Exploratory Models

مدل های اکتشافی

dayche.com . گروه دایکه

دای‌کِه DAYCHE

فرآیند داده کاوی

مدل های اکتشافی

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

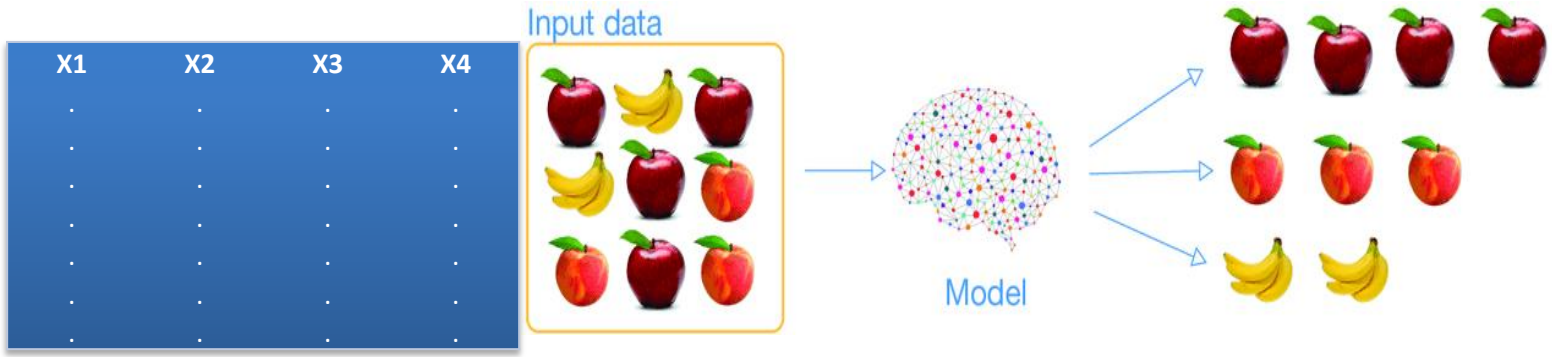
فرآیند داده کاوی

مدل های اکتشافی

❑ مدل های اکتشافی (Explorative Models)

مدل های اکتشافی در فرآیند داده کاوی که با عنوان مدل های توصیفی (Descriptive Models) نیز شناخته می شود، در دسته یادگیری بدون نظارت قرار می گیرد.

خوشه بندی
Clustering



قوانین انجمنی
Association Rules

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

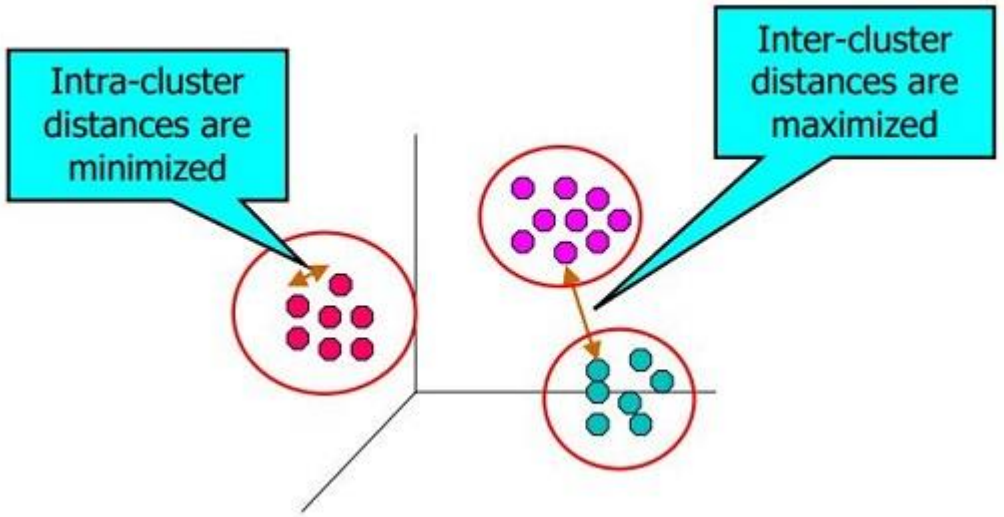
فرآیند داده کاوی

مدل های اکتشافی – خوشه بندی

❑ خوشه بندی (Clustering)

هدف از خوشه بندی، دسته بندی داده ها و تقسیم بندی آنها در چندین گروه می باشد بطوریکه اعضای داخل هر گروه دارای بیشترین شباهت (کمترین واریانس) به هم باشند و اعضای گروه های متفاوت دارای کمترین شباهت (بیشترین واریانس) باشند.

به هر یک از این گروه ها که زیر مجموعه ای از داده های شبیه به هم می باشند یک خوشه (Cluster) گفته می شود.



مثال:

- خوشه بندی مشتریان بر اساس رفتار خرید آنها (بخش بندی مشتریان)
- خوشه بندی مناطق زلزله خیز بر اساس اطلاعات لرزه های ثبت شده گذشته (تعیین زون های ساختاری زمین شناسی)

فرآیند داده کاوی

مدل های اکتشافی – خوشه بندی

❑ خوشه بندی (Clustering)

بطور کلی خوشه بندی به دو حالت مختلف می تواند وجود داشته باشد:

- **خوشه بندی سخت** (Hard Clustering): در این نوع خوشه بندی هر رکورد از داده ها صرفا به یک خوشه تعلق خواهد داشت.
- **خوشه بندی نرم/فازی** (Soft/Fuzzy Clustering): در این نوع خوشه بندی هر رکورد از داده ها، احتمال تعلق به هر یک از خوشه ها خواهد داشت.

الگوریتم های خوشه بندی نیز به انواع مختلف دسته بندی می شوند که سه رویکرد رایج در خوشه بندی این موارد هستند:

رویکرد سلسله مراتبی Hierarchical	رویکرد افرازی Partitional	رویکرد مبتنی بر تراکم Density Based

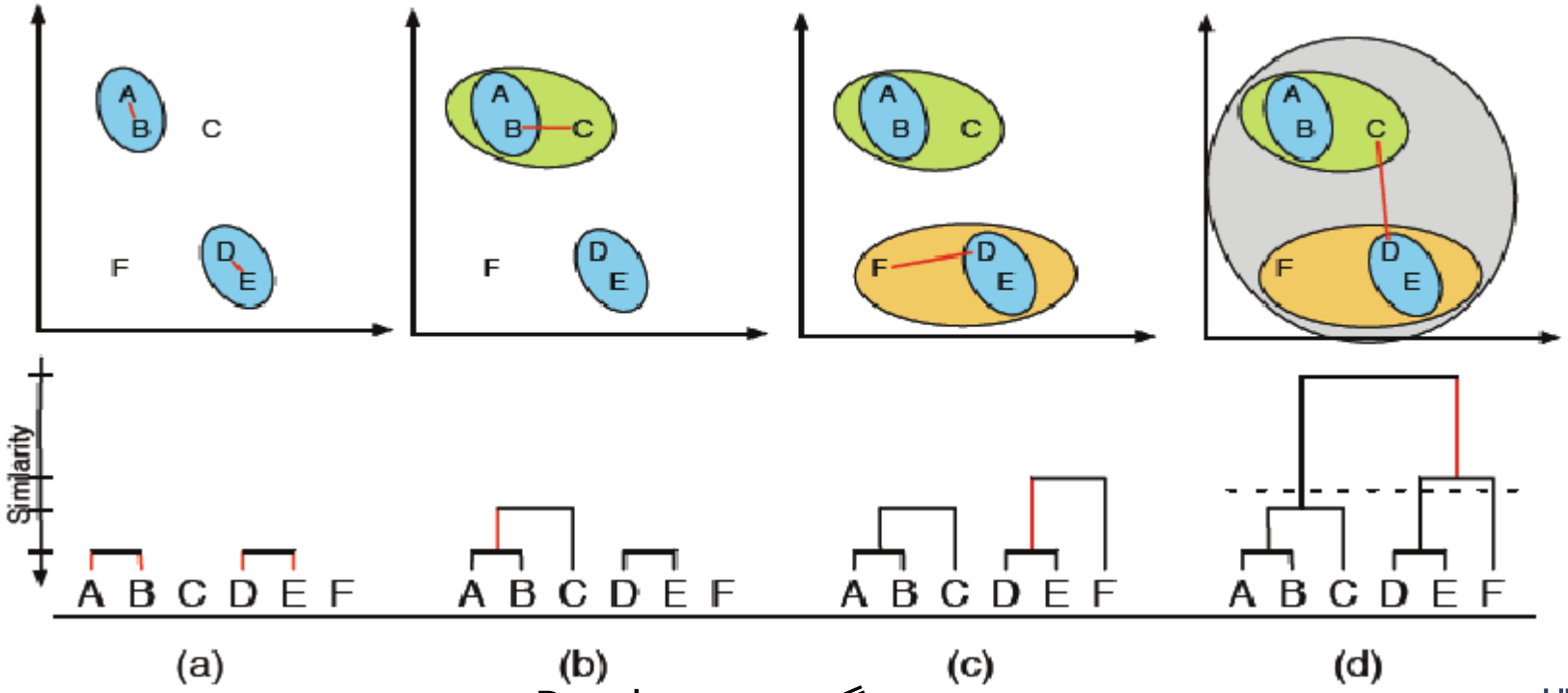
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های اکتشافی – خوشه بندی

❑ خوشه بندی سلسله مراتبی (Hierarchical Clustering)

همانطور که از اسم این الگوریتم مشخص است، این روش سلسله مراتبی از خوشه ها ایجاد می کند.



نمودار دندروگرام Dendrogram

در قدم اول هریک از رکوردها را به عنوان یک خوشه در نظر می گیرد. سپس نزدیکترین رکورد ها (شبیه ترین) را در یک سطح بالاتر در خوشه دیگری قرار می دهد و این عمل را آنقدر ادامه می دهد تا تمام رکورد ها در یک خوشه قرار گیرند.

با این روش سلسه مراتبی از خوشه ها ایجاد می شود و با انتخاب هر سطح می توان خوشه های بدست آمده در آن لایه را استخراج کرد.

رویکرد پایین به بالا
Agglomerative Approach

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ خوشه بندی سلسله مراتبی (Hierarchical Clustering)

در اجرای این الگوریتم به چند نکته می توان توجه نمود:

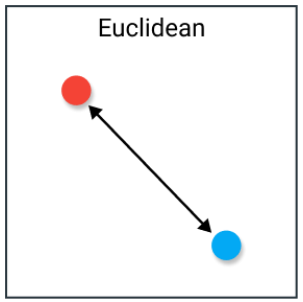
- می توان با رویکرد بالا به پایین (Top - Down) در قدم اول، همه رکوردها را در یک خوشه قرار داد و در هر قدم دورترین رکوردها (کمترین شباهت) را جدا کرده و به خوشه جدید منتقل کرد. این عمل را آنقدر ادامه داد تا تمام رکوردها به عنوان یک خوشه در نظر گرفته شوند.

- اجرای این الگوریتم نیاز به محاسبه ماتریس شباهت بین رکورد ها دارد. بنابراین با افزایش داده ها، زمان اجرای الگوریتم و نیاز به حافظه رم افزایش پیدا می کند.

- عامل تعیین کننده در پیاده سازی این الگوریتم نحوه محاسبه شباهت می باشد که دارای معیارها و رویکردهای متفاوت است.

❑ خوشه بندی سلسله مراتبی (Hierarchical Clustering)

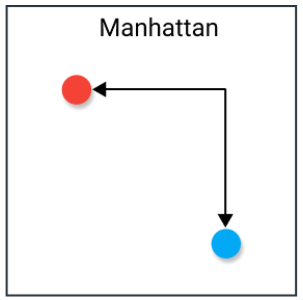
معیارهای متفاوتی برای اندازه شباهت بین دو رکورد می توان در نظر گرفت:



$$D(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

- **فاصله اقلیدسی**

رایج ترین معیار فاصله که در صورت کم بودن تعداد ویژگی ها به علت سادگی و شهودی بودن آن، نتایج بسیار خوبی خواهد داشت.



$$D(x,y) = \sum_{i=1}^{n}|x_i - y_i|$$

- **فاصله منهتن**

درک شهودی و تجسم این معیار نسبت به فاصله اقلیدسی سخت تر است ولی معمولا در داده های با ابعاد بالا عملکرد خوبی نشان می دهد.

نکته مهمی که در اندازه فاصله بایستی در نظر داشت، هم مقیاس بودن مقادیر ویژگی های مورد بررسی می باشد.

تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

فرآیند داده کاوی

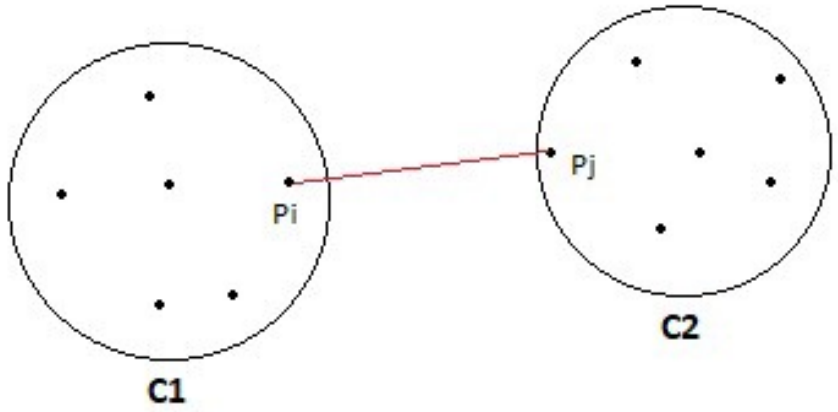
مدل های اکتشافی – خوشه بندی

- **خوشه بندی سلسله مراتبی (Hierarchical Clustering)**

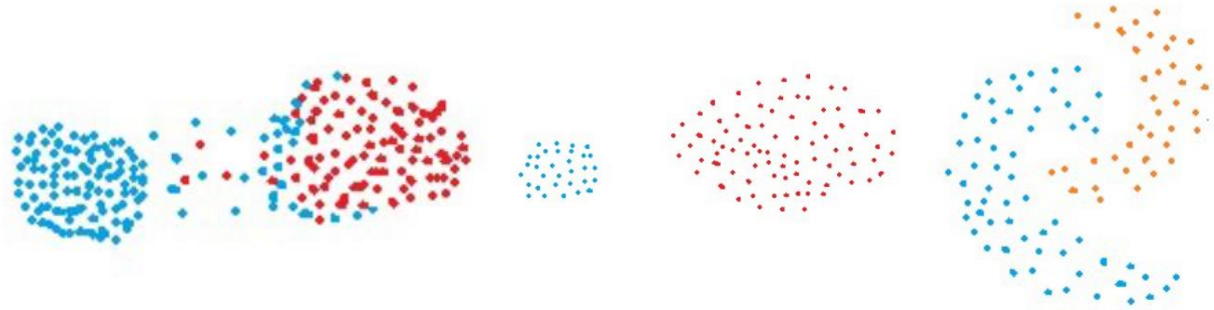
پس از تعیین معیار اندازه گیری شباهت، انتخاب رویکرد شباهت سنجی بین خوشه ها نیز دارای اهمیت می باشد:

- **رویکرد کمترین فاصله خوشه ها** Single Linkage (Min)

در این رویکرد فاصله بین دو خوشه، برابر با کوچکترین فاصله بین رکورد های دو خوشه می باشد.



