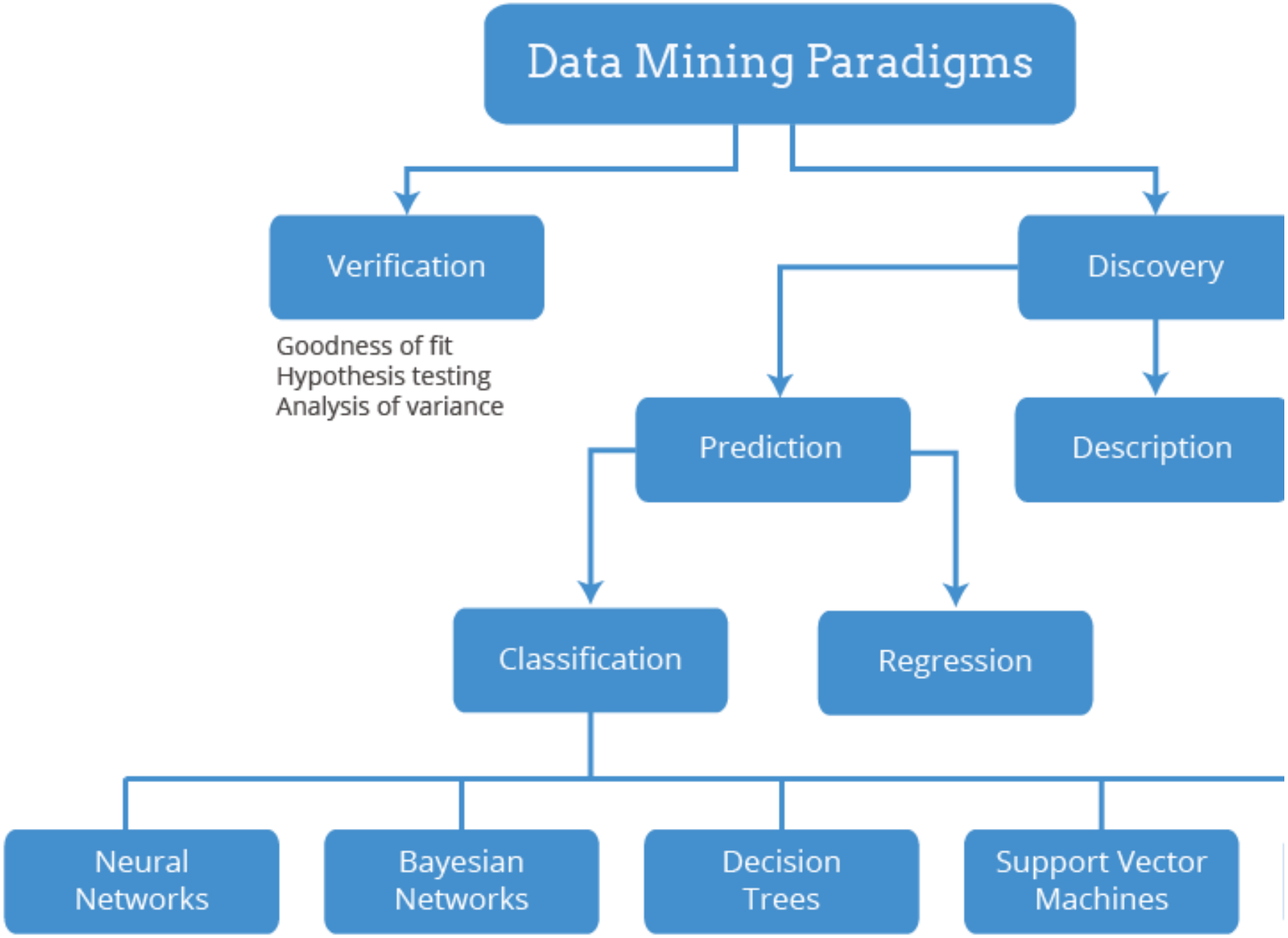

Data Mining Paradigms
Verification
Goodness of fit
Hypothesis testing
Analysis of variance
Discovery
Prediction
Description
Classification
Regression
Neural Networks
Bayesian Networks
Decision Trees
Support Vector Machines

Clustering

تحلیل خوشه‌ای

تحلیل خوشه‌ای یک تکنیک کاهش داده است. در این تکنیک تعداد بسیار زیادی از مشاهدات را می‌توان به تعداد بسیار کمتری از خوشه‌ها کاهش داد. در این خوشه‌ها، می‌بایست بیشترین شباهت در درون خوشه‌ها و کمترین شباهت در بین خوشه‌ها حاصل شود.

تعریف خوشه: خوشه مجموعه ای از اشیاء می‌باشد که در آن اشیاء با یکدیگر مشابه بوده و با اشیاء موجود در خوشه های دیگر غیر مشابه می‌باشند.

