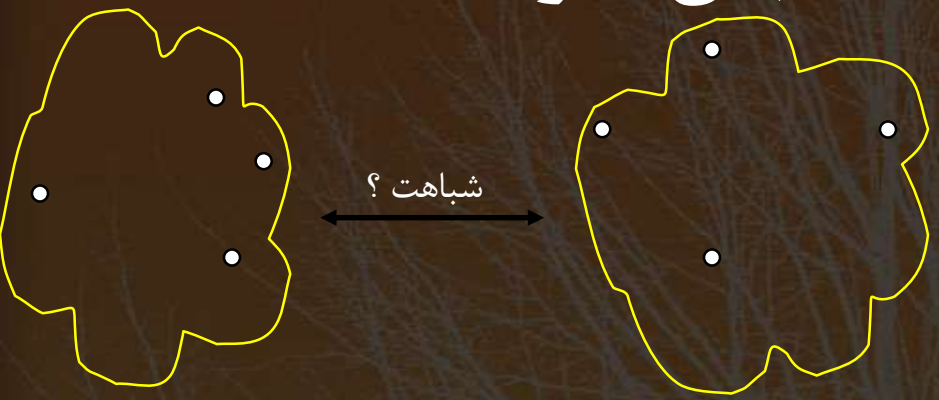


	C1		C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix



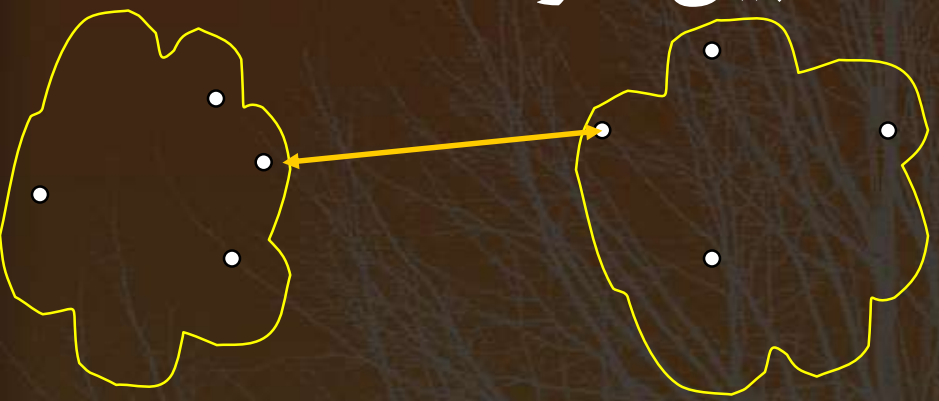
نحوه‌ی تعیین شباهت بین خوشه‌ها



- کمترین
- بیشترین
- میانگین گروه
- فاصله بین مرکز

	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

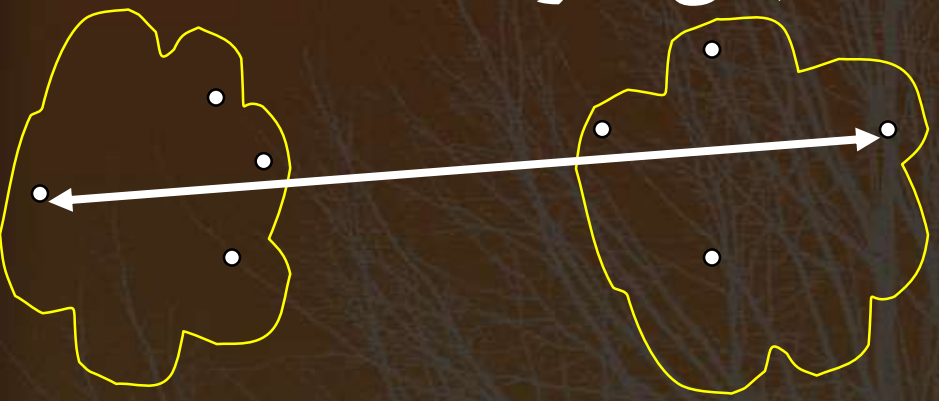
نحوه‌ی تعیین شباهت بین خوشه‌ها



	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

- کوچکترین:
  - کوچکترین فاصله در بین عناصر یک خوشه و عناصر دیگر به عنوان مثال: dist(Ki, Kj) = min(tip, tjq)
- بیشترین
- میانگین گروه
- فاصله بین مرکز

