خوشه بندی

تمرین درس داده کاوی بخش چهارم – استاد مهردوست

کیان رضایی

مقدمه

- تعریف خوشه بندی
- چرا خوشه بندی؟
- نگاهی کلی به متد های خوشه بندی
- کاربرد های خوشه بندی
- ....

خوشه بندی - تعریف

- خوشه بندی به عنوان یکی از مهمترین تکنیک های یادگیری بدون نظارت در نظر گرفته میشود. قبل از آنکه درباره خوشه بندی توضیحاتی را ارائه دهیم اجازه دهید با مفهوم خوشه کمی بیشتر آشنا بشویم.

- خوشه : خوشه مجموعه ای از اشیا داده که دارای ویژگی های مشابه که در یک گروه یا کلاس قرار میگیرند و همچنین از سایر اشیا در باقی خوشه ها متفاوت اند.

خوشه بندی - تعریف

- همانطور که قبل تر گفته شد خوشه بندی یک تکنیک یادگیری بدون نظارت است که در آن خوشه ها (کلاس ها) از پیش تعریف شده و اطلاعات قبلی چگونگی گروه شدن یا برچسب خوردن کلاس ها را معین میکند.
- همچنین خوشه بندی میتواند برای تجزیه و تحلیل داده های اکتشافی که به ما در کشف الگو های پنهان و ساختار های داده کمک میکنند، در نظر گرفته شود.
- خوشه بندی میتواند به عنوان یک ابزار مستقل برای دادن آگاهی از توزیع های داده و همچنین مرحله پیش پردازش داده در سایر الگوریتم ها به کار رود.

چرا خوشه بندی؟

- خوشه بندی به ما این امکان را میدهد که بتوانیم روابط پنهان میان داده های هدف و دیتاست ها را پیدا کنیم.
- در ادامه چند مثال از چرایی استفاده از خوشه بندی میزنیم...

خوشه بندی - مثال

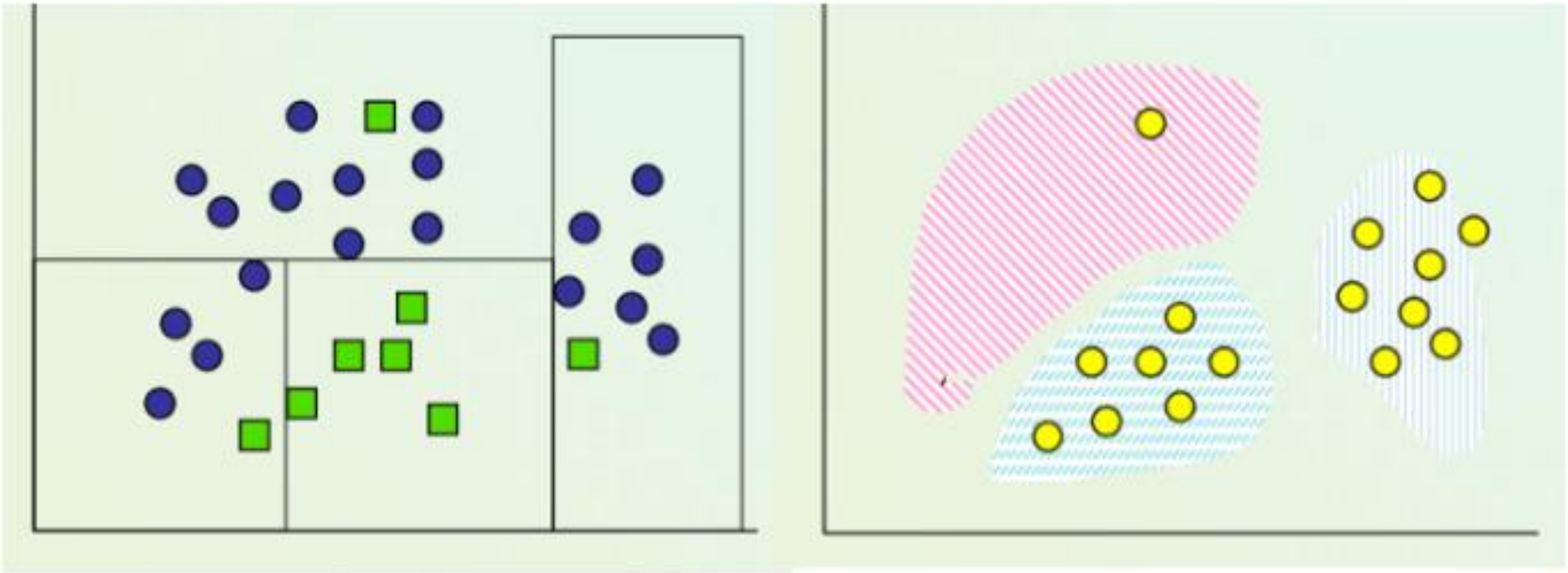
- در حوزه کسب و کار، مشتریان با توجه به بازار هدف خود دسته بندی میشوند. خوشه بندی میتواند با دسته بندی کردن مشتریان کمک شایانی به صاحبان کسب و کار بکند.
- خوشه بندی در نوشته _ فرض کنید مجموعه از نوشته ها (Text) را در اختیار داریم. به وسیله خوشه بندی میتوانیم این نوشته ها را بر اساس موضاعاتشان مرتب کنیم.
- یافتن الگوها در عکس – این زمینه به شدت حوضه زیست شناسی را در برگرفته است.

