دانشگاه سمنان
Semnan University
پردیس فرزانگان

بسمه تعالی

فصل پنجم داده کاوی

خوشه بندی
Clustering

مدرس

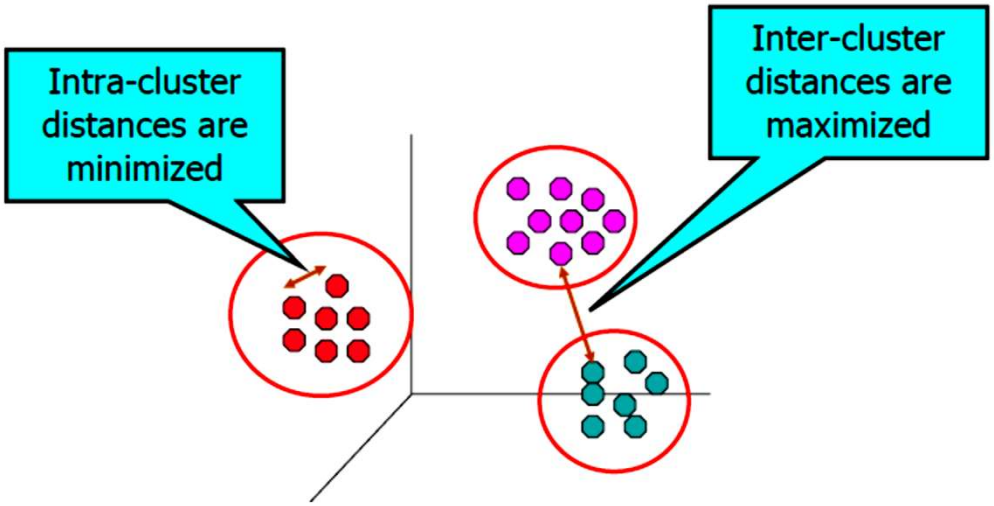
فاطمه دارائی

f_daraei@semnan.ac.ir

https://fdaraei.profile.semnan.ac.ir

تحلیل خوشه ای چیست؟

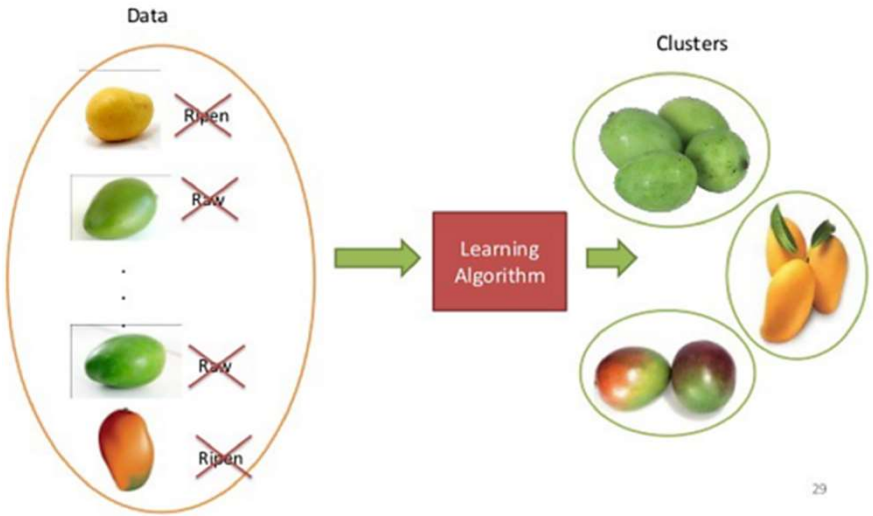
- خوشه‌بندی: مجموعه‌ای از اشیاء داده
- **شباهت (یا ارتباط)** بین اشیاء درون یک گروه: **شباهت درون‌کلاسی بالا**
- **عدم شباهت (یا عدم ارتباط)** با اشیاء گروه‌های دیگر: **شباهت بین‌کلاسی پایین**
- **چه اتفاقی می‌افتد اگر هر یک از این دو شرط را کاهش دهیم؟**
- **تحلیل خوشه‌ای (یا خوشه‌بندی، تقسیم‌بندی داده‌ها و ...)**
- یافتن شباهت‌ها بین داده‌ها براساس ویژگی‌های موجود در داده‌ها
- گروه‌بندی اشیاء مشابه داده به خوشه‌ها