نحوه‌ی تعیین شباهت بین خوشه‌ها

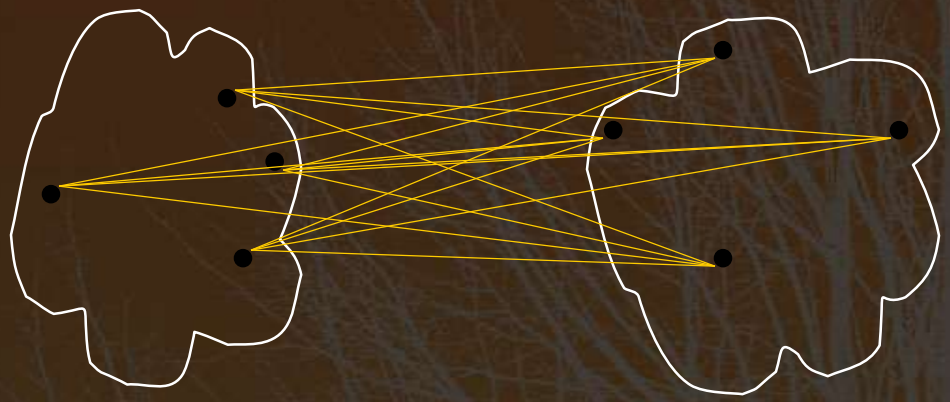


	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

- کوچکترین
- بیشترین:
  - کوچکترین فاصله در بین عناصر یک خوشه و عناصر دیگر به عنوان مثال:
  - dist(Ki, Kj) = max(tip, tjq)
- میانگین گروه
- فاصله بین مرکز

46

نحوه‌ی معیین شدن شباهت بین خوشه‌ها



	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

- کوچکترین
- بیشترین
- میانگین گروه:
  - میانگین فاصله در بین عناصر یک خوشه و عناصر دیگر به عنوان مثال:
  - dist(Ki, Kj) = avg(tip, tjq)
- فاصله بین مرکز

نحوه‌ی معیین شدن شباهت بین خوشه‌ها

	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

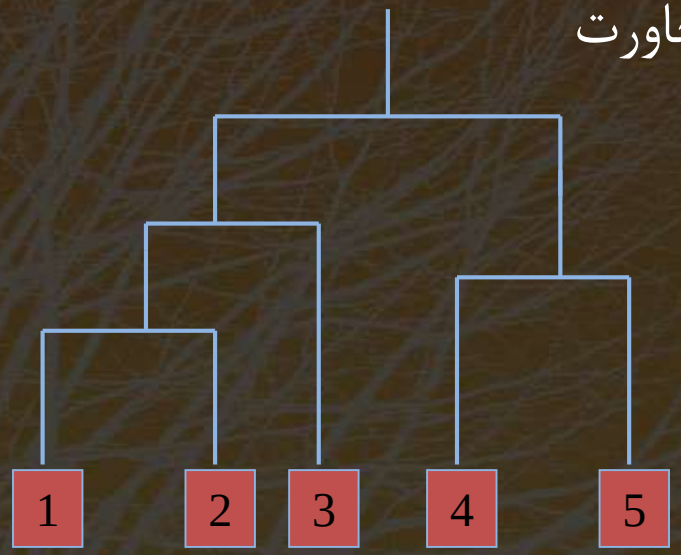
- کوچکترین
- بیشترین
- میانگین گروه
- فاصله بین مرکز:
- فاصله بین مرکز از دو خوشه مانند :

– dist(Ki, Kj) = dist(Ci, Cj)