و بسیاری مثال های دیگر به اهمیت خوشه بندی می افزاید...

خوشه بندی vs دسته بندی

- بیایید درک کنیم که چرا دسته بندی یک تکنیک یادگیری با نظارت و خوشه بندی یک تکنیک یادگیری بدون نظارت است.
- در یادگیری با نظارت مدل ما یک متد برای پیش بینی یک نمونه کلاس که از پیش تعریف شده (برچسب دارد) را یاد میگیرد.
- در یادگیری بدون نظارت مدل ما تلاش میکند که به صورت **طبیعی** نمونه ها را برای داده های بدون برچسب را گروه بندی کند.

خوشه بندی vs دسته بندی



Classification vs clustering

انواع متد های خوشه بندی

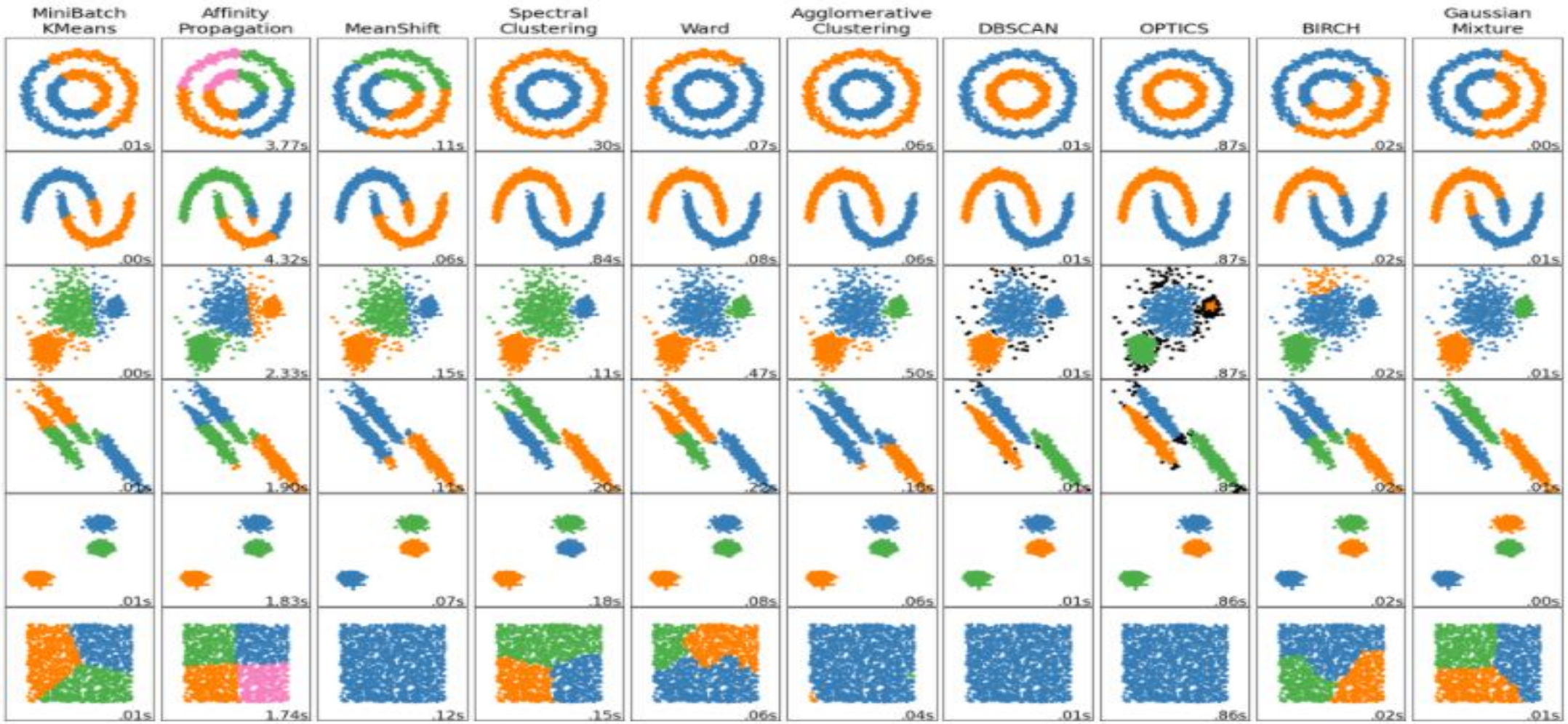
1. K-Means
2. Affinity propagation
3. Mean-Shift
4. Spectral clustering
5. Agglomerative clustering