کاربردها

- بهداشت و درمان،
- فنی و مهندسی،
- علوم اجتماعی و انسانی
- بازاریابی

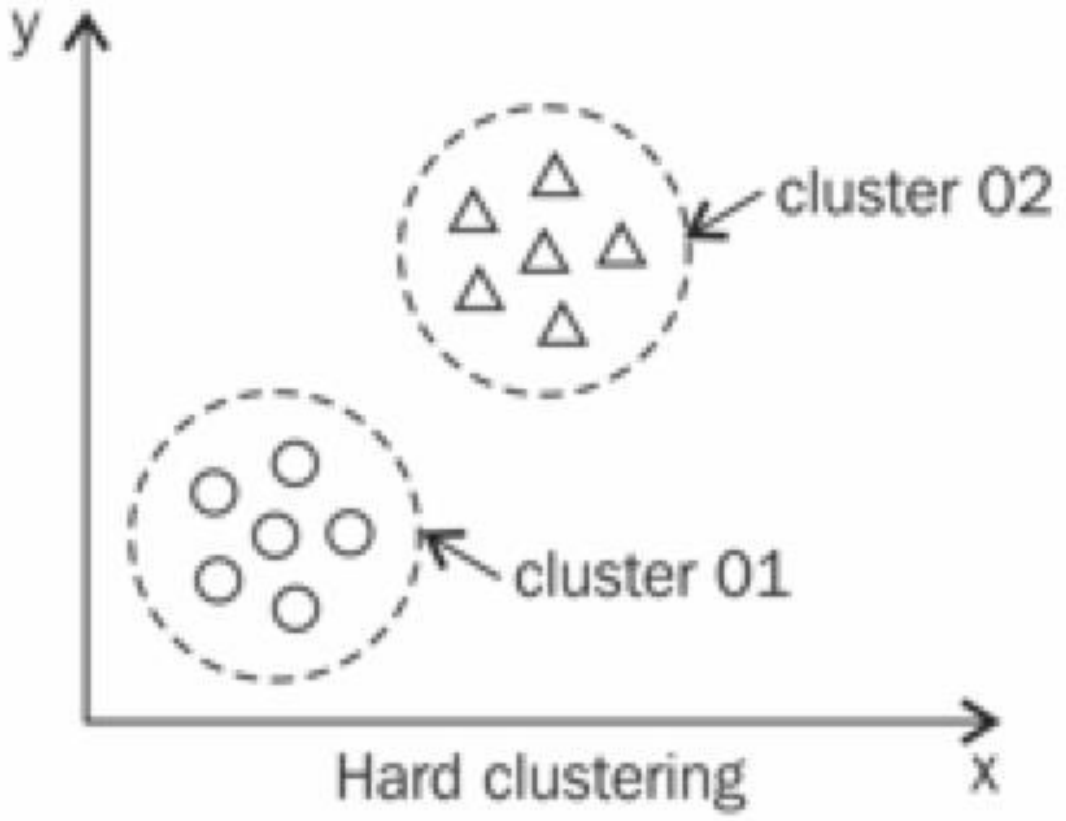
به عنوان مثال محققان بازاریابی از تحلیل خوشه‌ای به‌عنوان یک استراژی **تقسیم‌بندی مشتری** استفاده می‌کنند. مشتریان براساس شباهت رفتارهای خریدشان در خوشه‌های مربوطه قرار خواهند گرفت.

روش‌های خوشه‌بندی

➢ خوشه‌بندی سخت

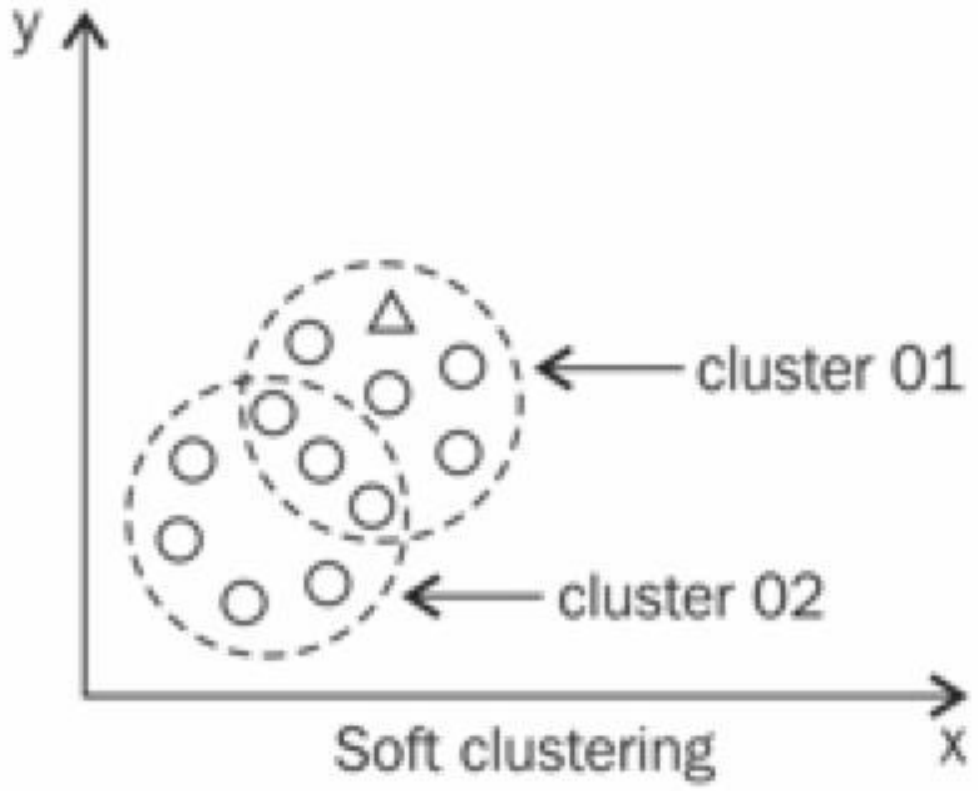
مانند:

- خوشه‌بندی سلسله مراتبی (hierarchical)
- خوشه‌بندی تفکیکی (partitioning)



➢ خوشه‌بندی نرم(soft)