امکان شناسایی الگوهای خوشه ای غیر بیضوی را فراهم می سازد، اما حساسیت به نویز در این روش بالاست و درصورت وجود نویز بین خوشه ها عملکرد مناسبی نخواهد داشت.



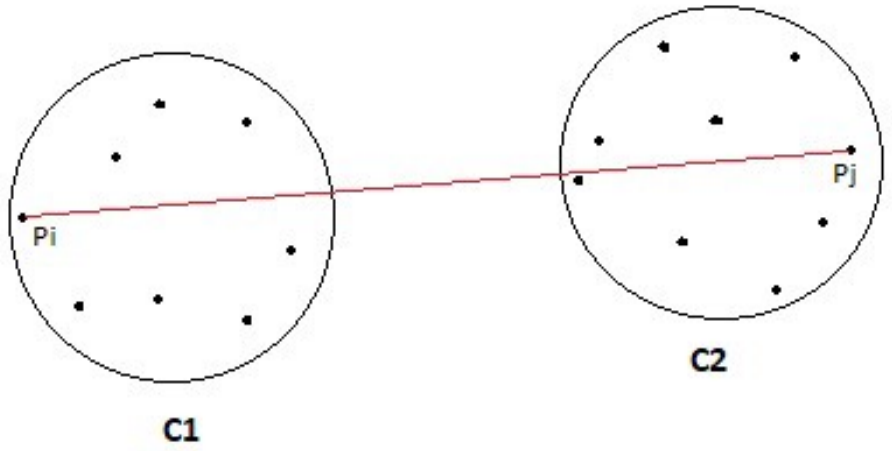
❑ خوشه بندی سلسله مراتبی (Hierarchical Clustering)

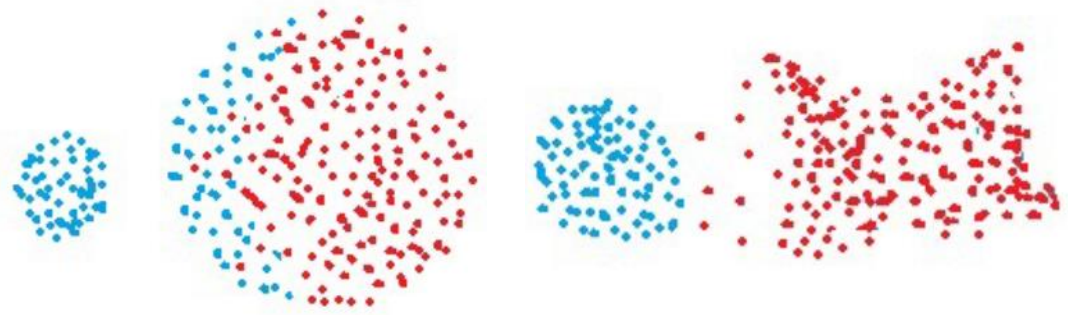
پس از تعیین معیار اندازه گیری شباهت، انتخاب رویکرد شباهت سنجی بین خوشه ها نیز دارای اهمیت می باشد:

- **رویکرد بیشترین فاصله خوشه ها** Complete Linkage (Max)

در این رویکرد فاصله بین دو خوشه، برابر با بزرگترین فاصله بین رکورد های دو خوشه می باشد.

این رویکرد نسبت به نویزهای موجود بین خوشه ها مقاوم هست و عموما سعی در تشخیص الگوهای کروی داشته و تمایل به شکستن خوشه های بزرگ دارد.





تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

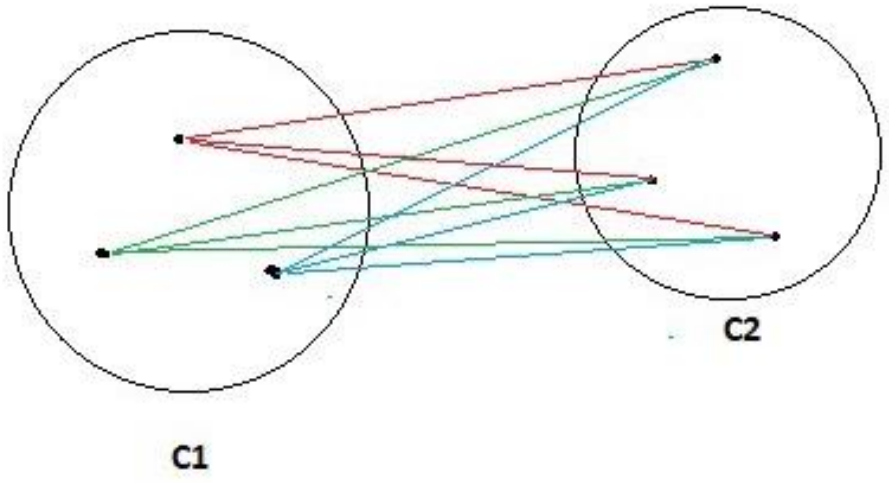
❑ خوشه بندی سلسله مراتبی (Hierarchical Clustering)

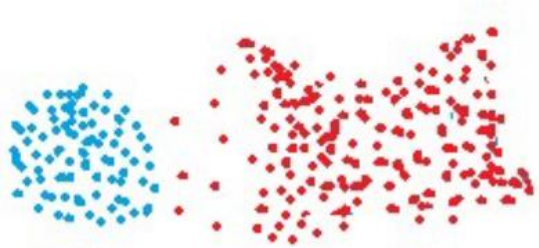
پس از تعیین معیار اندازه گیری شباهت، انتخاب رویکرد شباهت سنجی بین خوشه ها نیز دارای اهمیت می باشد:

- **رویکرد متوسط فاصله خوشه ها** (Group Average)

در این رویکرد فاصله بین دو خوشه، برابر با متوسط فاصله بین تمام رکورد های دو خوشه می باشد.

این رویکرد نسبت به نویزهای موجود بین خوشه ها مقاوم هست و عموما سعی در تشخیص الگوهای کروی دارد.





تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

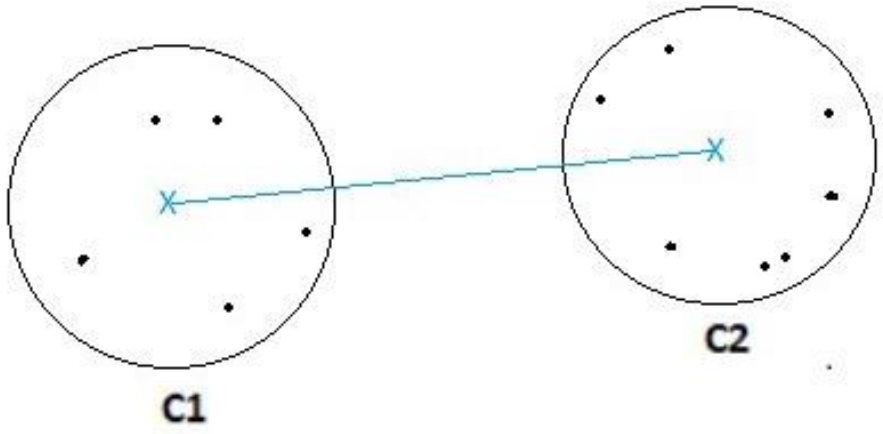
❑ خوشه بندی سلسله مراتبی (Hierarchical Clustering)

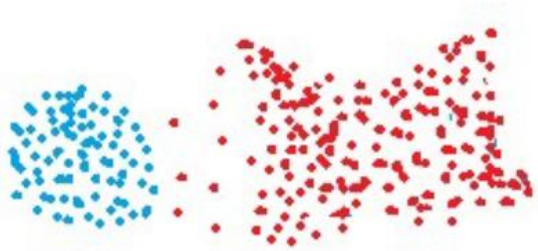
پس از تعیین معیار اندازه گیری شباهت، انتخاب رویکرد شباهت سنجی بین خوشه ها نیز دارای اهمیت می باشد:

- **رویکرد فاصله از مراکز خوشه** (Distance From Centroids)

در این رویکرد فاصله بین دو خوشه، برابر با فاصله بین مراکز دو خوشه می باشد.

این رویکرد نسبت به نویزهای موجود بین خوشه ها مقاوم هست و عموما سعی در تشخیص الگوهای کروی دارد.





تولید محتوا: زهرا ذوالقدر

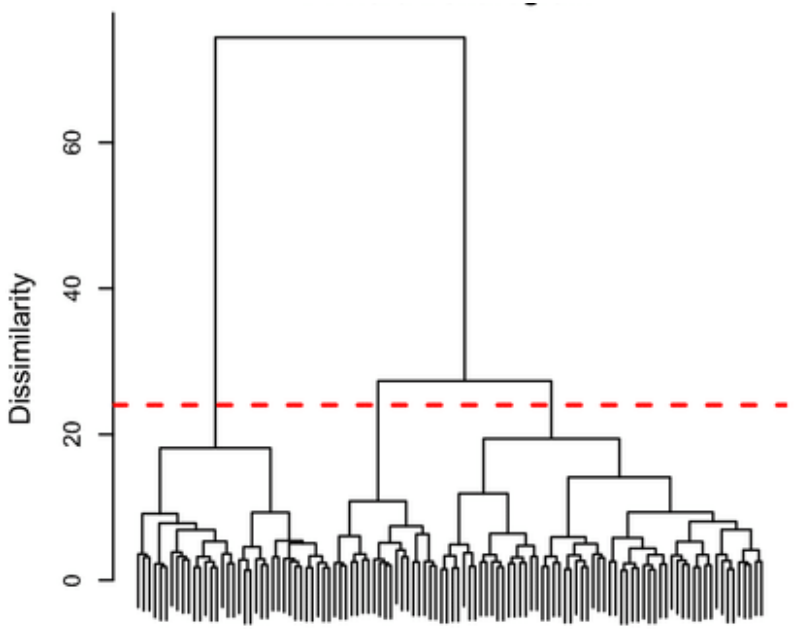
daychegroup

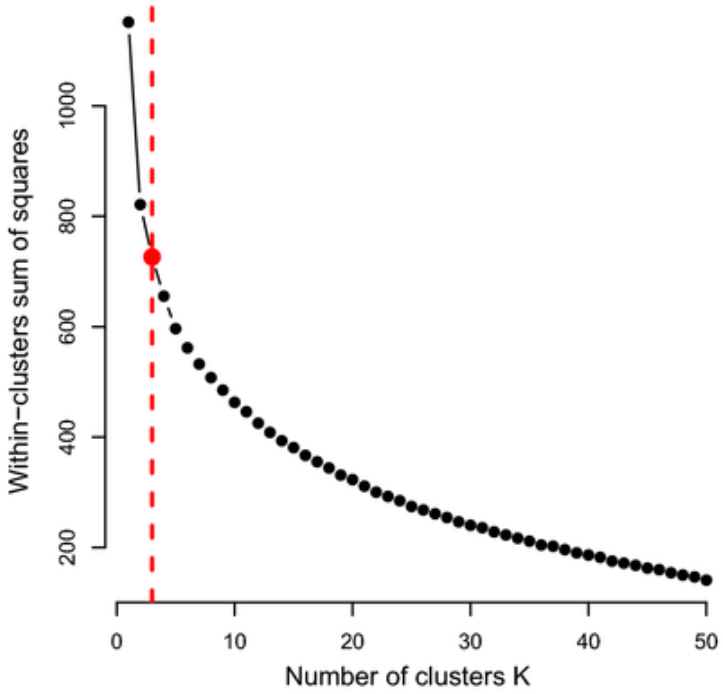
daychegroup

dayche.com | گروه دایکه

❑ خوشه بندی سلسله مراتبی (Hierarchical Clustering)

پس از اجرای الگوریتم با مقایسه شاخص های ارزیابی مجموع مربعات داخلی خوشه ها، تعداد خوشه بهینه انتخاب می‌گردد. برای این منظور استفاده از نمودار Elbow بسیار رایج است.





تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

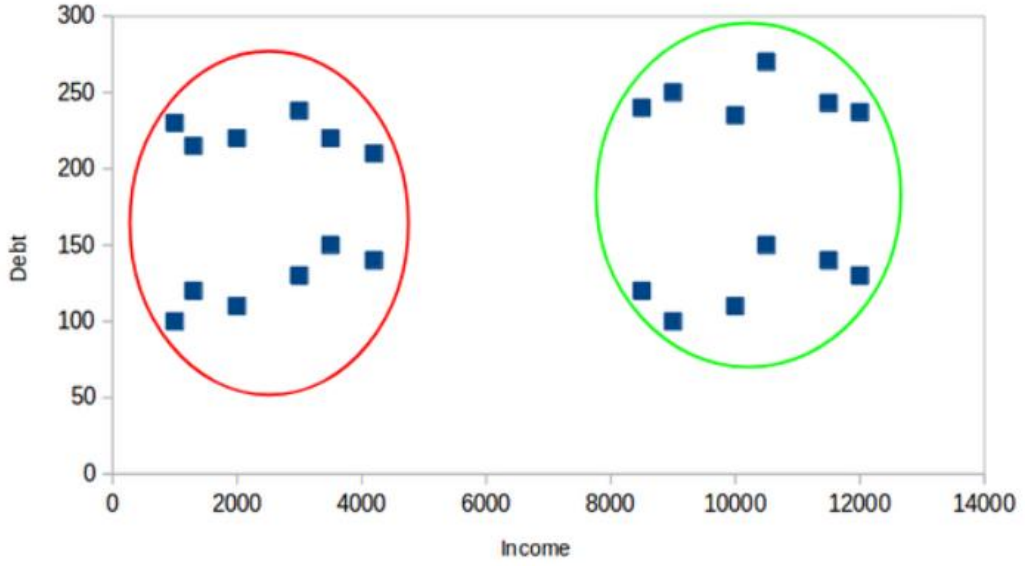
dayche.com | گروه دایکه

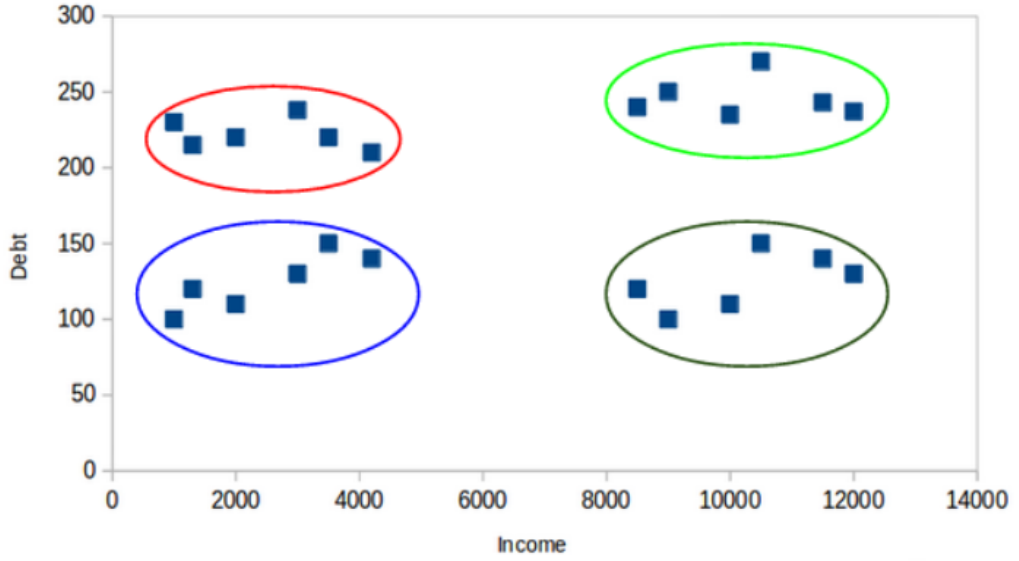
فرآیند داده کاوی

مدل های اکتشافی – خوشه بندی

❑ الگوریتم K-Means

این الگوریتم یکی از پرکاربردترین و محبوب ترین الگوریتم های خوشه بندی افرازی می باشد. در این الگوریتم، با در نظر گرفتن تعداد خوشه به اندازه K، الگوریتم سعی در تقسیم بندی داده ها در K خوشه می کند بطوریکه تمامی رکوردهای درون یک خوشه به مرکز آن خوشه نزدیکتر از مراکز خوشه های مجاور باشند.

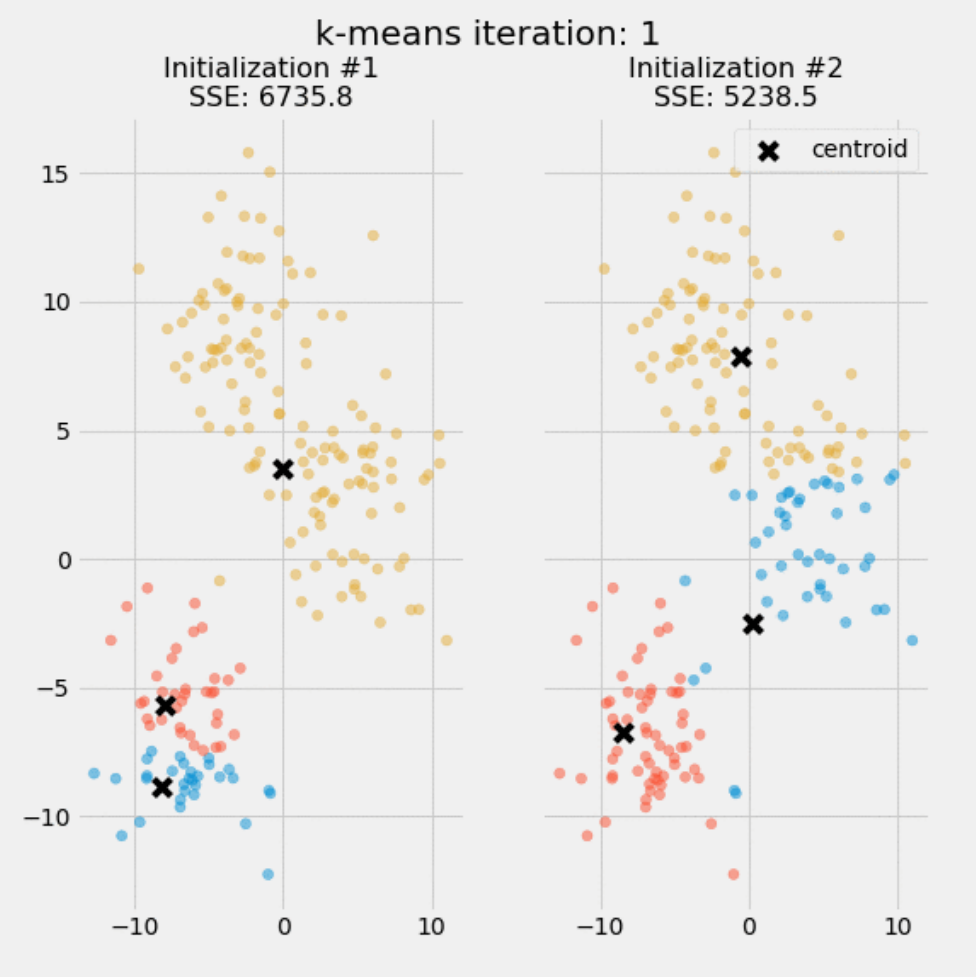




تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های اکتشافی – خوشه بندی



❑ الگوریتم K-Means

گام اول: انتخاب K رکورد از مجموعه داده ها به عنوان مراکز خوشه اولیه

معمولا مراکز خوشه اولیه بصورت تصادفی از بین داده ها انتخاب می شود. و این موضوع می تواند بعضا منجر به نتایج متفاوتی در شناسایی الگوها گردد. در واقع الگوریتم K-Means به نقاط اولیه حساس می باشد.

یکی از روش های رایج برای کنترل این مشکل، اجرای چندباره الگوریتم است. همچنین استفاده از الگوریتم خوشه بندی سلسله مراتبی و انتخاب نمونه ای از خوشه های بدست آمده در آن می تواند گزینه دیگر باشد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

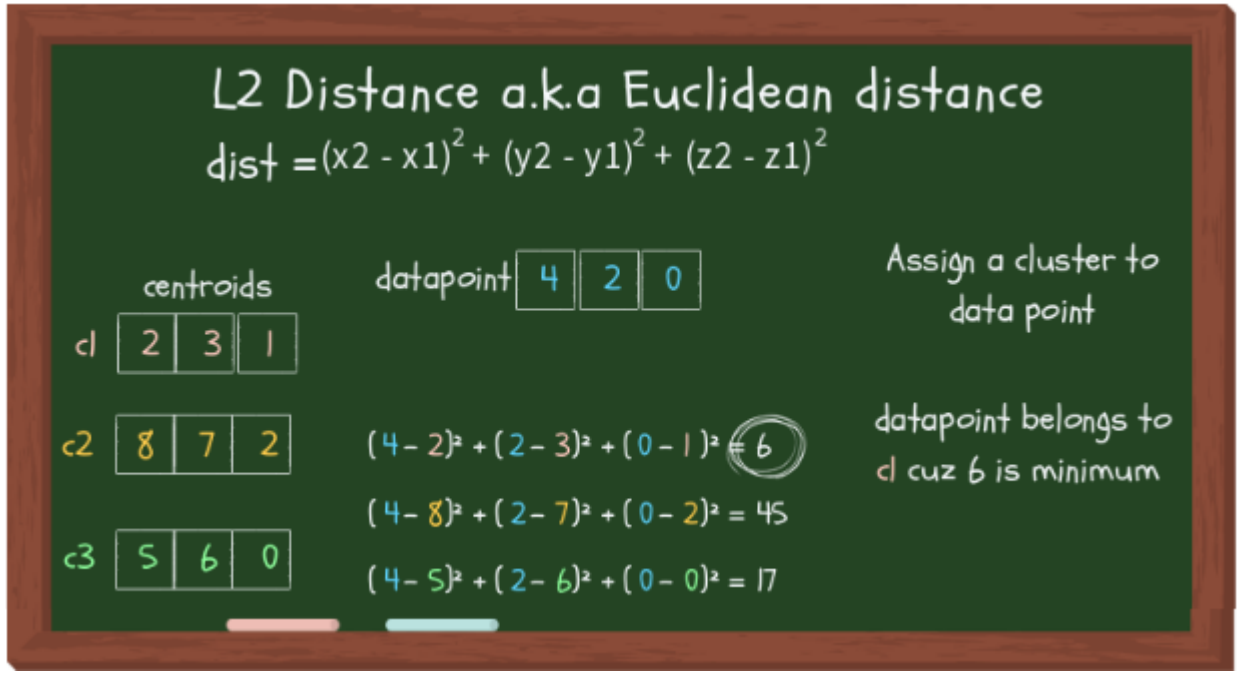
فرآیند داده کاوی

مدل های اکتشافی – خوشه بندی

❑ الگوریتم K-Means

گام دوم: برچسب گذاری تمام رکوردها بر مبنای فاصله از مراکز خوشه اولیه

با محاسبه معیار فاصله (بطور مثال فاصله اقلیدسی) تمامی رکوردها با هر یک از مراکز خوشه، برچسب خوشه ای که دارای کمترین فاصله باشد، به رکورد مورد نظر الحاق می گردد.



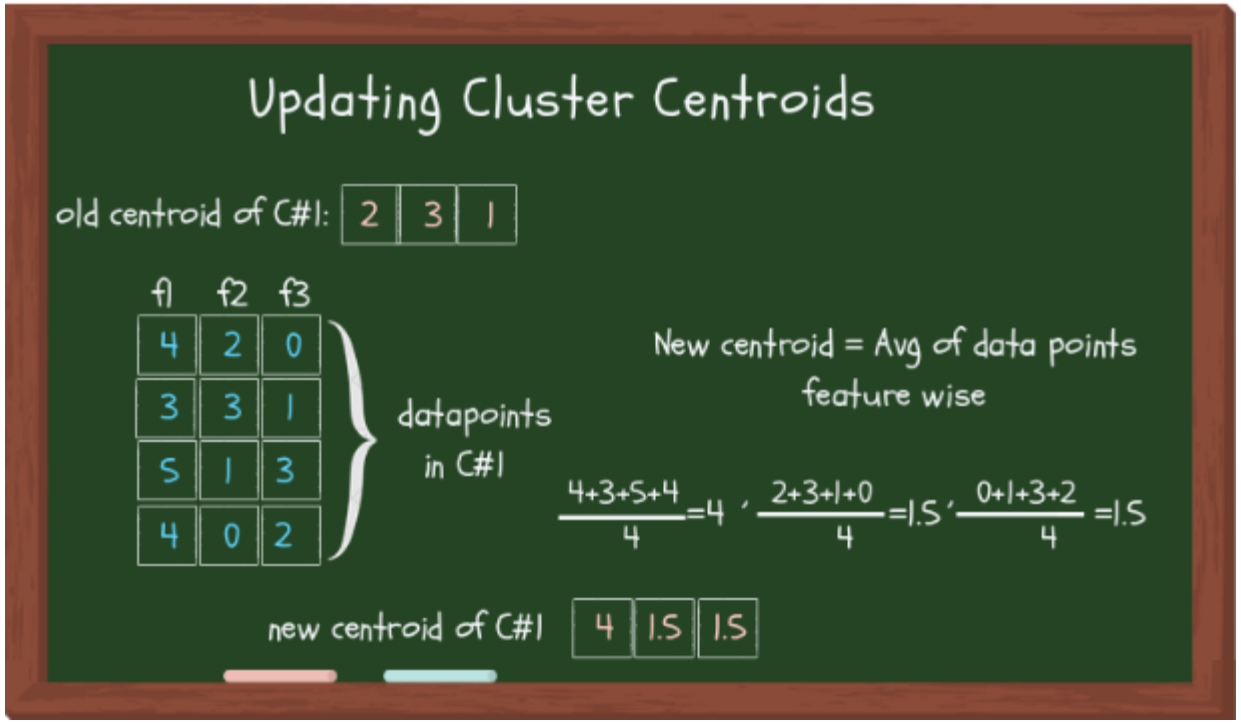
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
گروه دایکه | dayche.com