انواع متد های خوشه بندی

6. DBSCAN
7. OPTICS
8. Gaussian mixtures
9. BIRCH

…

مقایسه الگوریتم ها خوشه بندی در Scikit-learn



A comparison of the clustering algorithms in scikit-learn

انواع متد های خوشه بندی

- در ادامه دو روش خوشه بندی را به تفضیل شرح خواهیم داد و در فایل پیاده سازی ها این الگوریتم ها رو پیاده سازی خواهیم کرد...

انواع متد های خوشه بندی – K-Means

- این الگوریتم پارامتر k را به عنوان ورودی گرفته و مجموعه n شی را به k خوشه افراز میکند. به طوریکه سطح شباهت داخلی خوشه ها بالا بوده و سطح شباهت اشیا بیرون خوشه ها پایین باشد. شباهت هر خوشه نسبت به متوسط اشیا آن خوشه سنجیده شده که این متوسط، مرکز خوشه نیز نامیده میشود. این الگوریتم به صورت زیر کار میکند:
- ورودی : k، تعداد خوشه ها و یک پایگاه داده شامل n شی
- خروجی : یک مجموعه از k خوشه که معیار مربع خطا را حداقل میکند.

الگوریتم K-Means

- قدم ۱) به صورت تصادفی k نقطه دلخواه را به عنوان مراکز خوشه های ابتدایی انتخاب کن.
- قدم ۲) هر شی را با توجه به بیشترین شباهت آن به مراکز خوشه ها، به خوشه ها تخصیص بده.
- قدم ۳) مراکز خوشه ها را به روز کن به این معنی که برای هر خوشه میانگین اشیا آن خوشه را محاسبه کن
- قدم ۴) با توجه به مراکز جدید خوشه ها به قدم دوم برگرد تا هنگامی که هیچ تغییری در خوشه ها رخ ندهد. (در این حالت الگوریتم پایان یافته است.)

الگوریتم K-Means

- روش K-means تنها هنگامی کاربرد دارد که بتوان مراکز خوشه ها را تعریف کـرد. مـثلا برای داده هایی با ویژگیهای طبقه ای این روش کارا نیست. از معایب ایـن روش تعیـین K است که میبایست کاربر آن را معیین کند و راه خاصی برای تعیین آن مشخص نشده اسـت. یک راه امتحان k های مختلف و بررسی معیار مربع خطا برای هر k میباشد. همچنـین ایـن روش برای کشف خوشه هایی با شکلهای پیچیده مناسب نیست. یکـی از مهمتـرین نقـاط ضعف این روش این است که در برابر اغتشاشات و نقاط پرت حساس است زیرا ایـن داده ها به راحتی مراکز را تغییر می دهند و ممکن است نتایج مطلوبی حاصل نشود.

الگوریتم K-Means

- نکته بسیار مهم درالگوریتم های خوشه بندی تشخیص شباهت ها و عدم شباهت ها است.
- برای آنکه شباهت و عدم شباهت دو داده را تشخیص دهیم از تابع فاصله استفاده میکنیم. یک تابع فاصله پرکاربرد، تابع فاصله اقلیدسی است.
- تابع فاصله اقلیدسی : دو داده X_1 و X_2 را در نظر داشته باشید. که هر کدام n ویژگی داشته باشند.

$$X_1 = (x_1, x_2, x_3, \dots., x_n)$$
$$X_2 = (x_1, x_2, x_3, \dots., x_n)$$

در این صورت تابع اقلیدسی به صورت زیر تعریف میشود :

$$dist(X_1, X_2) = \sqrt{\sum_{i=1} ((x_{1i} - x_{2i}))^2}$$

$$1 \leq i \leq n$$

مثال

- فرض کنید دو قلم داده زیر را دارید. در این صورت فاصله این دو داده برابر است با :

Customer 1		
Age	Income	education
54	190	3

Customer 2		
Age	Income	education
50	200	8

$$\text{Dis}\,(x_1, x_2) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2}$$

$$= \sqrt{(54-50)^2+(190-200)^2+(3-8)^2} = 11.87$$

یک مثال از الگوریتم K-means

مثال ۱: به فرض مجموعهٔ {۲,٤,۱۰,۱۲,۳,۲۰,۳۰,۱۱,۲۵}را می‌خواهیم به $k=۲$ خوشه افراز کنیم. با استفاده از روش *k-means* مراحل زیر را طی می‌کنیم:

به‌طور تصادفی دو مرکز $m_۲=٤$ و $m_۱=۲$ را انتخاب کرده و بقیهٔ اعضاء مجموعه را با توجه به فاصلهٔ آنها از این دو مرکز تخصیص می‌دهیم. یعنی هر عضو را به خوشهای تخصیص میدهیمکه به مرکز آن نزدیکتر باشند. خوشه‌های حاصل عبارتند از:

$K_۱=\{۲,۳\}$ $K_۲=\{٤,۱۰,۱۲,۲۰,۳۰,۱۱,۲۵\}$

حال مراکز جدید را محاسبه می‌کنیم و تخصیص را نسبت به مراکز جدید انجام می‌دهیم. (مراکز در این مثال میانگین اعداد هر دسته می‌باشد):

$m_۱=۲/۵$, $m_۲=۱۶$

یک مثال از الگوریتم K-means -- ادامه

ادامه مثال ۱ خوشه‌های جدید عبارتند از:

$K_1 = \{2,3,4\}$, $K_2 = \{10,12,20,30,11,25\}$

روند فوق را آنقدر تکرار می‌کنیم تا اینکه دیگر تغییری در خوشه‌ها رخ ندهد:

$m_1 = 3$, $m_2 = 18$

$K_1 = \{2,3,4,10\}$, $K_2 = \{12,20,30,11,25\}$

$m_1 = 4.75$, $m_2 = 19.6$

$K_1 = \{2,3,4,10,11,12\}$, $K_2 = \{20,30,25\}$

$m_1 = 7$, $m_2 = 25$

$K_1 = (2,3,4,10,11,12\}$, $K_2 = \{20,30,25\}$

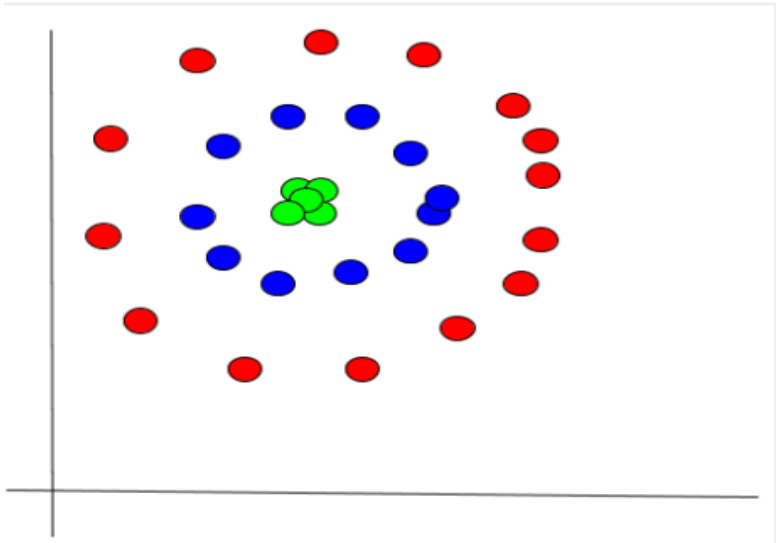
در این مرحله دیگر تغییری در خوشه‌ها رخ نمی‌دهد. لذا دو خوشهٔ فوق به‌دست آمده است والگوریتم خاتمه می‌یابد.

مزایای استفاده از روش k-means

۱. **دیتاست های بدون برچسب** : بسیاری از داده های جهان حقیقی بدون برچسب و بدون هیچگونه کلاس مشخص هستند. خوبی استفاده از الگوریتمی مانندk-means این است که ما گاهی نمیدانیم که چگونه داده ها در دیتاست باید گروه بندی شوند. برای مثال، در نظر بگیرید که شما میخواهید تماشاگران netflix را با توجه به شباهت ویدیوهای دیده خوشه بندی کنید. مدل های خطی نیز به هیچ وجه در این گونه مسائل به ما کمک نخواهد.

مزایای استفاده از روش k-means

۲. **جداسازی غیر خطی داده** : دیتاست زیر را شامل سه دسته از دوایر متحد المرکز را در نظر بگیرید. خوشه ها به صورت غیر خطی از یکدیگر هستند، به عبارت دیگر یعنی هیچ خط یا صفحه ای در فضا نمیتواند این جداسازی را انجام دهد. استفاده از الگوریتم k-means و تغییر متخصات کارتزین به مختصات قطبی این امکان را به ما میدهد که اطلاعات مربوط به شعاع ها را به دست آوردیم و بدین وسیله خوشه های محتدالمرکز را تشکیل دهیم.