یادگیری با نظارت Supervised learning

- **یادگیری بدون نظارت:**
- **کلاس‌های از پیش تعریف‌شده وجود ندارد** (یعنی یادگیری از طریق مشاهده به جای یادگیری از طریق مثال: یادگیری نظارت‌شده).
- **کاربردهای معمول:**

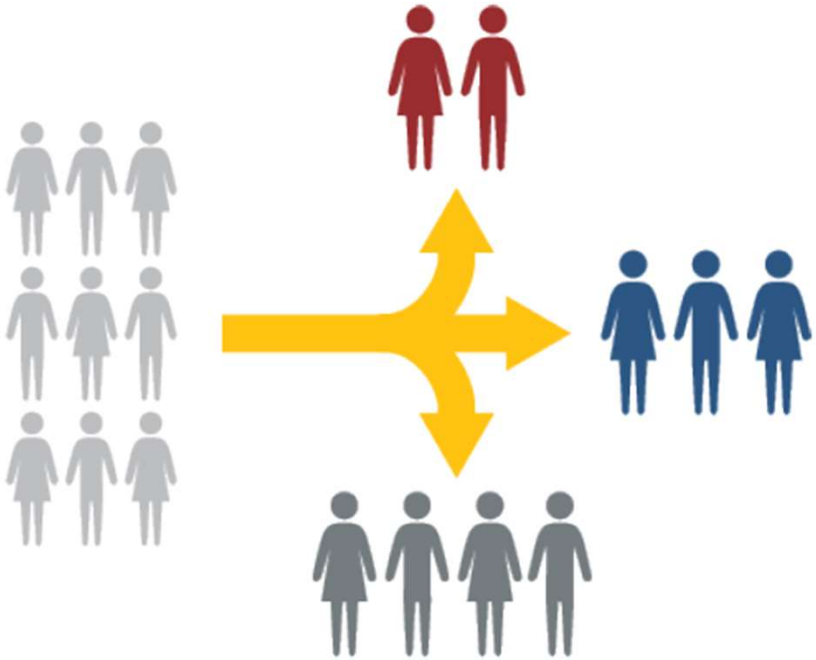
1. به‌عنوان یک ابزار مستقل برای **درک توزیع داده‌ها**
2. به‌عنوان یک مرحله **پیش‌پردازش برای الگوریتم‌های دیگر**



https://www.slideshare.net/sachinnagargoje1/introduction-to-machinelearningatsapthgiricollegebangalore

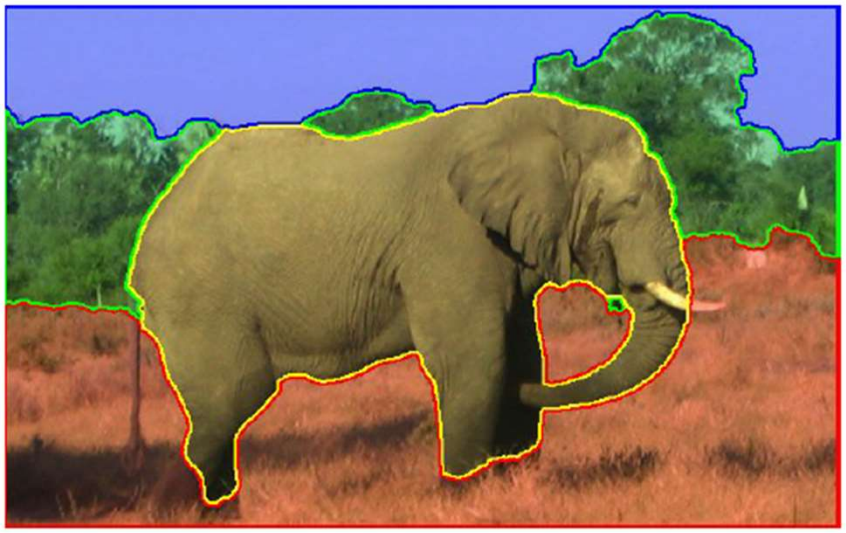
دانشگاه سمنان
Semnan University
پردیس فرزانگان