فرآیند داده کاوی

مدل های اکتشافی – خوشه بندی

❑ الگوریتم K-Means

گام سوم: بروز رسانی مراکز خوشه



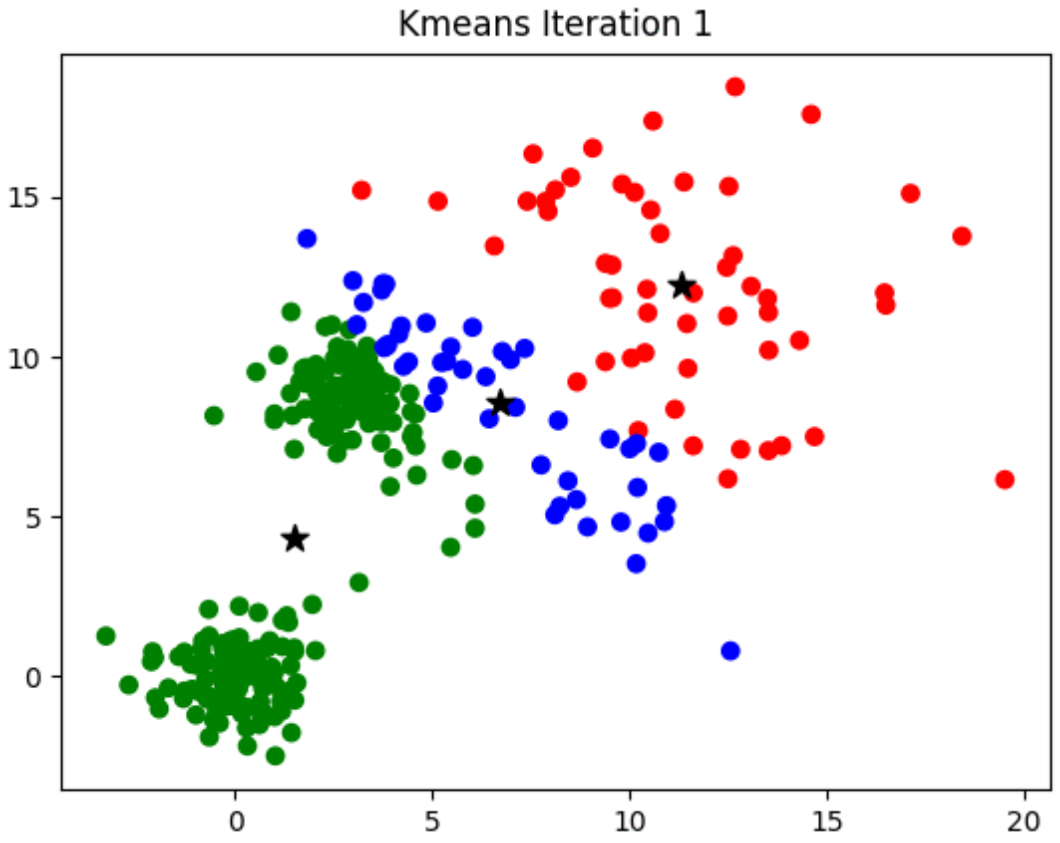
با میانگین گیری از مقادیر هر یک از ویژگی های ورودی به تفکیک خوشه ها، مراکز خوشه جدید محاسبه شده و جایگزین مراکز قبلی می شود.

وجود داده های پرت و نویزی در محاسبه میانگین، خطا ایجاد می کند؛ بنابراین یکی از راه های معرفی شده استفاده از میانه به جای میانگین هست که در الگوریتم K-Medoids استفاده می شود.

گام چهارم: تکرار گام های دوم و سوم تا رسیدن به شرایط توقف

فرآیند داده کاوی

مدل های اکتشافی – خوشه بندی



❑ الگوریتم K-Means

گام اول: انتخاب K رکورد از مجموعه داده ها به عنوان مراکز خوشه اولیه

گام دوم: برچسب گذاری تمام رکوردها بر مبنای فاصله از مراکز خوشه اولیه

گام سوم: بروز رسانی مراکز خوشه

گام چهارم: تکرار گام های دوم و سوم تا رسیدن به شرایط توقف

number of clusters — number of cases — centroid for cluster j — case i

objective function ← $J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$

Distance function

Sum of Squared Error (SSE)

تولید محتوا: زهرا ذوالقدر

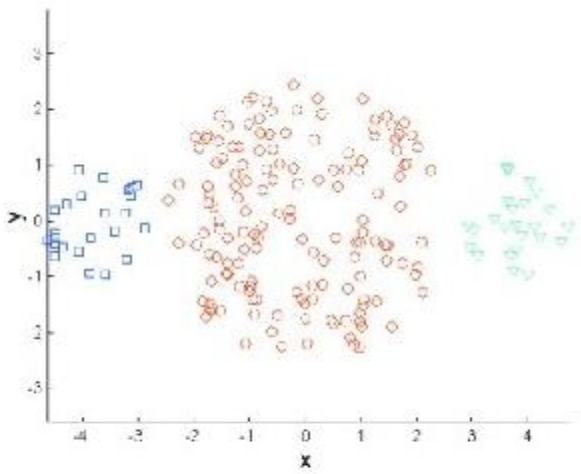
daychegroup

daychegroup

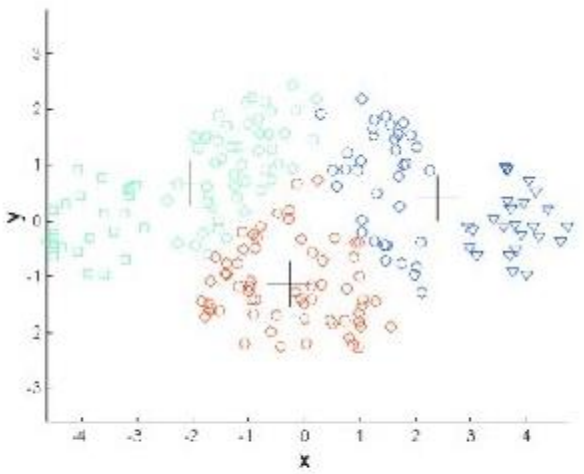
dayche.com | گروه دایکه

- فرضیات و محدودیت های الگوریتم K-Means
  - فرض هم اندازه بودن خوشه ها

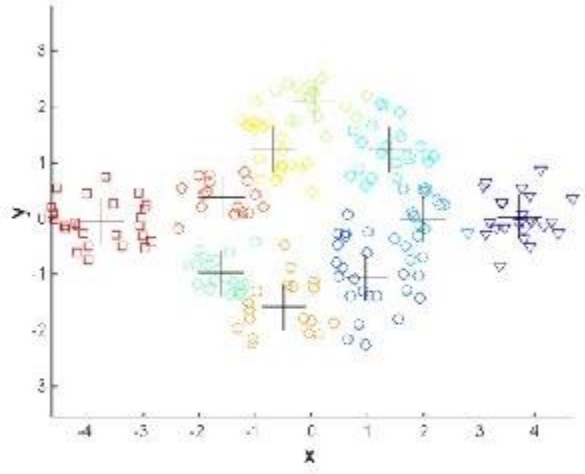
با توجه به اینکه الگوریتم K-Means به دنبال کمینه کردن مقدار SSE می باشد، بنابراین تمایل به شکستن خوشه های بزرگ و ایجاد خوشه های هم اندازه دارد. با افزایش تعداد K، می توان این مشکل را برطرف کرد.



Original Points



K-means (3 Clusters)



K-means Clusters

تولید محتوا: زهرا ذوالقدر

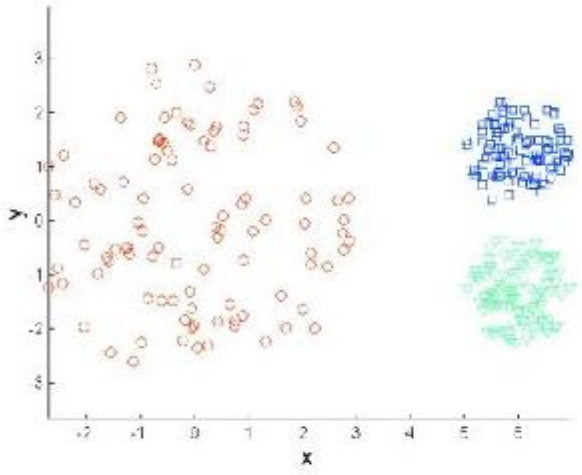
daychegroup

daychegroup

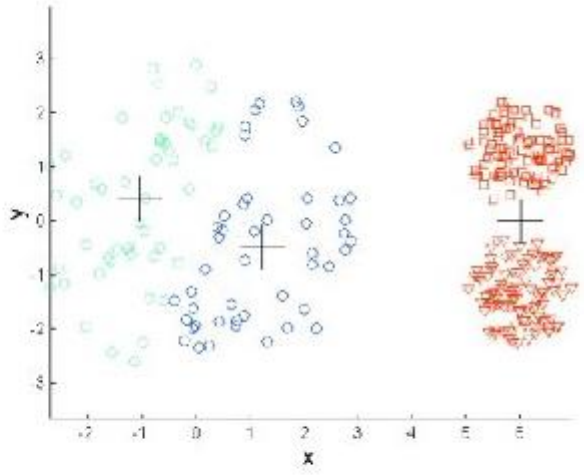
dayche.com | گروه دایکه

- فرضیات و محدودیت های الگوریتم K-Means
  - فرض یکسان بودن تراکم خوشه ها

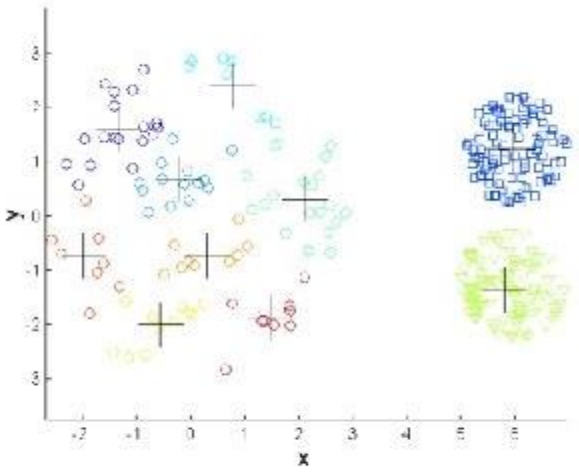
با توجه به اینکه الگوریتم K-Means به دنبال خوشه هایی با فاصله های کمتر (شبیه تر) به هم می باشد، تراکم داده ها تاثیری بر محاسبات آن ندارد و قادر به شناسایی همچین الگوهایی نیست. با افزایش تعداد K، می توان این مشکل را برطرف کرد.



Original Points



K-means (3 Clusters)



K-means Clusters

تولید محتوا: زهرا ذوالقدر

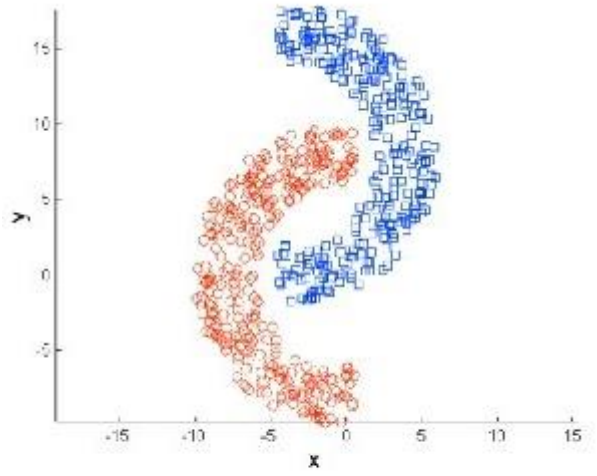
daychegroup

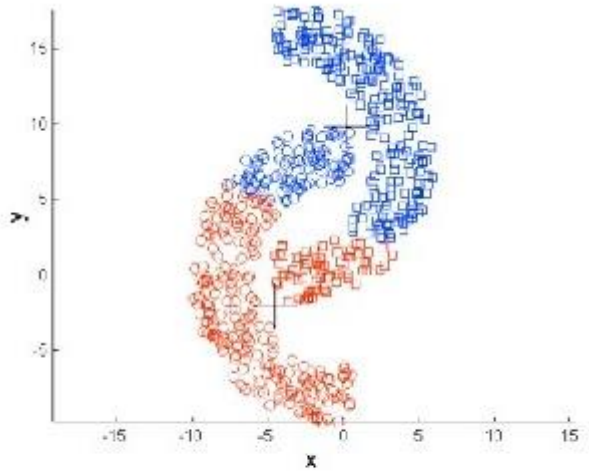
daychegroup

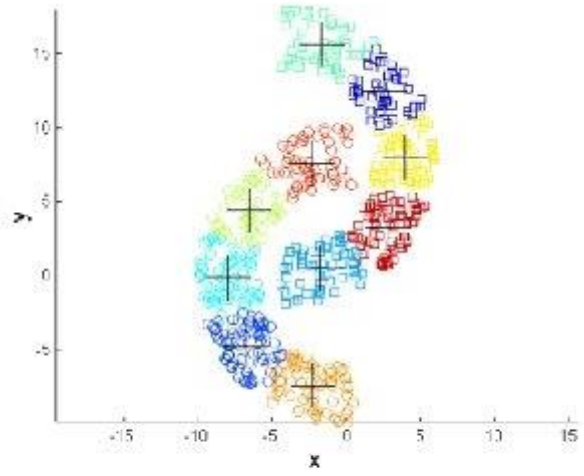
dayche.com | گروه دایکه

- فرضیات و محدودیت های الگوریتم K-Means
  - فرض کروی بودن ساختار خوشه ها

با توجه به اینکه الگوریتم K-Means فرض میکند واریانس هریک از ویژگی ها بصورت کروی هست، تنها قادر به شناسایی الگوی خوشه های کروی می باشد و نمی تواند الگوهای غیرکروی را به درستی خوشه بندی کند. با افزایش تعداد K، می توان این مشکل را برطرف کرد.







تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

- ❑ فرضیات و محدودیت های الگوریتم K-Means
  - کمی بودن ویژگی های ورودی

با توجه به اینکه الگوریتم K-Means با استفاده از میانگین گیری مراکز خوشه را محاسبه می کند، بنابراین این الگوریتم برای ویژگی های ورودی کمی قابل استفاده است. اما با توجه به اینکه در مسائل واقعی عموما ترکیبی از داده های کمی و کیفی وجود دارد، راهکار مورد استفاده برای حل این مشکل، استفاده از کد گذاری عددی برای داده های کیفی است.

عموما داده های ترتیبی پس از کد گذاری به روش Label Encoding و داده های اسمی با روش One Hot Encoding به الگوریتم K-Means وارد می شود.