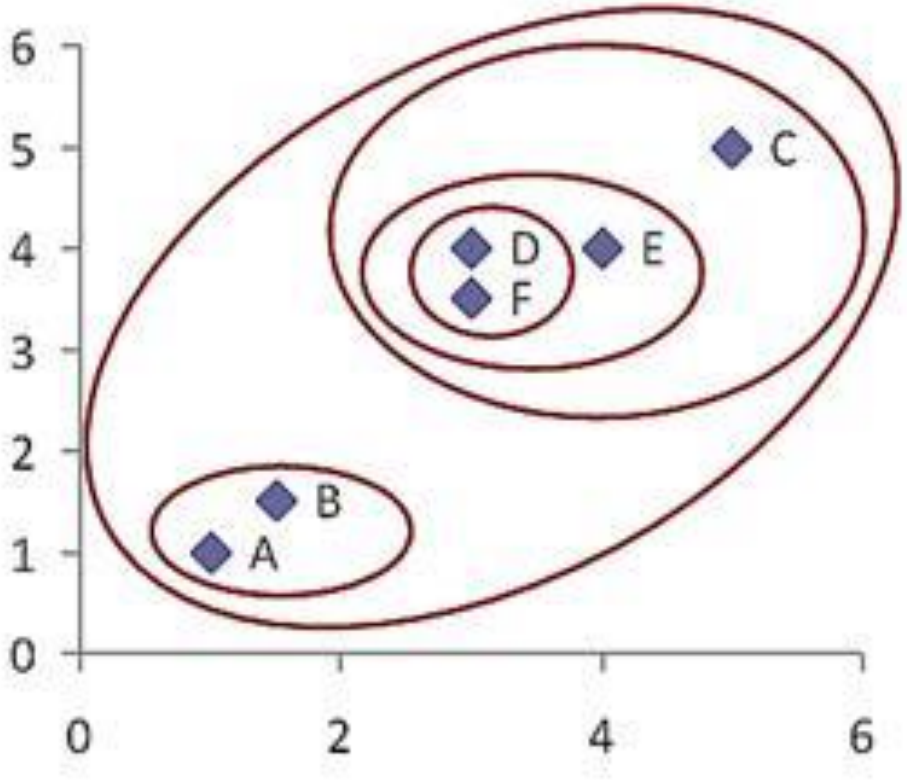
مانند خوشه‌بندی فازی

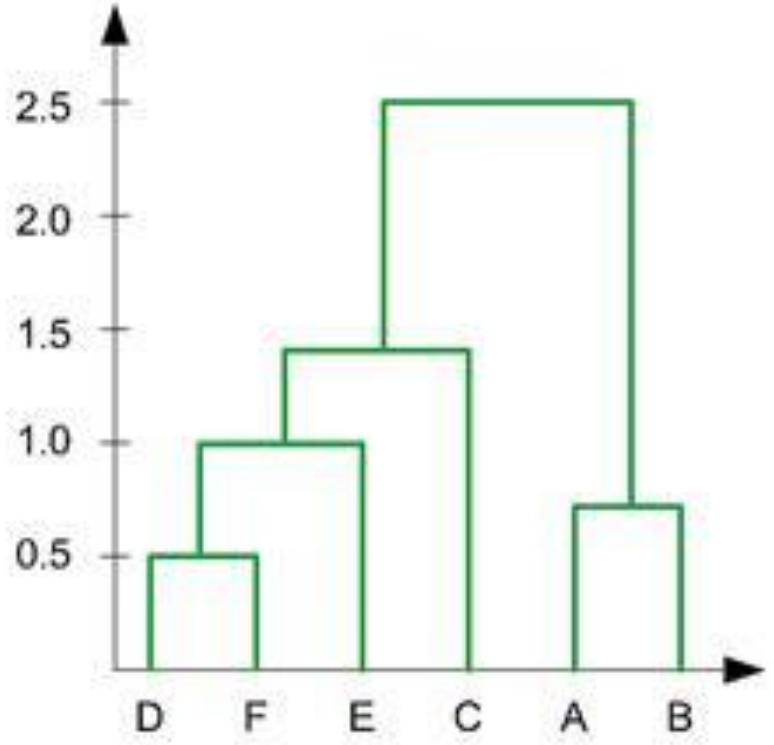


خوشه‌بندی سلسله مراتبی (hierarchical)