مزایای استفاده از روش k-means

۳. آسانی : بدین معنی زیرا که تنها در دو مرحله اجرا میشود. به عنوان یک الگوریتم از الگوریتم های یادگیری بدون نظارت برای پیاده سازی بسیار آسان هست و میتواند دیتاست ها با حجم بالا را مدیریت کند.

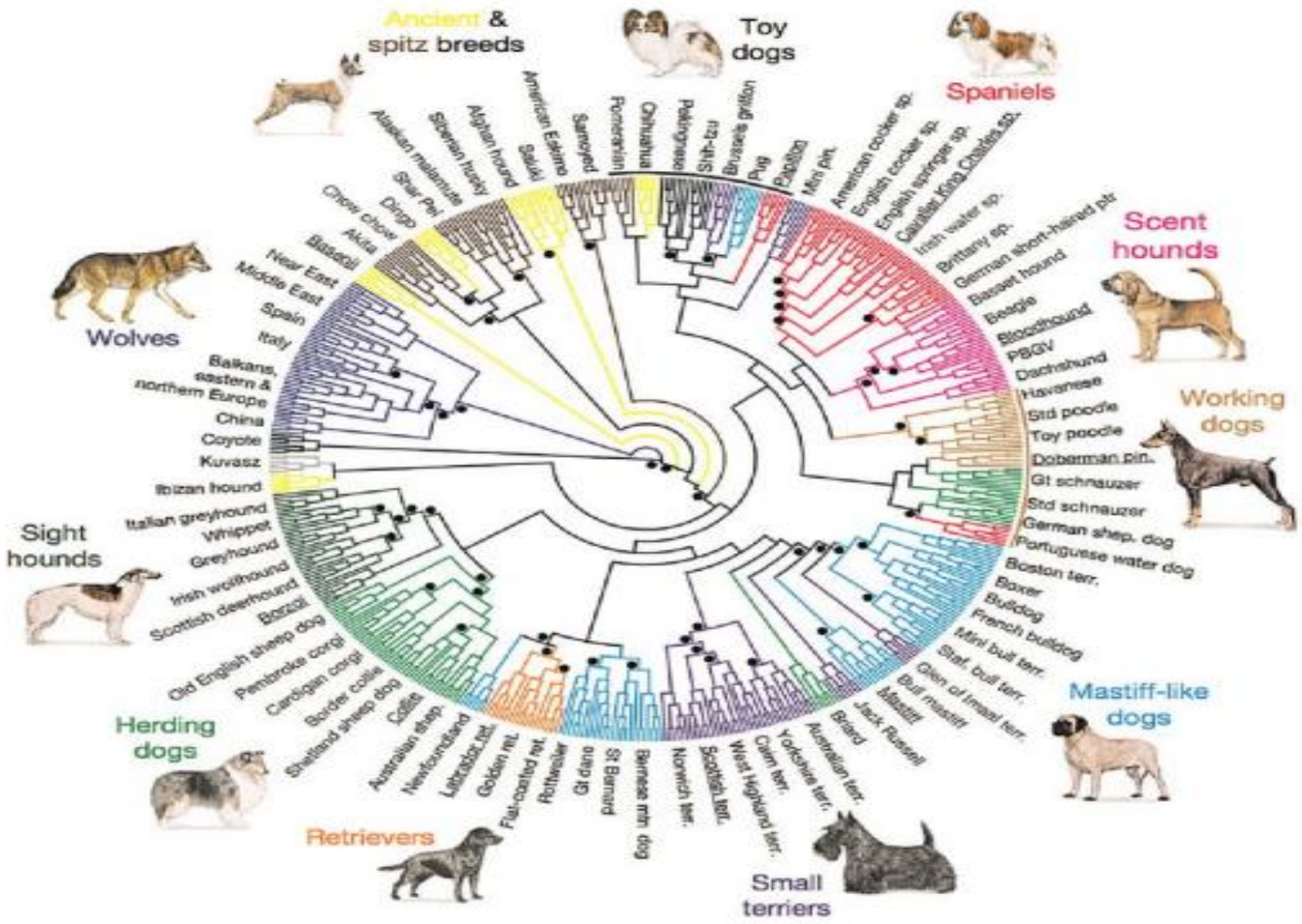
۴. دسترسی

۵. سرعت اجرا

انواع متد های خوشه بندی – سلسله مراتبی

- شکل زیر را در نظر بگیرید.
- از ۴۸۰۰۰ ژن این چارت بر اساس شباهت ها تشکیل شده است.
- هدف در این الگوریتم این است که یک سلسله مراتب از خوشه ها درست کنیم که در آن فرزندان یک خوشه شامل خوشه های فرزندان خود میشود.