Color
Yellow
Yellow
Blue
Yellow
Red
Yellow
Red
Red
Yellow
Blue

→

Color_Y	Color_B	Color_R	Target
1	0	0	0
1	0	0	1
0	1	0	1
1	0	0	1
0	0	1	1
1	0	0	0
0	0	1	1
0	0	1	0
1	0	0	1
0	1	0	0

One Hot Encoding

تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

❑ فرضیات و محدودیت های الگوریتم K-Means

○ کمی بودن ویژگی های ورودی

نکته: استفاده از روش One Hot Encoding با کد گذاری صفر و یک باعث می شود در محاسبه فاصله بین دو رکورد، یک تغییر در رده فیلد کمی به اندازه جذر 2 (به مقدار 1.41) فاصله ایجاد کند، در صورتی که یک واحد تغییر در داده های کمی به مقدار 1 واحد فاصله ایجاد کند. از این رو در برخی مواقع می توانیم برای ایجاد تاثیر یکسان از کد گذاری 0 و $\frac{\sqrt{2}}{2}$ استفاده کنیم.

ID	Age	Sex	F	M
1	23	F	1	0
2	23	M	0	1
3	24	F	1	0

$$d(1,2)=\sqrt{(23-23)^2+(1-0)^2+(0-1)^2}=\sqrt{2}=1.4$$

$$d(1,3)=\sqrt{(23-24)^2+(1-1)^2+(0-0)^2}=\sqrt{1}=1$$

ID	Age	Sex	F	M
1	23	F	$\sqrt{2}/2$	0
2	23	M	0	$\sqrt{2}/2$
3	24	F	$\sqrt{2}/2$	0

$$d(1,2)=\sqrt{(23-23)^2+(\sqrt{2}/2-0)^2+(0-\sqrt{2}/2)^2}=\sqrt{1}=1$$

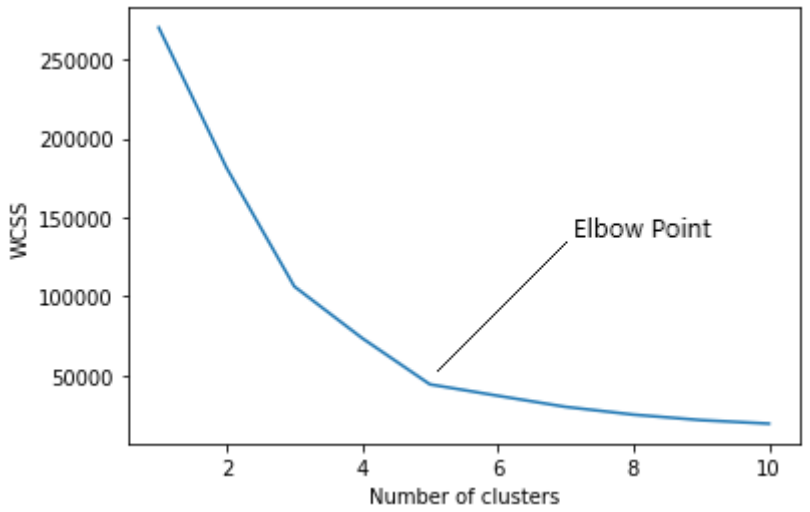
$$d(1,3)=\sqrt{(23-24)^2+(\sqrt{2}/2-\sqrt{2}/2)^2+(0-0)^2}=\sqrt{1}=1$$

❑ فرضیات و محدودیت های الگوریتم K-Means

○ تعیین تعداد خوشه مناسب

الگوریتم K-Means برای اجرا نیاز به معلوم بودن تعداد خوشه دارد. اما در واقعیت تعداد خوشه های مناسب یک پارامتر نامعلوم می باشد. استفاده از الگوریتم های دیگر مانند خوشه بندی سلسله مراتبی برای تعیین تعداد خوشه یکی از راهکارهای حل این مشکل هست. اما روش رایج برای تعیین تعداد خوشه مناسب، اجرای الگوریتم به ازای تعداد خوشه های متفاوت و مقایسه مقدار کارایی آنها بر اساس شاخص های ارزیابی الگوریتم های خوشه بندی می باشد.

در اغلب مسائل، دامنه جستجوی تعداد خوشه، بر اساس تجربیات پیشین در مسئله و الزامات کسب و کار بدست می آید و با مقایسه شاخص های ارزیابی آنها مانند میزان SSE، تعداد خوشه بهینه انتخاب می گردد. برای این منظور استفاده از نمودار Elbow بسیار رایج است.



تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

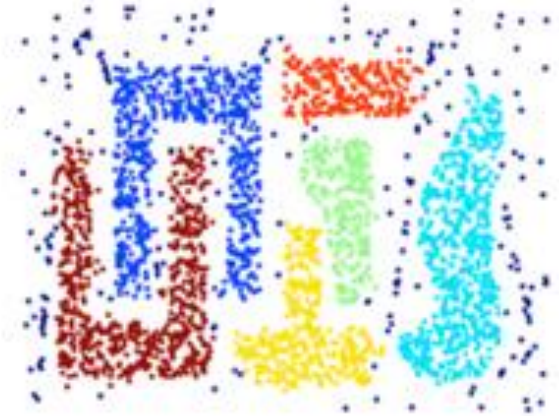
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های اکتشافی – خوشه بندی

❑ الگوریتم DBSCAN (Density Based Spatial Clustering of Applications with Noise)

الگوریتم DBSCAN رایج ترین الگوریتم خوشه بندی مبتنی بر تراکم می باشد که در مقابل نویز و داده های پرت مقاوم می باشد. همچنین با توجه به ساختار این الگوریتم، جهت شناسایی الگوهای پیچیده و غیرکروی مورد استفاده قرار می گیرد.



ایده اصلی در این الگوریتم اینست که یک رکورد به یک خوشه تعلق دارد در صورتی که به رکوردهای زیادی از آن خوشه نزدیک باشد.

بنابراین تعریف میزان تراکم داده ها، اهمیت اساسی در شناسایی ساختار الگوها دارد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

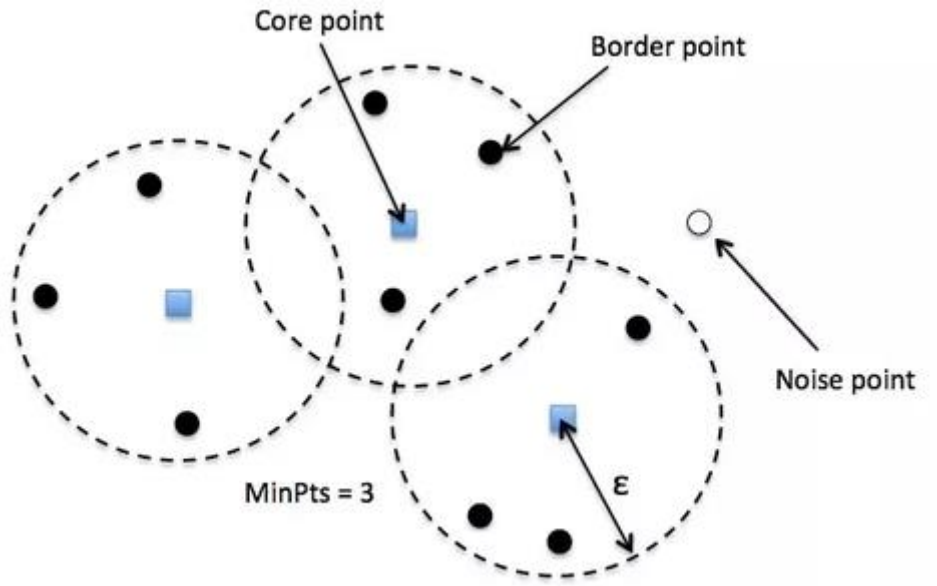
❑ الگوریتم DBSCAN (Density Based Spatial Clustering of Applications with Noise)

دو پارامتر اصلی برای اجرای الگوریتم وجود دارد:

- **eps یا شعاع همسایگی**: فاصله ای که برای تعریف همسایگی بکار می رود. اگر دو رکورد دارای فاصله کمتر از آن باشند، نقاط همسایه در نظر گرفته می شود.
- **minPts**: حداقل تعداد همسایه در محدوده یک شعاع تعریف شده جهت قرار گیری در یک خوشه

با توجه به پارامترهای تعریف شده، سه گروه از داده ها قابل تعریف است:

- **نقاط مرکزی Core Point**: نقاطی از داده ها که در شعاع همسایگی آنها حداقل به تعداد minPts همسایه وجود داشته باشد.
- **نقاط مرزی Border Points**: همسایگانی از نقاط مرکزی که قابلیت تبدیل به نقاط مرکزی ندارند.
- **نقاط پرت Noise Points**: نقاطی که در همسایگی هیچ نقطه مرکزی نیستند.



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

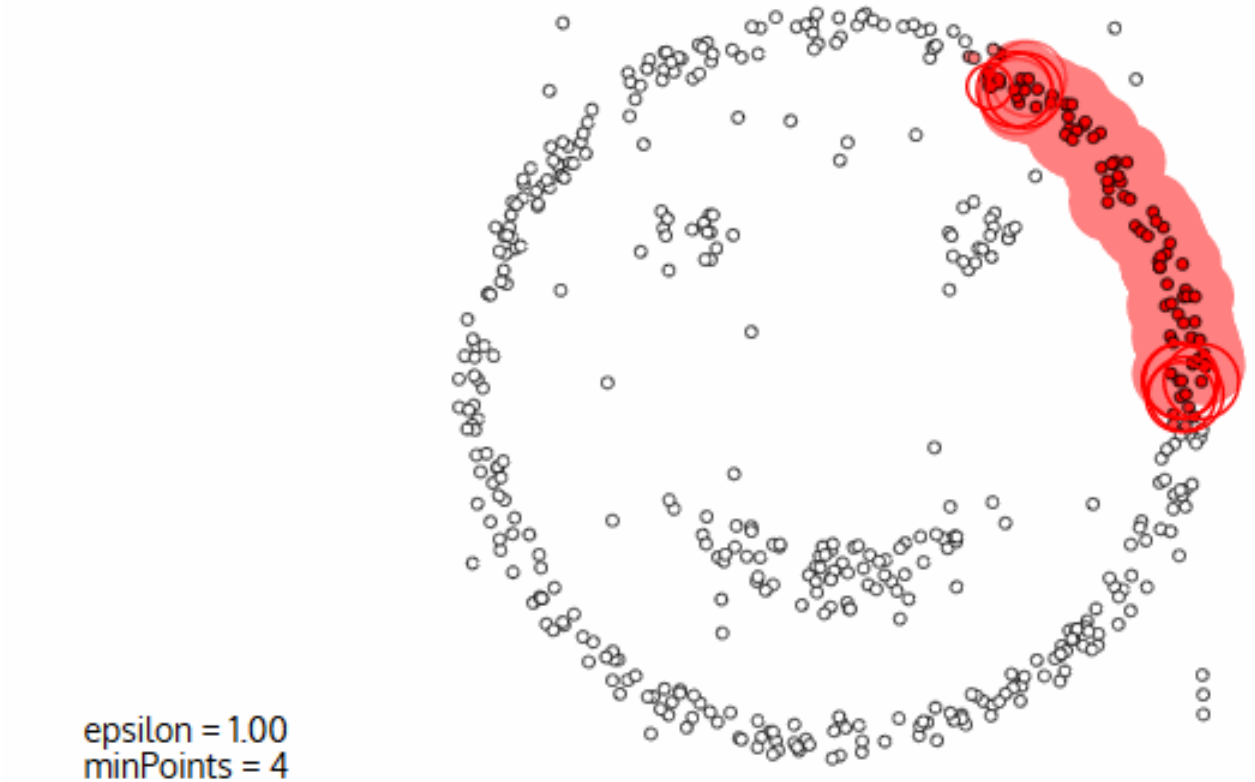
❑ الگوریتم DBSCAN (Density Based Spatial Clustering of Applications with Noise)

نحوه اجرای الگوریتم:

- یک نقطه به تصادف انتخاب می شود و با توجه به پارامترهای eps و minPts نوع داده تعیین می گردد. در صورتی که **نویز** تشخیص داده شود، کنار گذاشته شده و نقطه دیگری انتخاب می گردد.
- در صورتی که نوع داده، **نقطه مرکزی** تشخیص داده شود، به همراه تمام همسایه های خود متعلق به خوشه اول در نظر گرفته می شود و سپس **نقطه مرکزی** بودن همسایه ها مورد بررسی قرار می گیرد. در صورتی که تمام یا برخی از آنها نیز از نوع **نقاط مرکزی** باشد تمامی همسایه های مربوط به سایر **نقاط مرکزی** نیز متعلق به خوشه اول در نظر گرفته می شود و این روند ادامه پیدا می کند تا خوشه اول ساخته شود.
- برای شروع ساخت خوشه دوم از بین نقاط داده باقیمانده (از جمله داده های **نویزی** مرحله قبل)، مجددا نقطه ای به تصادف انتخاب شده و کلیه مراحل قبل تکرار می شود.
- این روند ادامه پیدا کرده تا هیچ **نقطه مرکزی** بر اساس پارامترهای تعریف شده eps و minPts در مجموعه داده ها وجود نداشته باشد.