خوشه‌بندی تودرتو

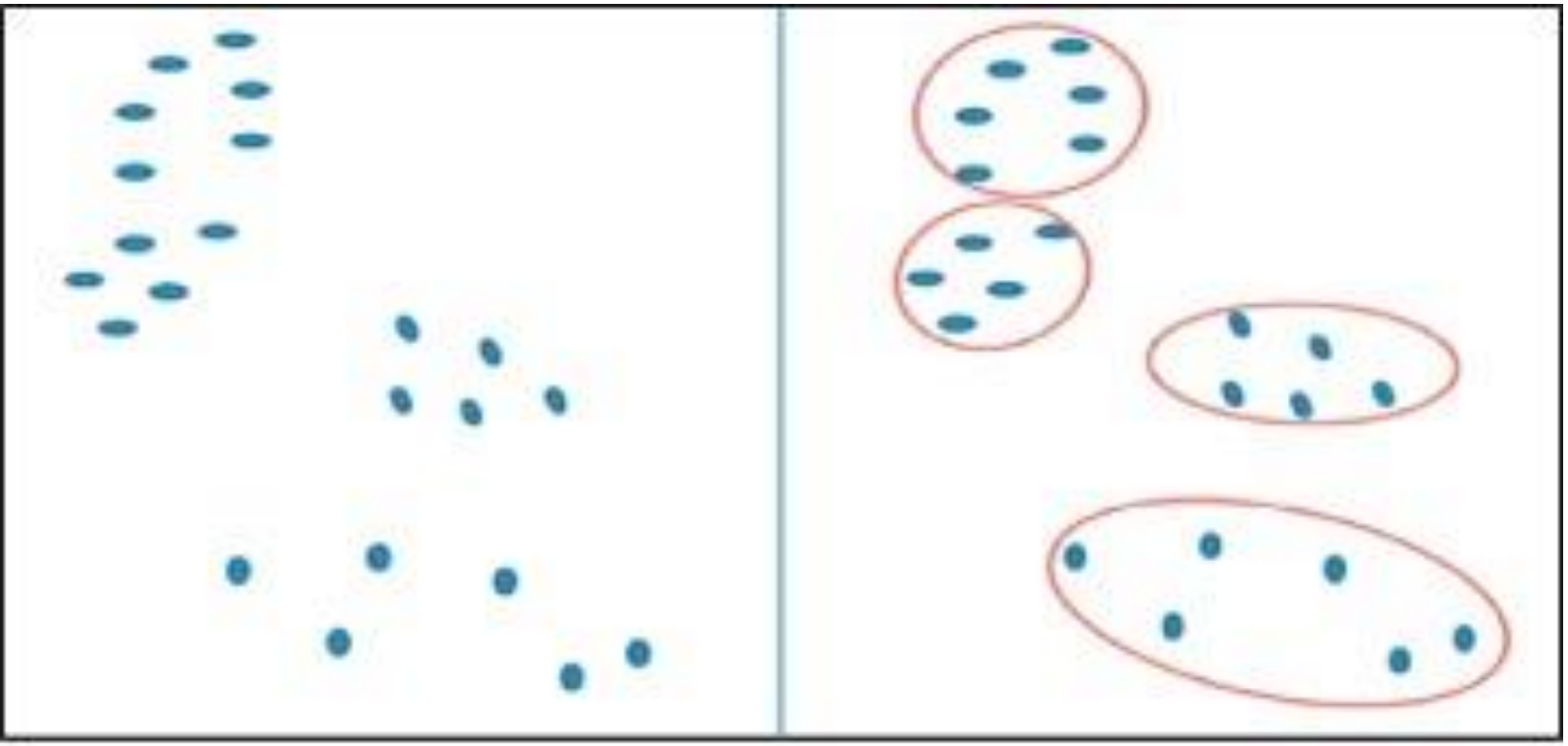
عدم تغییربرچسب‌ها در طی مراحل خوشه‌بندی





خوشه‌بندی تفکیکی (partitioning)

خوشه‌بندی بصورت گروه‌های مجزا



Data points

Partitional clusters

خوشه‌بندی تفکیکی را می‌توان هم در خوشه‌بندی سخت (مانند $k-means$) و هم در خوشه‌بندی نرم (مانند $c-means$) بکاربرد.