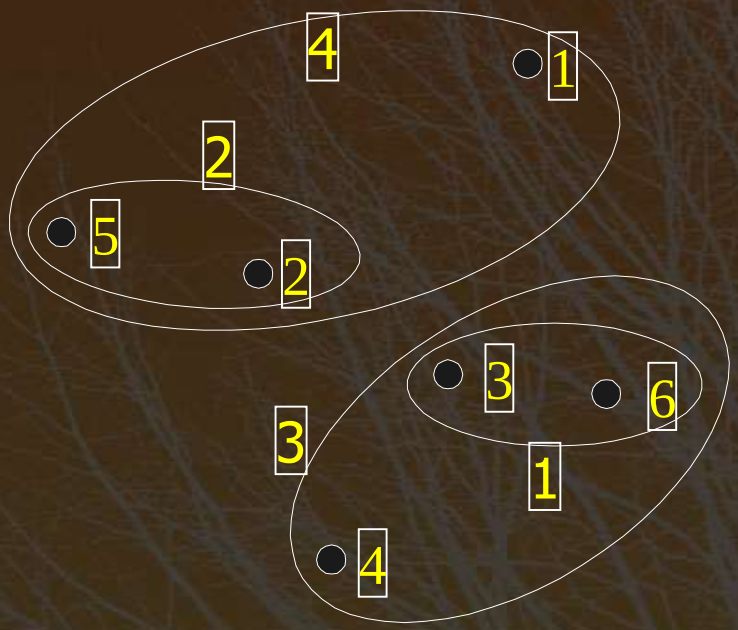
شباهت خوشه: MAX یا ارتباط کامل

- شباهت دو خوشه بر اساس (نزدیکترین) دو نقطه مشابه‌تر در دسته‌های مختلف
- لینک واحد: کوچکترین فاصله بین عنصر موجود در یک خوشه با عنصر خوشه‌ی دیگر به طور مثال dist(Ki, Kj) = min(tip, tjq)
  - تعیین شده توسط یک جفت از نقاط به طور مثال توسط یک لینک واحد در گراف مجاورت

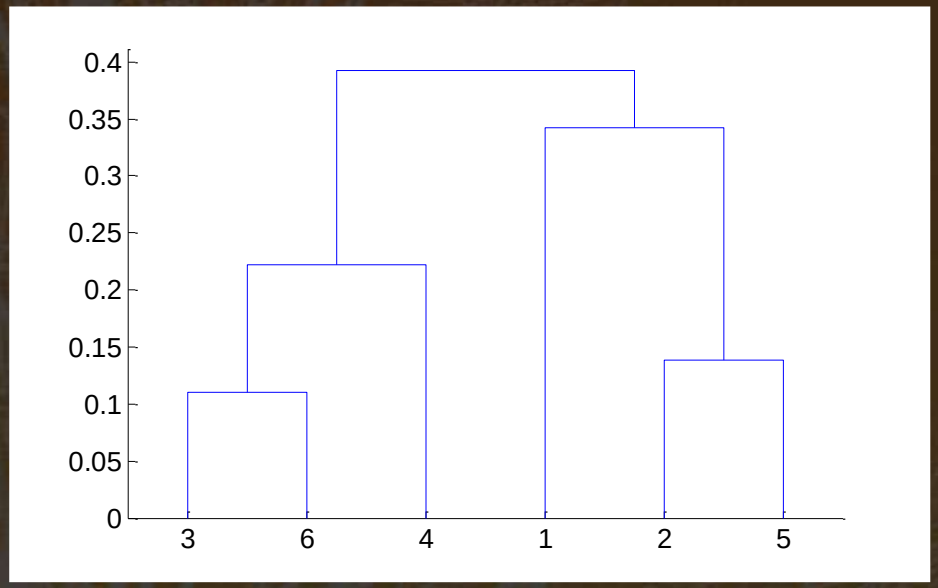
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



خوشه بندی سلسله مراتبی : MAX



خوشه های تودرتو

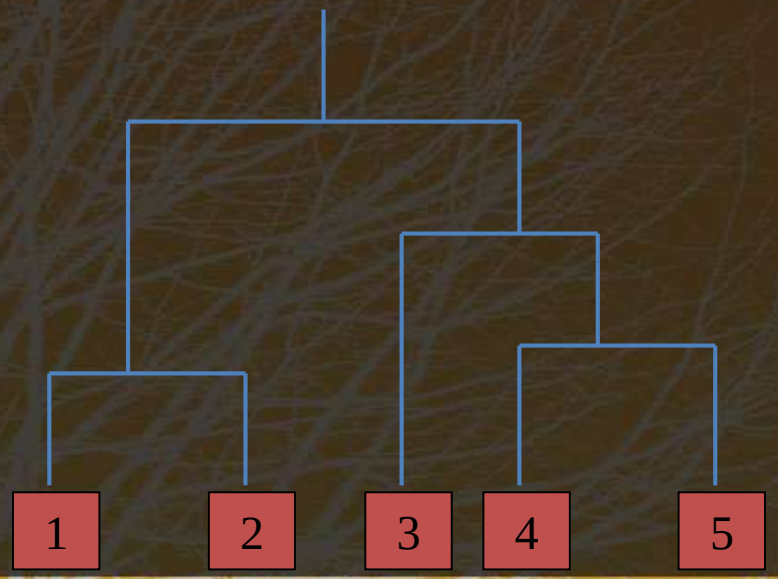


Dendrogram

شباهت خوشه‌ها: گروه میانگین

- شباهت دو خوشه براساس (دورترین) دو نقطه کمتر مشابه در دسته‌های مختلف
- لینک کامل: بیشترین فاصله بین عنصر یک خوشه با عنصر دیگر خوشه‌ها به طور مثال: dist(Ki, Kj) = max(tip, tjq)
  - با استفاده از تمام جفت نقاط در دو خوشه مشخص می‌شوند.

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



دانشگاه اصفهان

پایان

پیروز و عاقبت به خیر باشید