❑ الگوریتم DBSCAN (Density Based Spatial Clustering of Applications with Noise)

نحوه اجرای الگوریتم:



تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

❑ الگوریتم DBSCAN (Density Based Spatial Clustering of Applications with Noise)

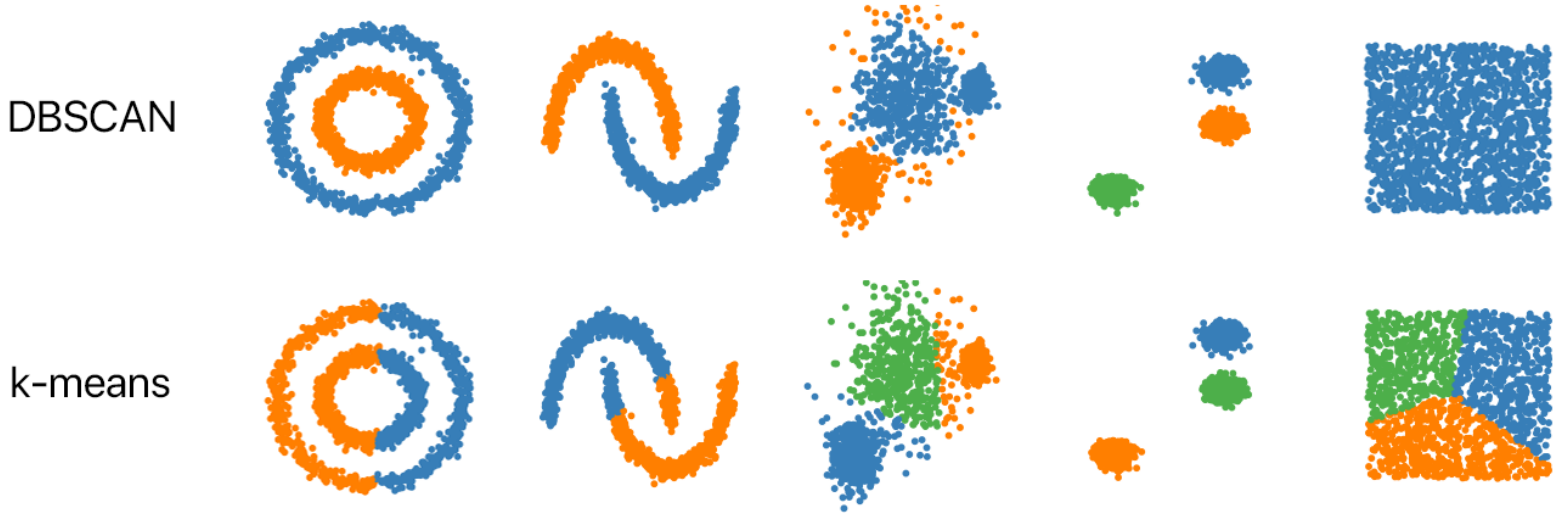
تنظیم پارامتر های الگوریتم:

- **minPts**: انتخاب مقادیر کوچک برای این پارامتر معمولا نتایج خوبی نخواهد داشت. بطور مثال انتخاب minPts =1 منجر به این خواهد شد که هر کدام از رکوردها به عنوان یک خوشه انتخاب شود. در صورتی که minPts =2 تنظیم گردد نتیجه خوشه بندی معادل خوشه بندی سلسله مراتبی با روش Single Linkage همراه با مقدار آستانه ای eps برای انتخاب خوشه می باشد. بنابراین مقدار این پارامتر معمولا بزرگتر از 2 و با توجه به ابعاد داده ها تعیین می گردد. به عنوان یک قاعده کلی مقدار این پارامتر بزرگتر از تعداد ویژگی های ورودی به الگوریتم انتخاب می گردد و انتخاب مقادیر بزرگ برای داده های نویزی و پیچیده مناسب تر است.
- **eps یا شعاع همسایگی**: در صورتی که پارامتر eps خیلی کوچک انتخاب گردد، بسیاری از داده ها در خوشه ها قرار نمی گیرند و در صورت انتخاب مقادیر بزرگ بخش زیادی از داده ها در یک خوشه ادغام می شود. به عنوان یک رویکرد کلی، تنها بخش کوچکی از داده ها بایستی در شعاع همسایگی قرار گیرند و انتخاب مقادیر کوچکتر پارامتر eps ارجحیت دارد.

❑ الگوریتم DBSCAN (Density Based Spatial Clustering of Applications with Noise)

الگوریتم DBSCAN در مقایسه با الگوریتم های افرازی مانند K-Means قابلیت شناسایی الگوهای پیچیده و تودرتو را از طریق تفکیک نواحی پرتراکم و کم تراکم ایجاد می کند.

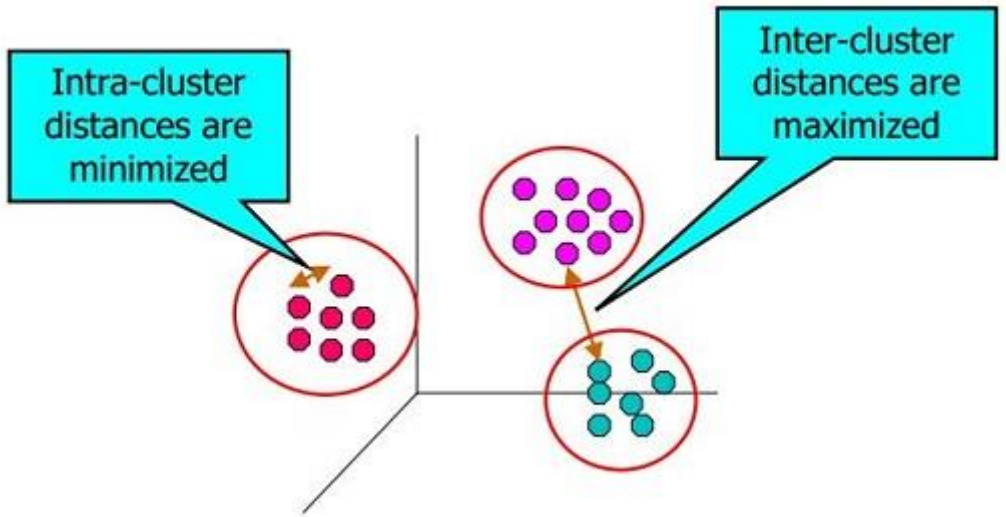
ولی بایستی توجه داشت در صورتی که تراکم داخلی خوشه ها تفاوت بسیاری داشته باشد، با توجه به نحوه عملکرد این الگوریتم بر اساس پارامترهای eps و minPts این الگوریتم توانایی شناسایی خوشه ها با تراکم های غیر همگن را ندارد. برای این منظور الگوریتم OPTICS معرفی شده است.



❑ معیار های ارزیابی خوشه بندی

هدف خوشه بندی ایجاد گروه هایی از داده هاست که اعضای درون هر گروه دارای بیشترین شباهت (کمترین پراکندگی) و اعضای بین گروه ها دارای کمترین شباهت (بیشترین پراکندگی) باشند.

بنابراین اندازه گیری کارایی مدل های خوشه بندی را می توان بر اساس همین هدف بدست آورد. در واقع اغلب معیارهای ارزیابی خوشه بندی بر اساس دو فاکتور انسجام خوشه ای (Cohesion) و جدایی خوشه ای (Separation) تعریف می گردد که معمولا به آنها معیارهای داخلی می گویند.



❑ معیار های ارزیابی خوشه بندی

- معیار مجموع مربعات درون خوشه ای (WSS)

این شاخص همان تابع زیان الگوریتم K-Means است (SSE) و با کمینه کردن آن، خوشه ها شناسایی می شود.

Within cluster: $$WSS(C,k)=\sum_{i=1}^{N}\| x_i - c_{p(i)} \|^2$$

Between clusters: $$BSS(C,k)=\sum_{j=1}^{k} n_j \| c_j - \bar{x} \|^2$$

Total Variance of data set:

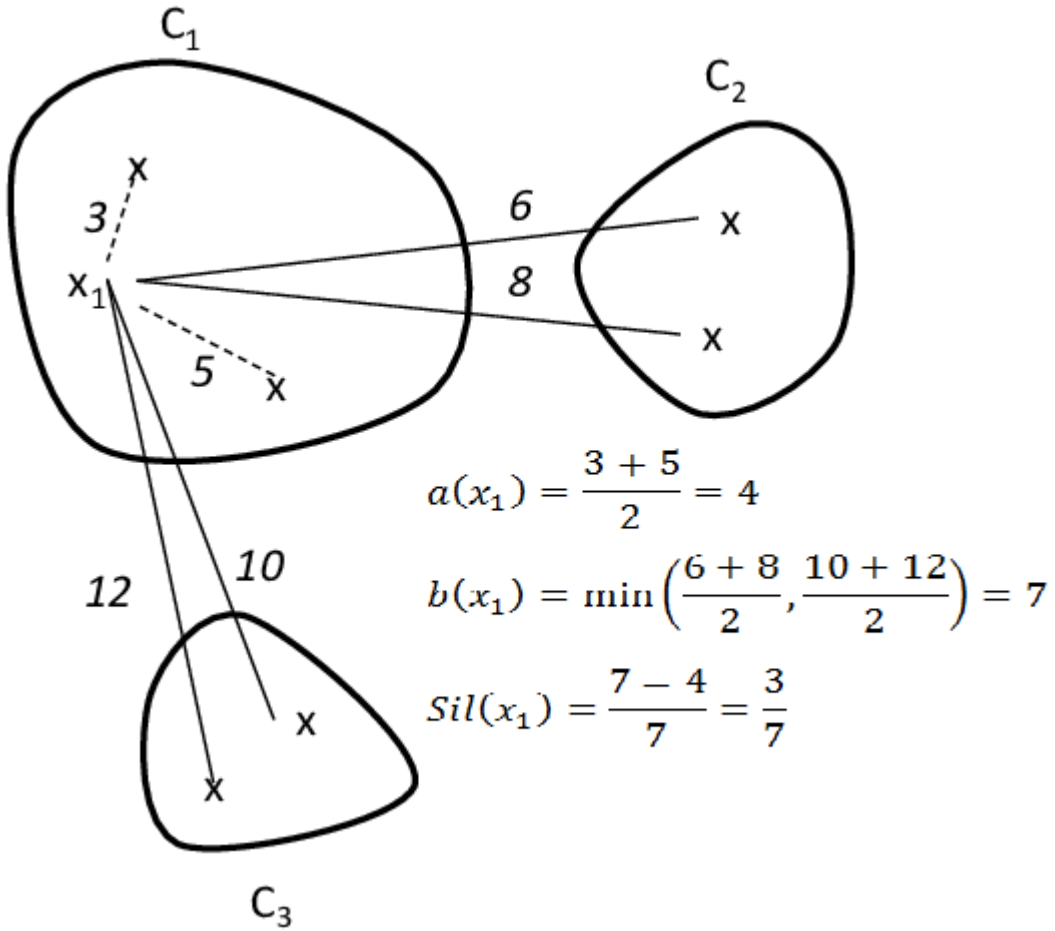
$$\sigma(X)=\underbrace{\sum_{i=1}^{N}\| x_i - c_{p(i)} \|^2}_{WSS}+\underbrace{\sum_{j=1}^{k} n_j \| c_j - \bar{x} \|^2}_{BSS}$$

ویژگی بسیار جالب این معیار این است که علاوه بر سادگی و سرعت بالای محاسبه، در ذات خود فاکتورهای انسجام خوشه ای و جدایی خوشه ای را به صورت همزمان داراست.

از آنجا که مقدار واریانس کل برای هر مجموعه داده مقدار ثابتی هست، بنابراین کمینه کردن فاکتور انسجام خوشه ای WSS به معنی بیشینه کردن فاکتور جدایی خوشه ای BSS می باشد.

فرآیند داده کاوی

مدل های اکتشافی – خوشه بندی



$$a(x_1) = \frac{3+5}{2} = 4$$

$$b(x_1) = \min\left(\frac{6+8}{2}, \frac{10+12}{2}\right) = 7$$

$$Sil(x_1) = \frac{7-4}{7} = \frac{3}{7}$$

معیار های ارزیابی خوشه بندی

- معیار نیم رخ (Silhouette)

این معیار میزان تعلق هر رکورد به خوشه ای که در آن قرار گرفته، در مقایسه با خوشه مجاور را اندازه گیری می کند.

محاسبه معیار سیلوئت برای هر رکورد انجام می شود:

a: میانگین فاصله هر رکورد از تمامی رکوردهای هم خوشه

b: میانگین فاصله هر رکورد از تمامی رکوردهای نزدیکترین خوشه مجاور

$$Silhouette - score = \frac{b_i - a_i}{\max(bi, a_i)}$$

تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

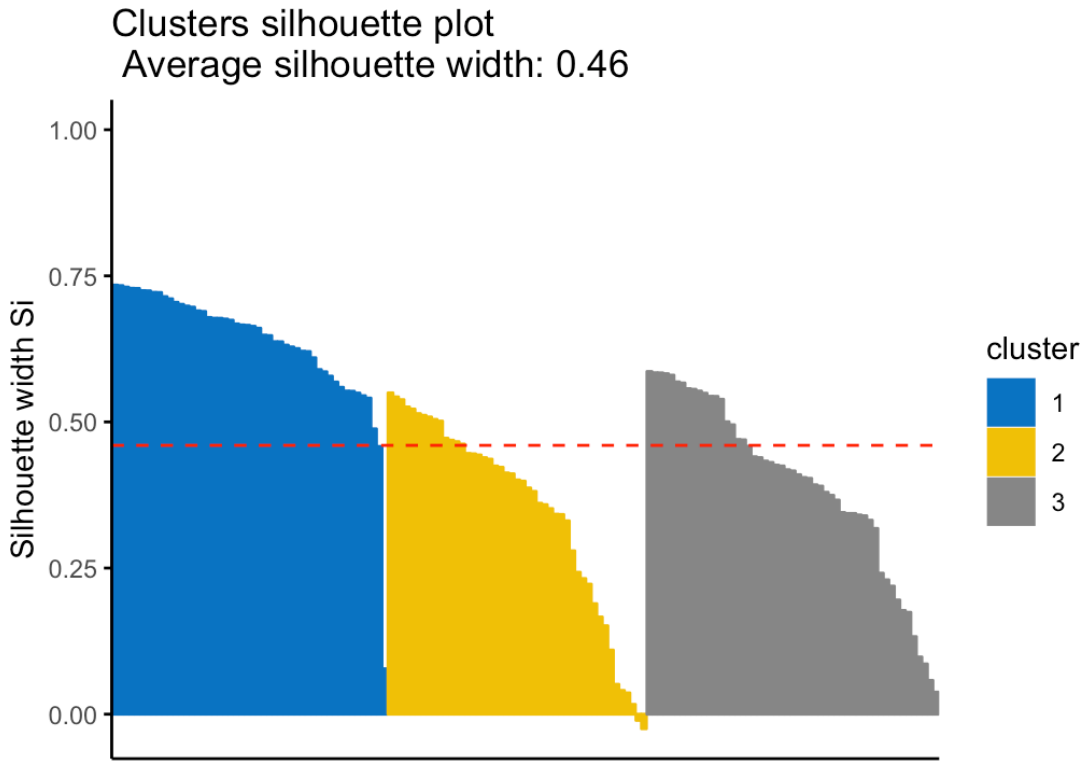
فرآیند داده کاوی

مدل های اکتشافی – خوشه بندی

معیار های ارزیابی خوشه بندی

- معیار نیم رخ (Silhouette)

دامنه معیار سیلوئت در بازه 1- تا 1 قرار می گیرد. مقادیر مثبت و نزدیک به مقدار یک به معنی خوشه بندی درست و با کیفیت خوب است. همچنین منفی بودن این معیار نیز به معنی نادرست بودن خوشه بندی می باشد.