معیار تشابه در داده‌های کمی

Distance functions

Euclidean

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Manhattan

$$\sum_{i=1}^{k}|x_i - y_i|$$

Minkowski

$$\left(\sum_{i=1}^{k}\left(|x_i - y_i|\right)^q\right)^{1/q}$$

Chebyshev

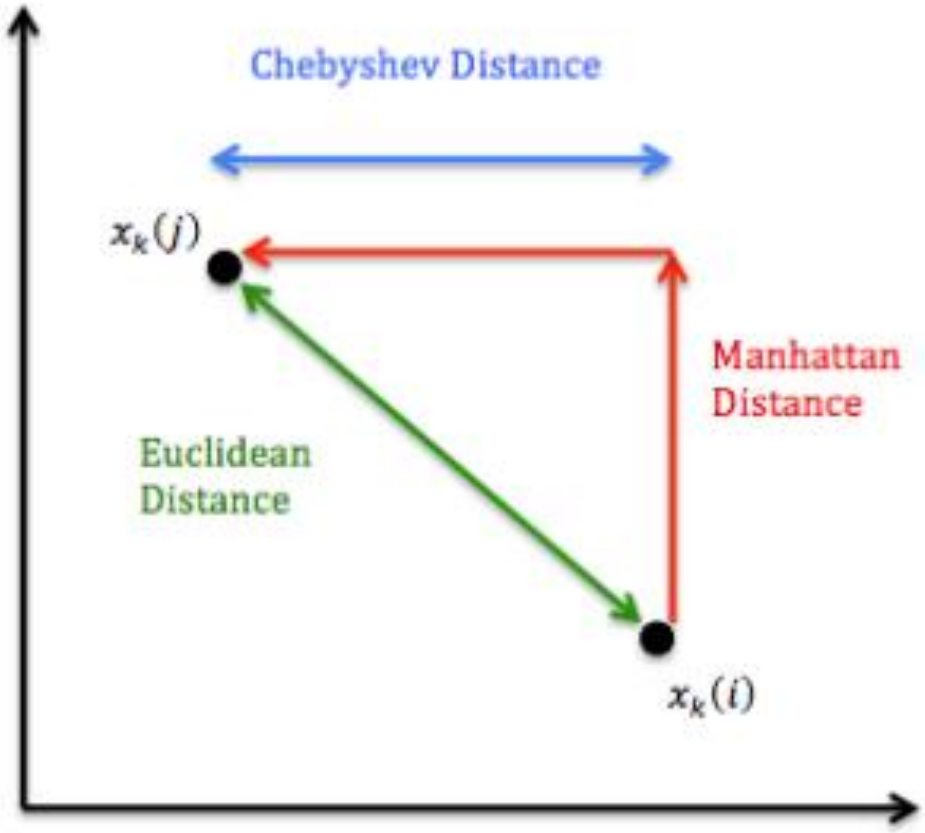
$$D_{ij} = max|x_{ik} - x_{jk}|$$

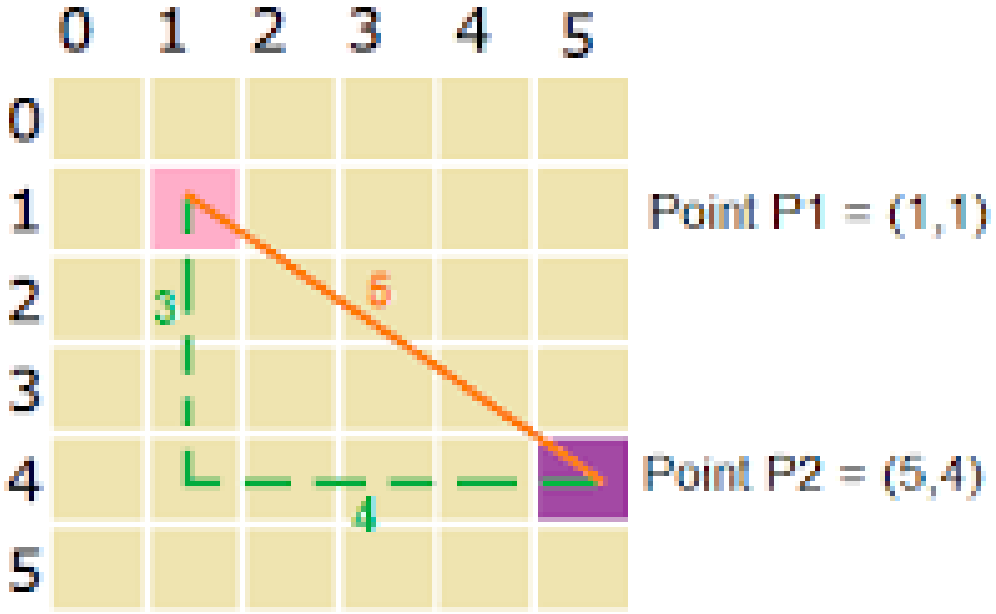
معیار تشابه در داده‌های دودویی $(Simple\ Maching)$

$$SM = \frac{c}{T}$$

تعداد صفات منطبق

تعداد کل صفات


Chebyshev Distance
$x_k(j)$
Manhattan
Distance
Euclidean
Distance
$x_k(i)$

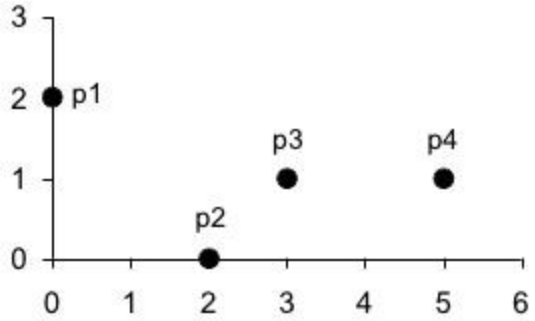


Euclidean distance = $\sqrt{(5-1)^2 + (4-1)^2} = 5$

Manhattan distance = $|5-1| + |4-1| = 7$

ماتریس فاصله

Euclidean Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

62

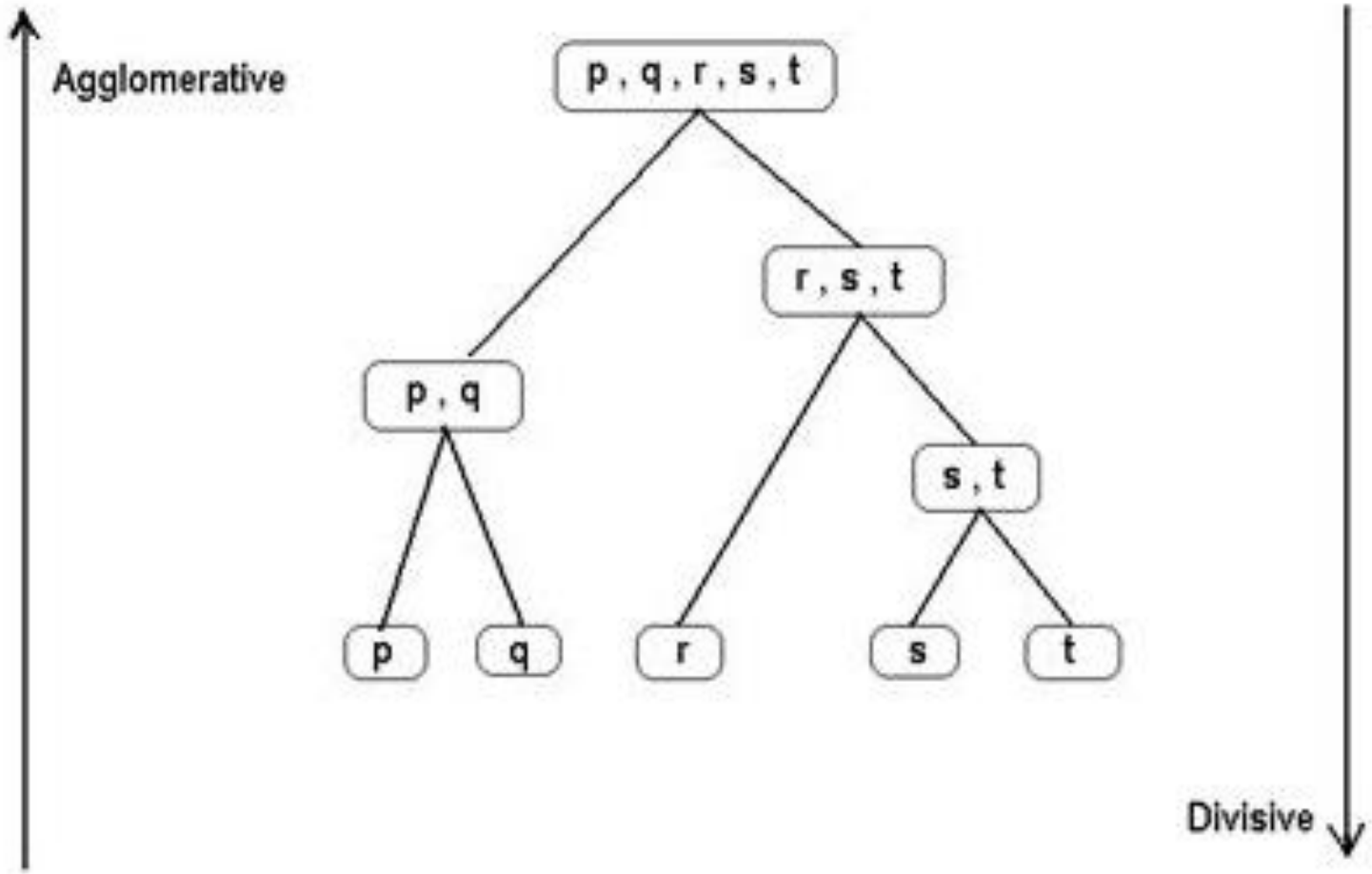
خوشه‌بندی سلسله مراتبی (hierarchical)