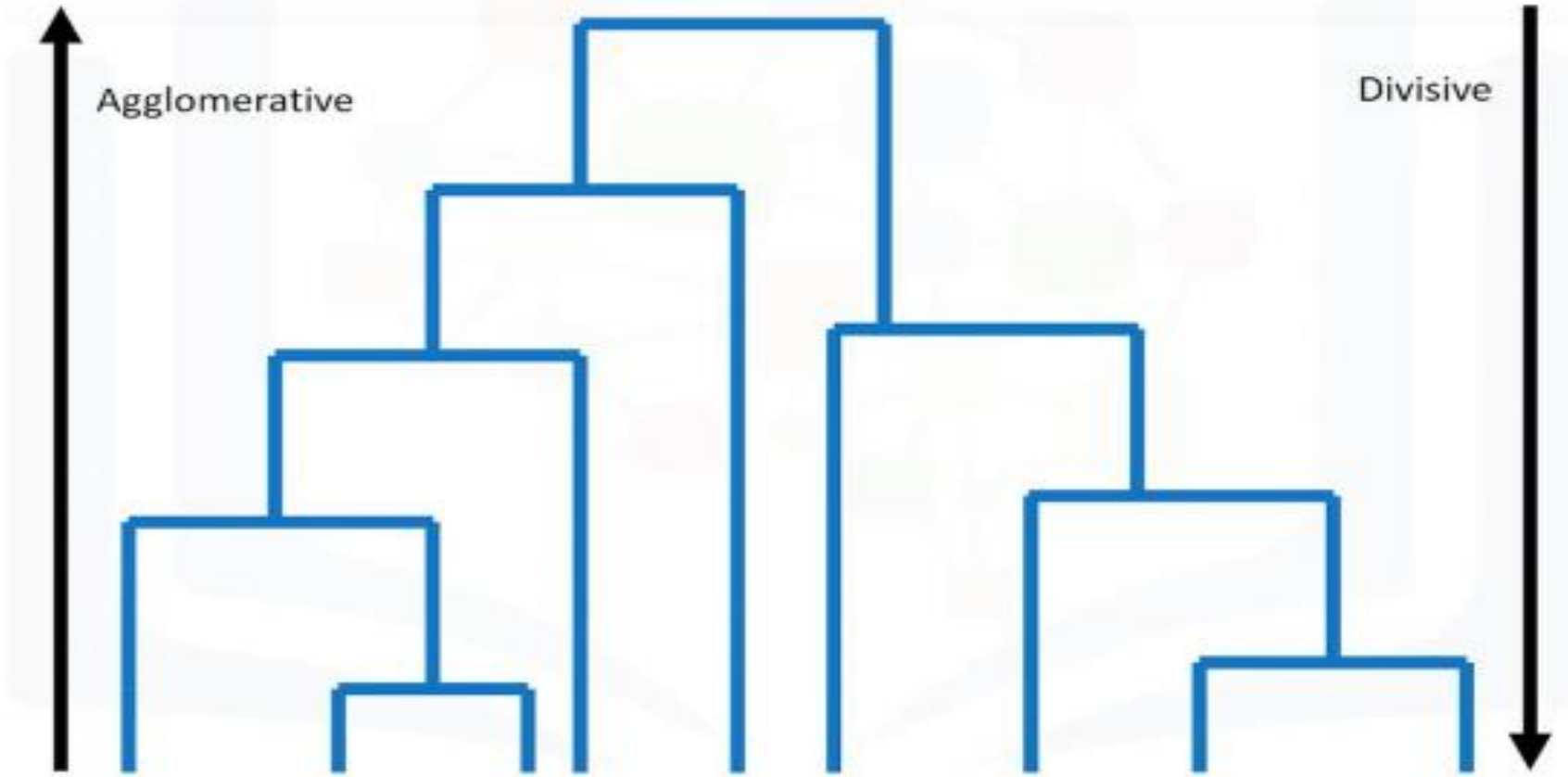
انواع متد های خوشه بندی – سلسله مراتبی



انواع متد های خوشه بندی – روش های سلسله مراتبی

این روش ساختاری سلسله مراتبی از اشیا یک مجموعه معلوم ایجاد میکند. روش سلسله مراتبی میتواند خوشه بندی را به صورت تجمیعی و یا به صورت تقسیمی انجام دهد. به رویکرد تجمیعی، رویکرد پایین به بالا نیز گفته میشود. این روش با شکل دهی گروه های مجزا که هر یک شامل حداقل یک شی میباشند شروع میشود. سپس اشیا یا گروه های نزدیک به هم را یکی میکند تا این که در نهایت یک گروه کلی در بالاترین سطح ایجاد شود. در روش تقسیمی کل اشیا در یک خوشه در نظر گرفته شده و در هر تکرار یک خوشه به دو خوشه کوچکتر تقسیم میشوند.