از مزایای این معیار می توان گفت، قابلیت ارزیابی کلی مدل، ارزیابی هر خوشه مجزا و همچنین ارزیابی هر رکورد را فراهم می سازد.

در ارزیابی کلی مدل خوشه بندی مناسب معمولا میانگین معیار سیلوئت بایستی بالاتر از 0.2 در نظر گرفته شود.

به علت نیاز به محاسبه فاصله بین تمام جفت رکوردها، سرعت محاسبه این معیار کند بوده و معمولا در مجموعه داده های بزرگ استفاده نمی شود.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های اکتشافی – خوشه بندی

معیار های ارزیابی خوشه بندی

- معیار دان (Dunn Index)

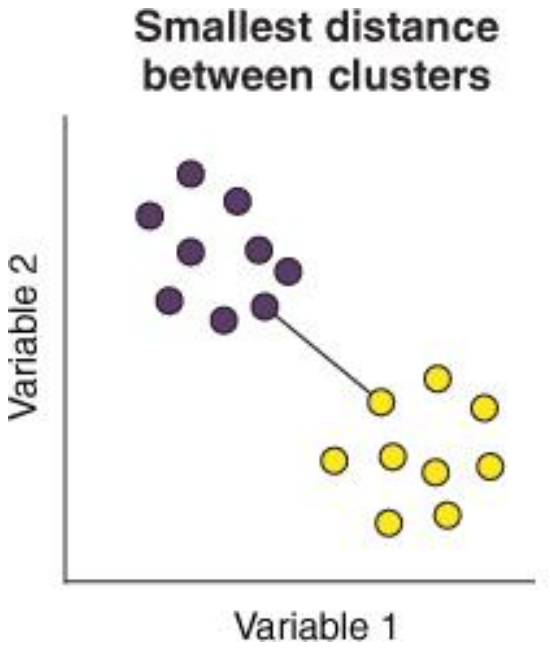
این شاخص نیز مانند سایر شاخص های معرفی شده، از ترکیب فاکتورهای انسجام خوشه ای و جدایی خوشه ای استفاده می کند.

محاسبه معیار دان:

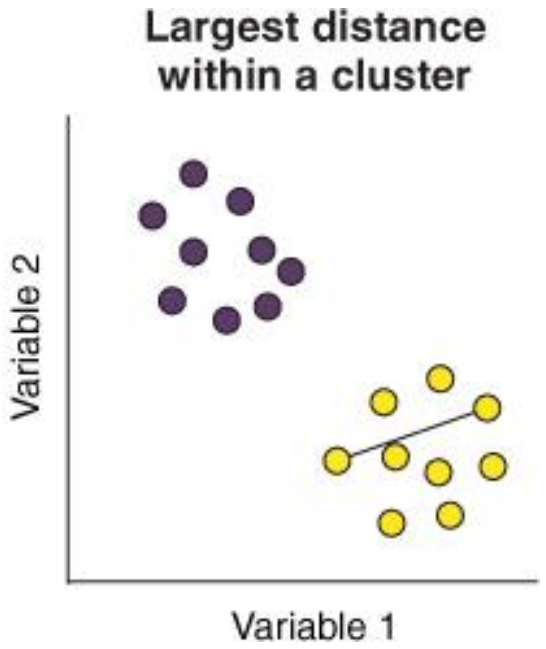
جدایی خوشه ای: کوچکترین فاصله بین اعضای دو خوشه

انسجام خوشه ای: بزرگترین فاصله بین اعضای یک خوشه (قطر خوشه)

شاخص دان با نسبت کمترین فاصله بین خوشه ها به بیشترین میزان قطر خوشه محاسبه می شود.



$$D(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y)$$



$$diam(C_l) = \max_{x,y \in C_l} d(x, y)$$

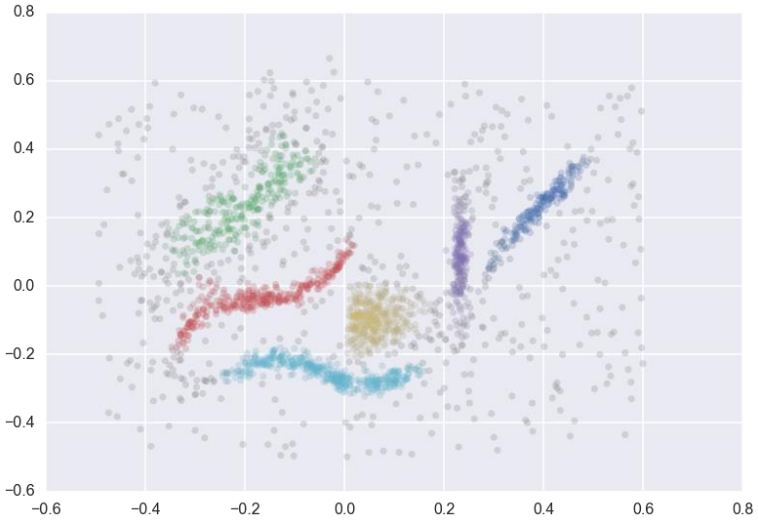
$$V_D = \left[\frac{\min_{i=1 \le j \le k} D(C_i, C_j)}{\max_{1 \le l \le k} diam(C_l)}\right]$$

مقدار شاخص دان در بازه صفر تا بی نهایت هست که بزرگ بودن آن به معنای تفکیک پذیری و در نتیجه خوشه بندی بهتر می باشد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

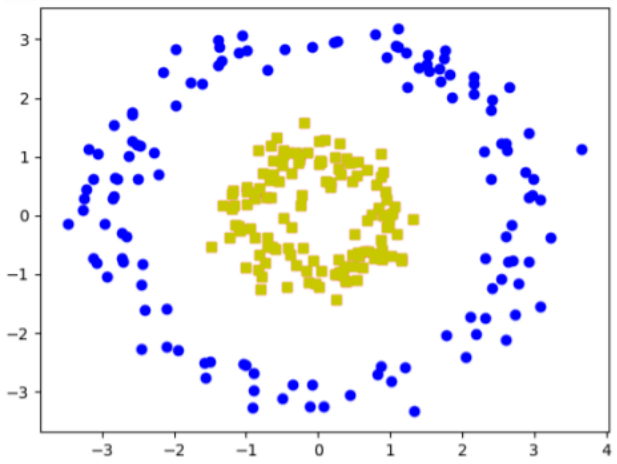
مدل های اکتشافی – خوشه بندی



معیار های ارزیابی خوشه بندی

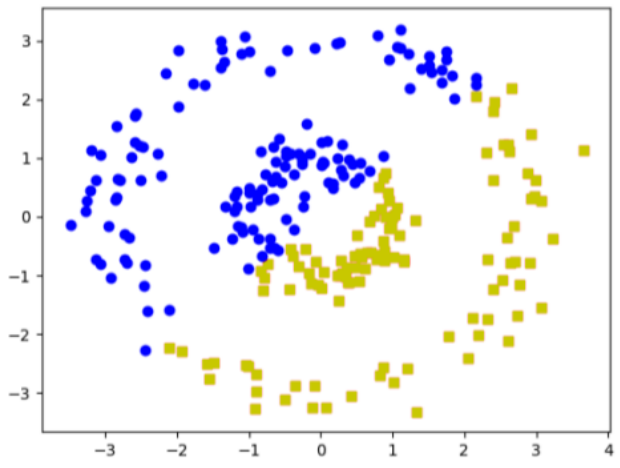
- معیار DBCV (Density Based Clustering Validation)

این شاخص به عنوان معیاری برای ارزیابی خوشه های پیچیده و غیر کروی و بر مبنای تراکم داده ها محاسبه می شود. ایده کلی این شاخص بر اساس ترکیب تراکم داده ها در داخل و بین خوشه ها هست. محاسبه این شاخص نسبت به سایر روش های گفته شده، دارای پیچیدگی محاسباتی و زمان بیشتر می باشد و همانند شاخص سیلوئت دارای دامنه 1- تا 1 بوده و برای هر خوشه امکان ارزیابی کیفیت جداگانه آن نیز وجود دارد.



silhouette: 0.1847

DBCV: 0.42204



silhouette: 0.2617

DBCV: -0.8969

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های اکتشافی – قوانین انجمنی

❑ مدل های اکتشافی (Explorative Models)

مدل های اکتشافی در فرآیند داده کاوی که با عنوان مدل های توصیفی (Descriptive Models) نیز شناخته می شود، در دسته یادگیری بدون نظارت قرار می گیرد.

خوشه بندی
Clustering

قوانین انجمنی
Association Rules

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
گروه دایکه | dayche.com

فرآیند داده کاوی

مدل های اکتشافی – قوانین انجمنی

❑ قوانین انجمنی (Association Rules)

الگوهای تکرار شونده، یکی از انواع الگوهای جذاب در مجموعه داده ها می باشد که شامل ترکیبی از اقلام یا اشیا است که به صورت مکرر، باهم یا در طول هم اتفاق می افتند. مانند مجموعه ای از اقلام در فروشگاه که بصورت مکرر باهم در سبد خرید مشتریان قرار می گیرد.(تحلیل سبد بازار: Market Basket Analysis)

الگوهای تکرار شونده به معنی وجود وابستگی در میان داده ها می باشد و به قوانینی که چنین روابطی را نشان می دهند، قوانین وابستگی یا قوانین انجمنی گفته می شود.

مثال 1: در صورت خرید تلفن همراه، با احتمال 80% محافظ صفحه نمایش هم خریداری می شود.

مثال 2: در صورت وجود ریسک فاکتورهای X و Y در بیمار، با احتمال 45% شانس وقوع عارضه Z پس از عمل جراحی قلب وجود دارد.

مثال 3: در صورت خرید یک دستگاه کامپیوتر شخصی توسط مشتری، با احتمال 35% پس از یک ماه برای خرید پرینتر خانگی برخواهد گشت.

مثال 4: در صورت وجود نشتی مایع خنک کننده از رادیاتور خودرو، با احتمال 65% ترموستات نیاز به تعمیر یا تعویض دارد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های اکتشافی – قوانین انجمنی

❑ قوانین انجمنی (Association Rules)

شناسایی الگوهای تکرار شونده، با جستجو در تراکنش های ثبت شده در پایگاه داده، به دنبال روابط تکراری بین اقلام تراکنش ها می باشد.

معمولا تراکنش ها، بصورت برداری از اقلام مورد بررسی با مقادیر بولین (Boolean) نمایش داده می شود و هدف الگوریتم، یافتن روابط تکراری در وقوع همزمان زیرمجموعه ای از اقلام و استخراج قوانین انجمنی می باشد.

TID	Transaction	a	b	c	d	e
1	{a , c}	1	0	1	0	0
2	{b , c , e}	0	1	1	0	1
3	{b , d , e}	0	1	0	1	1
4	{a , c , d}	1	0	1	1	0
5	{c , e}	0	0	1	0	1
6	{b}	0	1	0	0	0
7	{a , c , e}	1	0	1	0	1
8	{a , c , e}	1	0	1	0	1
9	{b , c , d}	0	1	1	1	0
10	{b , e}	0	1	0	0	1

بنابراین شناسایی الگو ها در این مسئله شامل دو گام اصلی می باشد:

- تعیین مجموعه اقلام مکرر (Frequent Itemset)
- استخراج قانون (Rule Extraction)

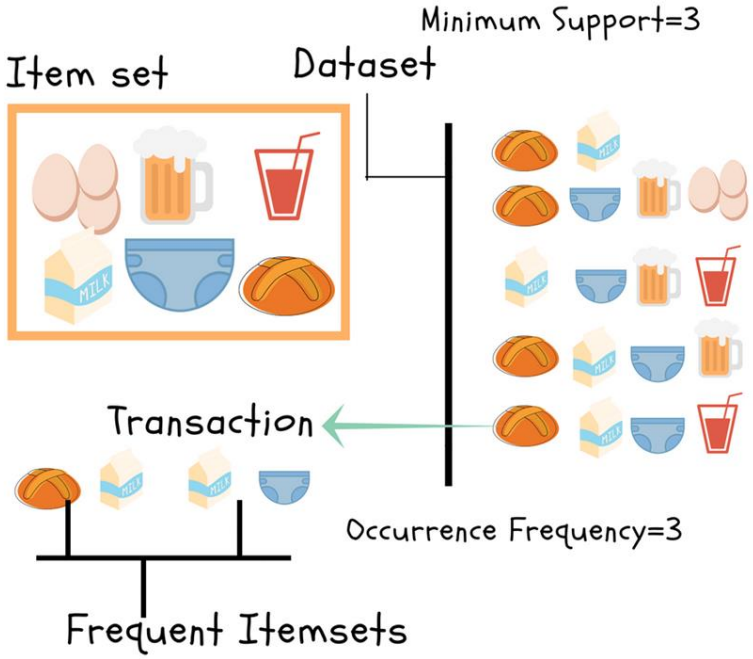
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های اکتشافی – قوانین انجمنی

❑ قوانین انجمنی (Association Rules)

ساده ترین روش تولید مجموعه اقلام مکرر اینست که تمامی ترکیب های ممکن بین اقلام در مجموعه داده ها، اسکن شده و فراوانی تکرار آنها مورد بررسی قرار گیرد. بدین معنی که برای n قلم کالا، بایستی دو به توان n حالت ممکن در مجموعه داده ها جستجو شود. طبیعی است هزینه محاسباتی این فرآیند در تعداد زیاد اقلام بسیار بالاست و نیاز به روش های سریعتر وجود دارد.