۱. پایین به بالا $(Bottom - Up)$ یا تجمیعی $(Agglomerative)$

در این روش ابتدا تعداد خوشه‌ها برابر با تعداد داده‌ها در نظرگرفته می‌شود و در طی فرایندی تکراری در هر مرحله خوشه‌هایی که شباهت بیشتری با یکدیگر دارند با هم ترکیب می‌شوند تا در نهایت یک خوشه که شامل همه‌ی داده‌ها می‌باشد، حاصل شود.

۲. بالا به پایین $(Top - Down)$ یا تقسیم کننده :$(Divisive)$

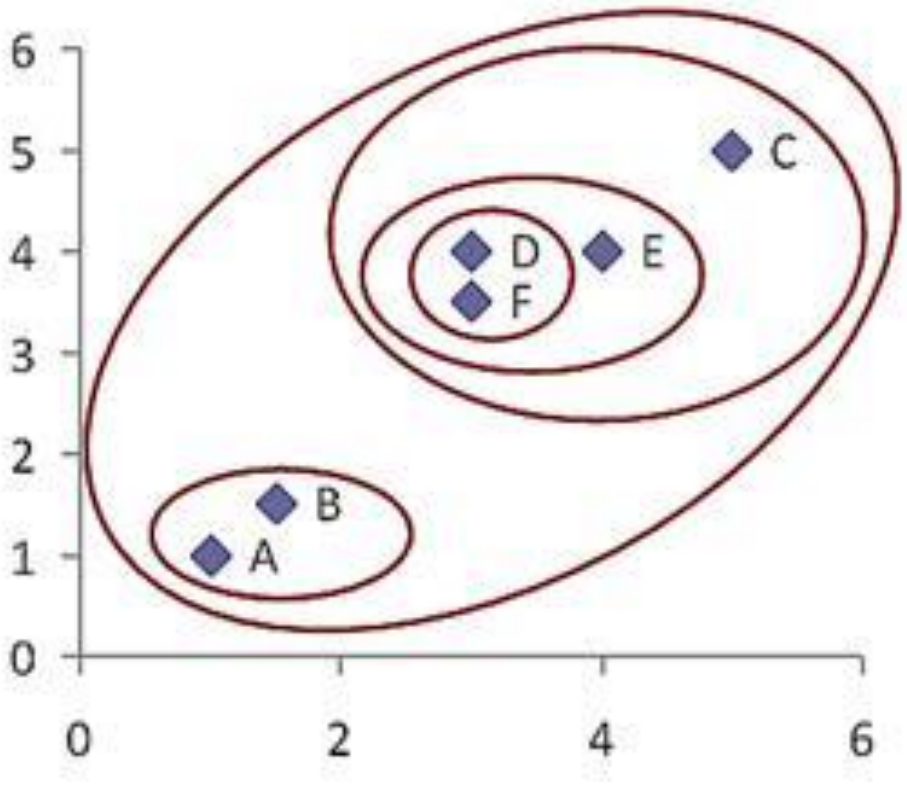
در این روش ابتدا تمام داده‌ها به عنوان یک خوشه در نظر گرفته می‌شوند و سپس در طی یک فرایند تکراری در هر مرحله داده‌هایی که شباهت کمتری به هم دارند به خوشه‌های مجزایی شکسته می‌شوند و این روال تا رسیدن به خوشه‌هایی که دارای یک عضو هستند ادامه پیدا می‌کند.