انواع متد های خوشه بندی – روش های تجمیعی و تقسیمی



مثال

- فرض کنید شما به عنوان یک رئیس وظیفه استان بندی استان های کشور کانادا را دارید.

هر شهر را یک خوشه در نظر میگیریم. ابتدا ماتریس فاصله را تشکیل میدهیم. دو استان که کمترین فاصله را از یکدیگر دارند را به عنوان یک خوشه درنظر میگیریم.

TO OT MO VA ED WI

	TO	OT	VA	MO	WI	ED
TO		351	3363	505	1510	2699
OT			3543	167	1676	2840
VA				3690	1867	819
MO					1824	2976
WI						1195
ED						

ادامه مثال...

- در ادامه قصد داریم به روش تجمیعی این سلسله مراتب را تشکیل دهیم. به علت نزدیکی استان اتاوا با مونترال آن را در یک خوشه قرار میدهیم و مجدد ماتریس فاصله را محاسبه میکنیم.
- حال که دو استان اتاوا و مونترال را در یک خوشه قرار دادیم، مسئله این است حال فاصله این خوشه با سایر خوشه های باقی مانده چقدر خواهد بود. تعریف های بسیاری برای این مسئله وجود دارد که به domain expert این قضیه مربوط میشود. ما در اینجا نصف فاصله بین دو شهر را به عنوان (مرکز) در نظر میگیریم.

ادامه مثال...

TO OT MO VA ED WI

	TO	OT/MO	VA	WI	ED
TO		351	3363	1510	2699
OT/MO			3543	1676	2840
VA				1867	819
WI					1195
ED					

ادامه مثال...

- این ایجاد خوشه را هر بار انجام میدهیم.

	TO/OT/MO	VA	WI	ED
TO/OT/MO		3543	1676	2840
VA			1867	819
WI				1195
ED				

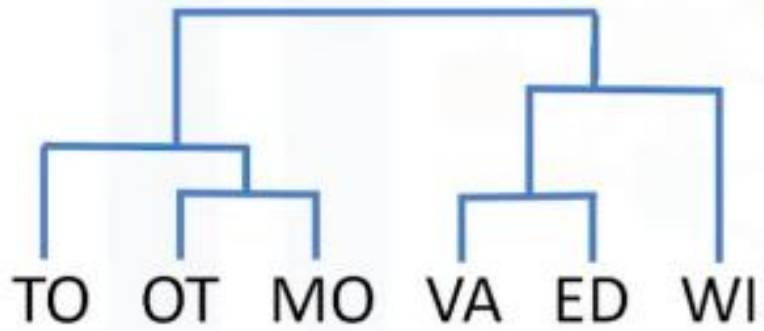
ادامه مثال...

	TO/OT/MO	VA	WI	ED
TO/OT/MO		3543	1676	2840
VA			1867	819
WI				1195
ED				

ادامه مثال...