الگوریتم Apriori در سال 1994 به عنوان روشی سریع، جهت تولید قوانین انجمنی توسط آگراوال معرفی شد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ الگوریتم Apriori

این الگوریتم با هدف تولید قوانین انجمنی، در دو مرحله شناسایی مجموعه اقلام مکرر و تولید قوانین توسعه یافته است. در گام اول، به منظور شناسایی مجموعه اقلام مکرر، نیازمند دانستن پارامتر حداقل پشتیبانی (min Support) می باشد.

پشتیبانی (Support): نسبتی از تمامی تراکنش ها که در آنها یک قلم یا ترکیبی از اقلام مورد نظر وجود داشته باشد.

$$Support\,(X) = \frac{freq(X)}{N}$$

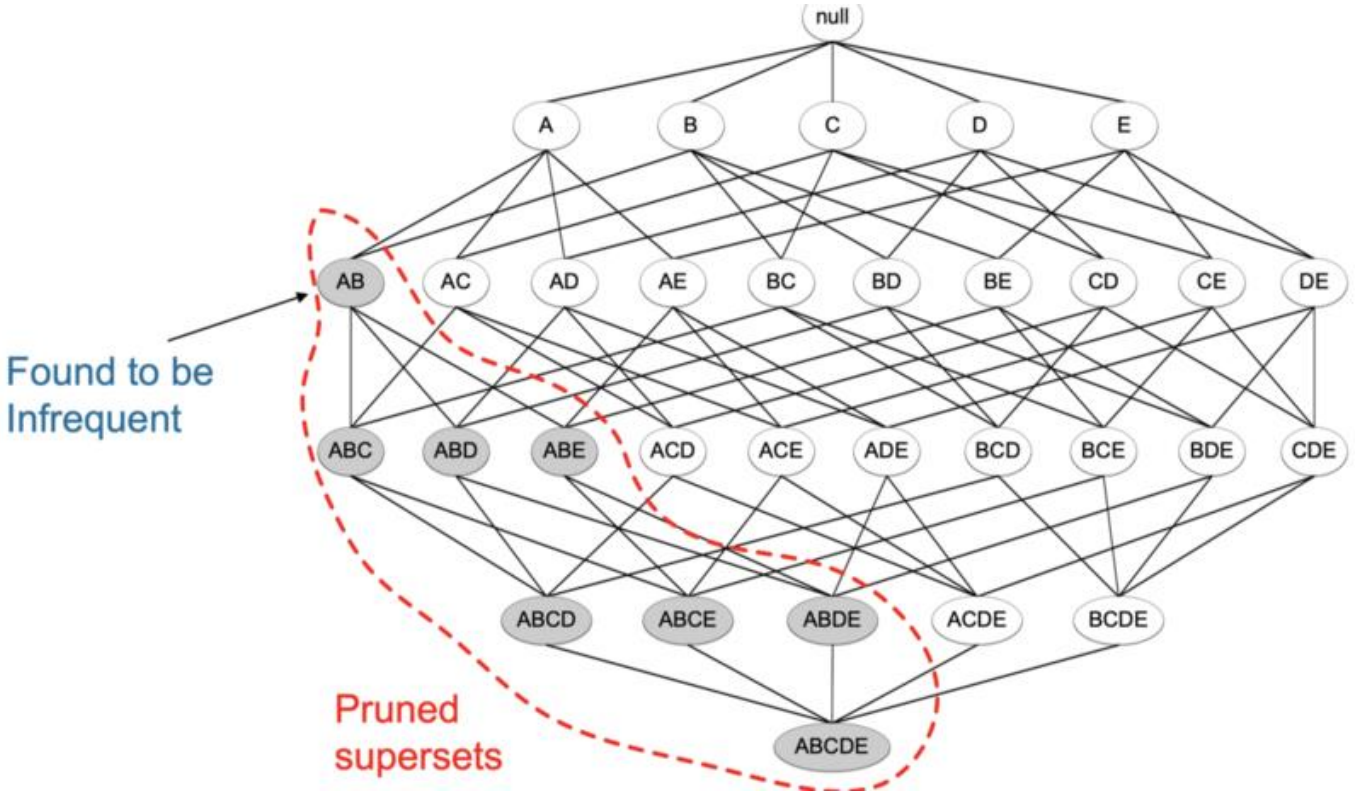
می توان معیار پشتیبانی را بصورت مطلق نیز بیان نمود. در این حالت فقط به تعداد تراکنش های شامل زیرمجموعه X اشاره می شود.

تعیین مقدار حداقل پشتیبانی، بر اساس تجربه افراد خبره و نیاز کسب و کار تعیین می گردد. بدیهی هست با انتخاب مقادیر بزرگتر، مجموعه اقلام کمتری دارای شرایط تعریف شده می شوند و فضای جستجوی الگوریتم کوچکتر شده و در زمانی سریعتر به تعداد محدودتری از قوانین انجمنی خواهیم رسید.

فرآیند داده کاوی

مدل های اکتشافی – قوانین انجمنی

❑ الگوریتم Apriori



اصل کلیدی در الگوریتم Apriori که منجر به کوچک شدن فضای جستجو و سرعت بالای شناسایی مجموعه اقلام مکرر می شود:

"هر زیرمجموعه غیرتهی از مجموعه اقلام مکرر، بایستی شرایط اقلام مکرر را داشته باشد."

بنابراین الگوریتم Apriori، از بررسی اقلام تکی شروع کرده و با محاسبه پشتیبانی آنها، در لایه بعدی صرفا به بررسی ترکیب دوتایی اقلام مکرر قبلی می پردازد و سپس ترکیب سه تایی و به همین صورت تا جاییکه ترکیب اقلام مکرر وجود داشته باشد، ادامه می دهد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های اکتشافی – قوانین انجمنی

❑ الگوریتم Apriori

TID	Transaction	a	b	c	d	e
1	{a , c}	1	0	1	0	0
2	{b , c , e}	0	1	1	0	1
3	{b , d , e}	0	1	0	1	1
4	{a , c , d}	1	0	1	1	0
5	{c , e}	0	0	1	0	1
6	{b}	0	1	0	0	0
7	{a , c , e}	1	0	1	0	1
8	{a , c , e}	1	0	1	0	1
9	{b , c , d}	0	1	1	1	0
10	{b , e}	0	1	0	0	1

مثال:

در مجموعه داده های روبرو، با در نظر گرفتن حداقل پشتیبانی برابر با مقدار 4،

مجموعه اقلام مکرر را بیابید:

$S(a) = 4$ +

$S(b) = 5$ +

$S(c) = 7$ +

$S(d) = 3$ ×

$S(e) = 6$ +

$S(a,b) = 0$ ×

$S(a,c) = 4$ +

$S(a,e) = 2$ ×

$S(b,c) = 2$ ×

$S(b,e) = 3$ ×

$S(c,e) = 4$ +

Frequent Itemset:

$\{a,c\}, \{c,e\}$

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ الگوریتم Apriori

در گام دوم پس از شناسایی مجموعه اقلام مکرر، قوانین انجمنی استخراج می گردد. قوانین بدست آمده، نوع وابستگی بین اقلام مکرر را نشان می دهد. قوانین انجمنی به شکل اگر - آنگاه (X –> Y) نشان داده می شود و به این معنی است که اگر X اتفاق بیفتد آنگاه Y نیز اتفاق خواهد افتاد. بدین منظور نیاز به تعیین مقدار پارامتر دوم الگوریتم یعنی حداقل اطمینان (min Confidence) می باشد.

اطمینان (Confidence): نسبتی از تراکنش های قسمت مقدم (X) که دارای اقلام قسمت موخر (Y) می باشد.

$$Confidence\,(X \rightarrow Y) = \frac{freq(X,Y)}{freq(X)}$$

این معیار در واقع میزان اطمینان از درستی قانون را اندازه می گیرد.

تعیین مقدار حداقل اطمینان، بر اساس تجربه افراد خبره و نیاز کسب و کار تعیین می گردد. بدیهی هست با انتخاب مقادیر بزرگتر، در زمانی سریعتر مجموعه قوانین کمتری استخراج خواهد شد.

فرآیند داده کاوی

مدل های اکتشافی – قوانین انجمنی

TID	Transaction	a	b	c	d	e
1	{a , c}	1	0	1	0	0
2	{b , c , e}	0	1	1	0	1
3	{b , d , e}	0	1	0	1	1
4	{a , c , d}	1	0	1	1	0
5	{c , e}	0	0	1	0	1
6	{b}	0	1	0	0	0
7	{a , c , e}	1	0	1	0	1
8	{a , c , e}	1	0	1	0	1
9	{b , c , d}	0	1	1	1	0
10	{b , e}	0	1	0	0	1

$S(a) = 4$ +
$S(b) = 5$ +
$S(c) = 7$ +
$S(d) = 3$ ×
$S(e) = 6$ +

$S(a,b) = 0$ ×
$S(a,c) = 4$ +
$S(a,e) = 2$ ×
$S(b,c) = 2$ ×
$S(b,e) = 3$ ×
$S(c,e) = 4$ +

$Frequent\ Itemset$:

$\{a,c\}, \{c,e\}$

سوال: آیا هر قانون انجمنی قوی، جذابیت نیز دارد؟

❑ الگوریتم Apriori

مثال:

در مجموعه داده های روبرو، با در نظر گرفتن حداقل اطمینان برابر با مقدار 0.6 قوانین انجمنی را استخراج نمایید:

$$Confidence(a \rightarrow c) = \frac{S(a,c)}{S(a)} = \frac{4}{4} = 1 \quad +$$

$$Confidence(c \rightarrow a) = \frac{S(a,c)}{S(c)} = \frac{4}{7} = 0.57 \quad \times$$

$$Confidence(c \rightarrow e) = \frac{S(c,e)}{S(c)} = \frac{4}{7} = 0.57 \quad \times$$

$$Confidence(e \rightarrow c) = \frac{S(c,e)}{S(e)} = \frac{4}{6} = 0.67 \quad +$$

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ الگوریتم Apriori

استخراج قوانین انجمنی قوی با دارا بودن حداقل پشتیبانی و حداقل اطمینان، همیشه منجر به تولید قوانین جذاب نمی شود. هدف از استخراج قوانین انجمنی، شناسایی روابط تکرارشونده و وابستگی بین اقلام می باشد بطوریکه نسبت به دانش اولیه موجود، ارزش افزوده ای ایجاد نماید؛ به این معنی که استفاده از قوانین بدست آمده، منجر به ارتقای شانس وقوع برخی اقلام شود.

ارتقا (Lift): نسبت اطمینان یک قانون انجمنی به احتمال وقوع اولیه.

$$Lift\,(X \rightarrow Y) = \frac{Confidence(X \rightarrow Y)}{P(Y)}$$

مقدار احتمال Y برابر با میزان پشتیبانی نسبی Y در مجموعه داده های اولیه می باشد.

بدیهی است هرچه مقدار Lift به عدد یک نزدیکتر باشد، نشان دهنده عدم جذابیت قانون است. معمولا به دنبال اعداد بزرگتر از یک هستیم که نشان می دهد شرایط گفته شده در قانون، شانس وقوع آن را ارتقا می دهد. همچنین اعداد کوچکتر از یک به معنی وابستگی منفی می باشد.

فرآیند داده کاوی

مدل های اکتشافی – قوانین انجمنی

❑ الگوریتم Apriori

مثال:

در مجموعه داده های روبرو، با محاسبه معیار ارتقا، قوانین جذاب را لیست کنید:

TID	Transaction	a	b	c	d	e
1	{a , c}	1	0	1	0	0
2	{b , c , e}	0	1	1	0	1
3	{b , d , e}	0	1	0	1	1
4	{a , c , d}	1	0	1	1	0
5	{c , e}	0	0	1	0	1
6	{b}	0	1	0	0	0
7	{a , c , e}	1	0	1	0	1
8	{a , c , e}	1	0	1	0	1
9	{b , c , d}	0	1	1	1	0
10	{b , e}	0	1	0	0	1

$$S(a) = 4$$
$$S(b) = 5$$
$$S(c) = 7$$
$$S(d) = 3$$
$$S(e) = 6$$

$$Confidence(a \rightarrow c) = \frac{S(a,c)}{S(a)} = \frac{4}{4} = 1$$

$$Confidence(e \rightarrow c) = \frac{S(c,e)}{S(e)} = \frac{4}{6} = 0.67$$

$$Lift(a \rightarrow c) = \frac{Confidence(a \rightarrow c)}{P(c)} = \frac{1}{0.7} = 1.43 \quad +$$

$$Lift(e \rightarrow c) = \frac{Confidence(e \rightarrow c)}{P(c)} = \frac{0.67}{0.7} = 0.96 \quad \times$$

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
گروه دایکه | dayche.com

فرآیند داده کاوی

مدل های اکتشافی – قوانین انجمنی

❑ الگوریتم Apriori

○ رتبه بندی قوانین

در مجموعه داده های بزرگ معمولا تعداد قوانین بسیار زیادی که دارای شرایط حداقلی در معیارهای پشتیبانی، اطمینان و ارتقا می باشند استخراج می گردد. همچنین با رجوع به منابع می توان معیارهای متنوع دیگری همچون کای – دو، کسینوسی، کولژیبنسکی و ... را مورد استفاده قرار داد. رتبه بندی قوانین بر اساس هریک از شاخص ها می تواند منجر به ترتیب های متفاوتی گردد. معمولا استفاده از افراد خبره و بررسی قوانین منتخب بر اساس هریک از شاخص ها، ایده انتخاب شاخص اصلی در رتبه بندی قوانین مسئله مورد نظر را بوجود می آورد.

○ بهینه سازی الگوریتم

علی رغم ایده تولید مجموعه اقلام مکرر که در الگوریتم Apriori بکار رفته است، اما هنوز در داده های بسیار بزرگ منجر به فضای جستجوی گسترده و کندی الگوریتم می گردد. بنابراین اغلب توسعه های انجام شده در این حوزه، روی ساختار مشابه الگوریتم Apriori و بهبود سرعت آن بوده است. الگوریتم FP-Growth توانسته بصورت کارآمدی در داده های بسیار بزرگ فرآیند شناسایی اقلام مکرر را بهبود داده و با سرعت بیشتری به خروجی برسد.