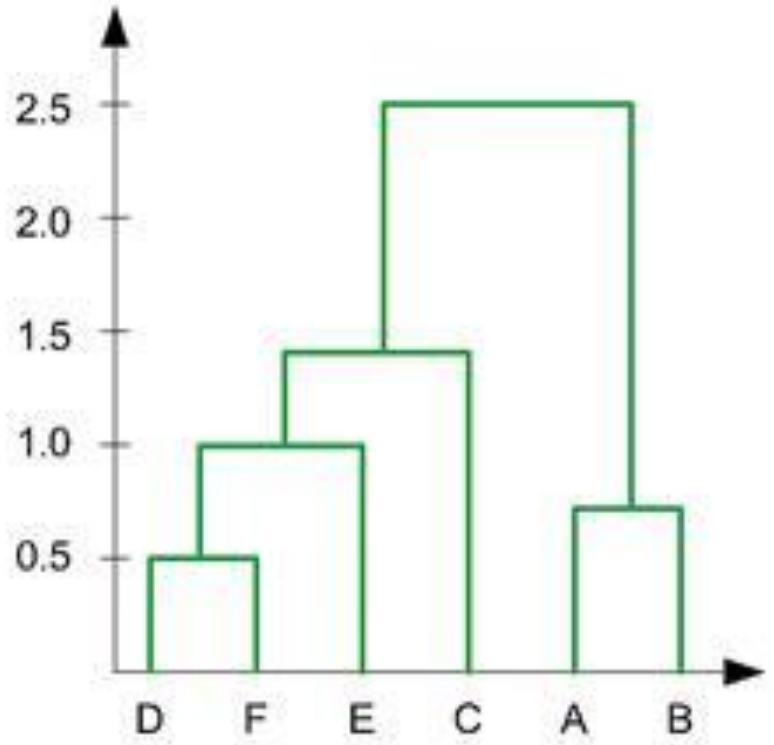

Agglomerative
p , q , r , s , t
r , s , t
p , q
s , t
p
q
r
s
t
Divisive

نمودار درخت‌واره

خوشه‌بندی تودرتو

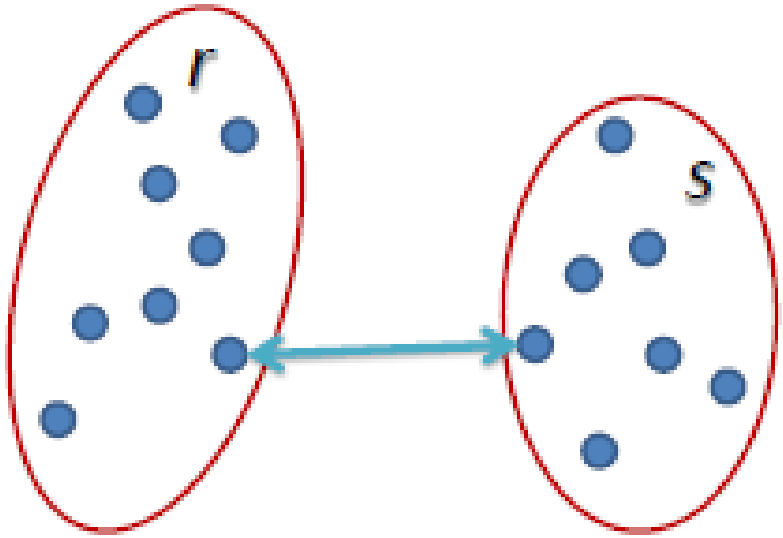
عدم تغییربرچسب‌ها در طی مراحل خوشه‌بندی





معیارهای پیوند خوشه‌ها

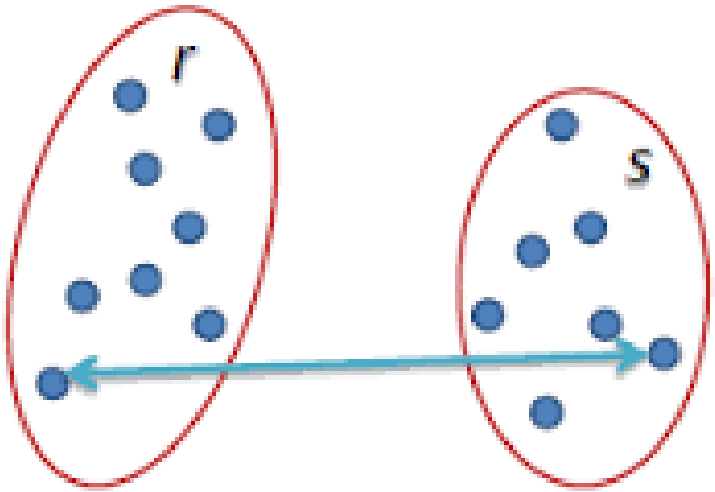
Single Linkage (= nearest neighbor) پیوند تکی – نزدیکترین همسایه