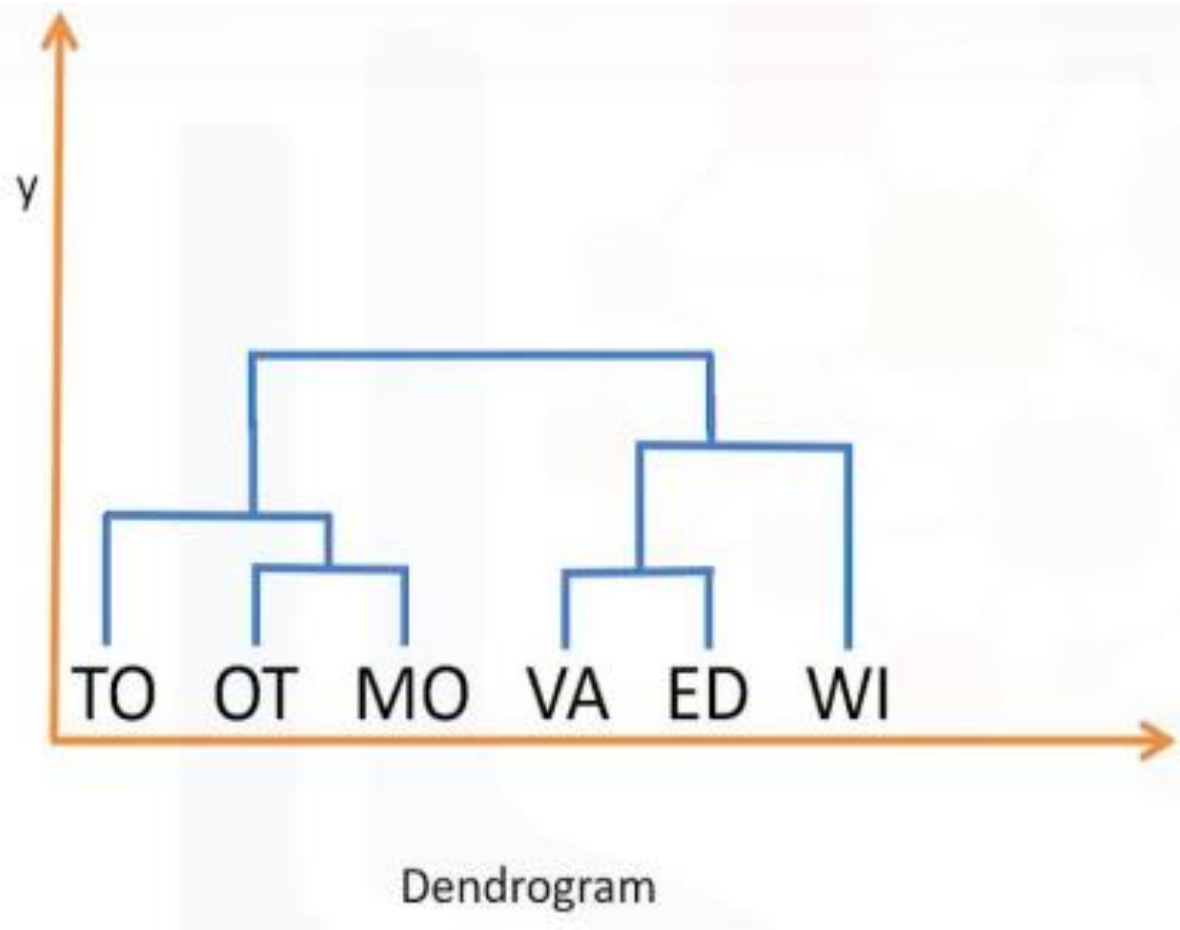
	TO/OT/MO	VA/ED	WI
TO/OT/MO		2840	1676
VA/ED			1667
WI			

ادامه مثال...



	TO/OT/MO	VA/ED/WI
TO/OT/MO		1676
VA/ED/WI		

ادامه مثال...



Dendrogram

الگوریتم سلسله مراتبی

1. Create *n* clusters, one for each data point
2. Compute the Proximity Matrix
3. **Repeat**
    i. Merge the two closest clusters
    ii. Update the proximity matrix
4. **Until** only a single cluster remains

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

مزایا و معایب متد سلسله مراتبی

- **مزایا :**

۱. **با n شی ناشناس کار میکند. (یادگیری بدون نظارت)**
۲. **پیاده سازی این الگوریتم ساده است.**
۳. **سلسله مراتب ایجاد شده، بسیار مناسب برای فهمیدن است.**

- **معایب :**

۱. **توانایی undo کردن را نداریم. یعنی اینکه این الگوریتم به شما سلسله مراتب را میدهد و از روند ایجاد این سلسله مراتب ما بی خبریم و توانایی دستکاری آن را نداریم.**
۲. **زمان اجرای طولانی دارد.**
۳. **در برخی موارد شناسایی تعداد خوشه ها بسیار سخت خواهد بود. (به ویژه برای دیتاست ها بزرگ)**

سلسله مراتب vs k-means

- سلسله مراتبی میتواند زمان اجرای کند تری داشته باشد.
- روش سلسله مراتبی نیازی به تعداد خوشه (k) نیازی ندارد.
- روش سلسله مراتبی بخش بندی مناسب تری نسبت به k-means را ارائه میکند.
- هر چند باری که روش سلسله مراتبی اجرا شود، دقیقا جواب یکسانی خواهد داد. چرا؟
- مثال بالا را در نظر بگیرید، فاصله شهر ها در هر بار اجرا تغییر نمیکنند برای همین است که در هر بار اجرا جواب یکسانی را تولید میکند. اما در روش k-means با هر بار اجرا جواب متفاوتی ممکن است به ما بدهد.

منابع

- Amazon.com: Data Clustering: Algorithms and Applications (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series): 0001466558210: Aggarwal, Charu C., Reddy, Chandan K.: Books
- Clustering: Wunsch, Don, Xu, Rui: 978047027680 8: Amazon.com: Books
- Cluster analysis – Wikipedia
- 2.3. Clustering — scikit-learn 1.0.2 documentation
- Clustering in Machine Learning – GeeksforGeeks
- kiyan-rezaee/machine_learning_with_python_jadi: The notebooks we use on ML course (github.com)