$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

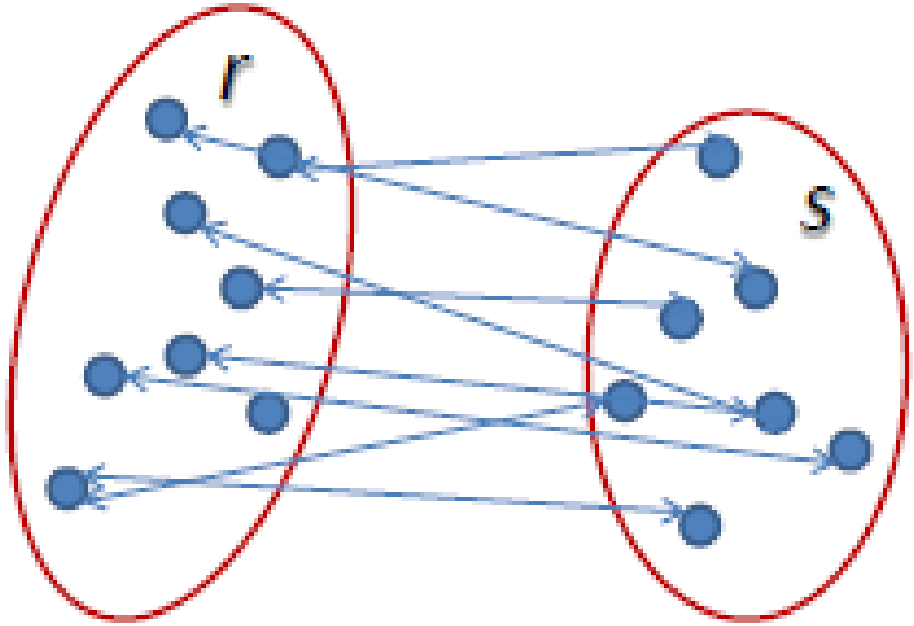
پیوند کامل- دورترین همسایه

Complete linkage- farthest neighbor



$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

Average linkage between groups پیوند میانگین بین خوشه‌ها

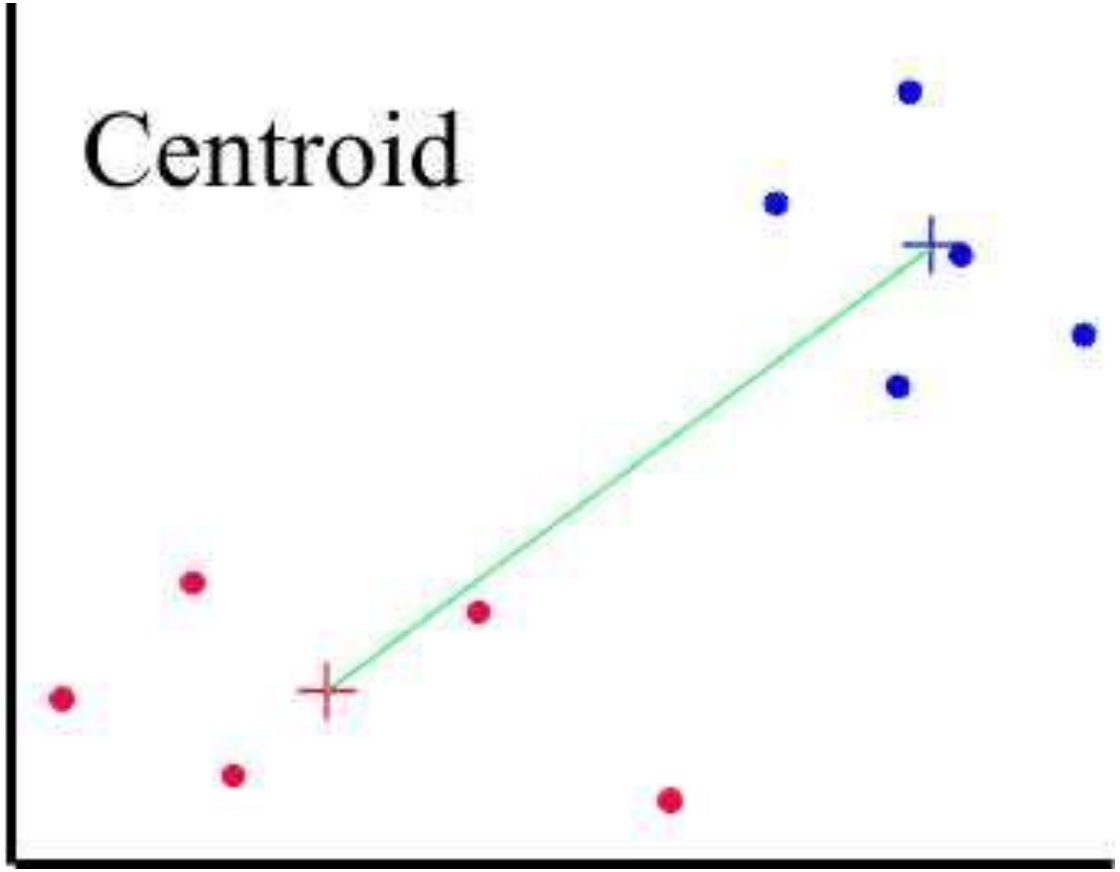


$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

Average linkage within groups

پیوند میانگین درون خوشه‌ها

پیوند گرانیکاه



پیوند میانه

$$D(A,B) = d(Med(A), Med(B))$$

روش وارد *Ward's method*

انتخاب دوخوشه برای ادغام باکمترین تغییر در میزان مجموع مربعات

خوشه بندی $k-means$

تعیین تعداد خوشه‌ها یا تعیین مقدار بهینه k یکی از چالش‌های استفاده از این روش است.

برای تعیین میزان عدم شباهت(فاصله‌ی بین مشاهدات) از فاصله‌ی اقلیدسی استفاده می‌شود.

بر خلاف روش سلسله مراتبی می‌تواند برای داده‌های حجیم نیز بکار رود

الگوریتم:

۱. انتخاب k نقطه از مشاهدات بطور تصادفی بعنوان مراکز خوشه‌ها

۲. محاسبه‌ی فاصله‌ی بین نقاط و مراکز انتخاب شده

۳. مشاهدات را با توجه به نقاطی که کمترین فاصله را از مرکز انتخاب شده دارند، به k خوشه تقسیم می‌کنیم.

٤. تکرار مراحل ۲ و ۳ (با این تفاوت که میانگین خوشه‌های انتخاب شده را بعنوان مراکز انتخاب می‌کنیم) تا زمانی که تغییری در مراکز خوشه‌ها ایجاد نشود.

تذکر: در مرحله ٤ چنانچه از میانه خوشه‌ها استفاده شود روش خوشه بندی را $k - medians$ و اگر از مد یا نما استفاده شود، روش مربوطه را $k - modes$ می‌نامیم.

خوشه بندی $Two\ Step$

مزیت‌ها

۱. تعین مقدار بهینه k (تعداد خوشه‌ها) با استفاده از معیارهای AIC و BIC

۲. قابل بکارگیری برای خوشه‌بندی داده‌های حجیم $(Big\ data)$

۳. امکان انجام خوشه‌بندی در حضور همزمان متغیرهای کمی و کیفی

٤. در این روش می‌توان مقدار k را معلوم فرض کرد. در این صورت این روش با روش $k-means$ یکسان خواهد بود با این تفاوت که می‌توان روش $k-means$ را در حضور متغیرهای کیفی هم انجام داد.

محدودیت‌ها:

۱. توزیع نرمال متغیرهای کمی

۲. توزیع چندجمله‌ای متغیرهای کیفی

۳. مستقل بودن مشاهدات

تذکر: در این روش چنانچه تمامی متغیرها کمی باشند، از فاصله‌اقلیدسی برای معیار فاصله می‌توان استفاده کرد. در غیر اینصورت، چنانچه حداقل یکی از متغیرها کیفی باشد، از تابع درستنمایی برای سنجش فاصله استفاده خواهد شد.