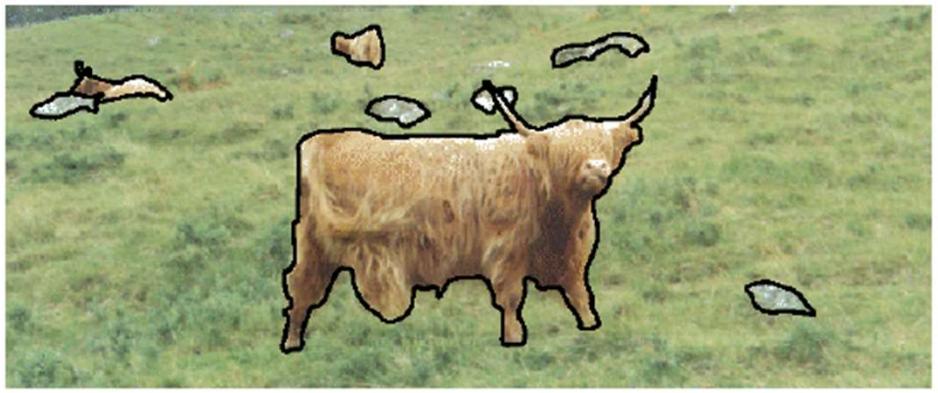
مثال: تقسیم بندی مشتری



دانشگاه سمنان
Semnan University
پردیس فرزانگان

مثال: تقسیم بندی تصویر

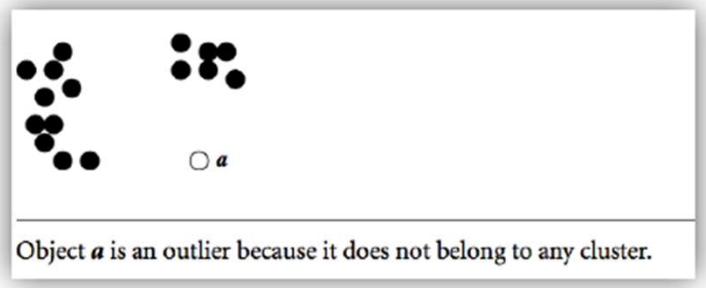




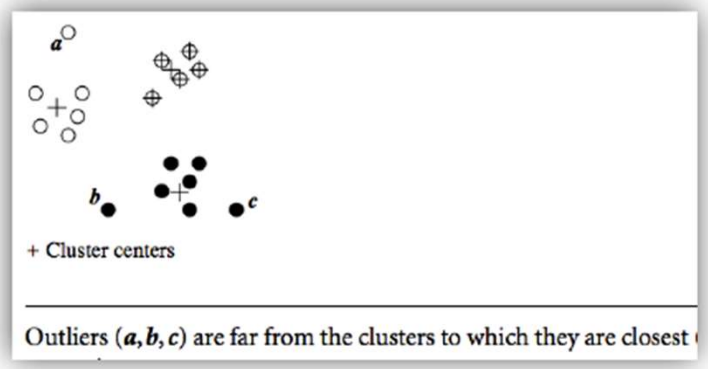
مثال: شناسایی داده های پرت

- **نقاط پرت Outliers چیست؟**
- مجموعه‌ای از اشیاء که به طور قابل توجهی با بقیه داده‌ها **متفاوت هستند.**
- **کاربردهای تشخیص نقاط پرت:**
- **تشخیص تقلب**، مانند تشخیص تقلب در کارت‌های اعتباری.
- به عنوان یک ابزار **پیش‌پردازش** نیز مفید است.
- **یکی از روش‌های تشخیص نقاط پرت با استفاده از خوشه‌بندی:**
- **اشیائی که به هیچ خوشه‌ای تعلق ندارند.**
- **اشیائی که فاصله زیادی از سایر اشیاء در همان خوشه دارند.**
- **خوشه‌هایی با تعداد اعضای بسیار کم**

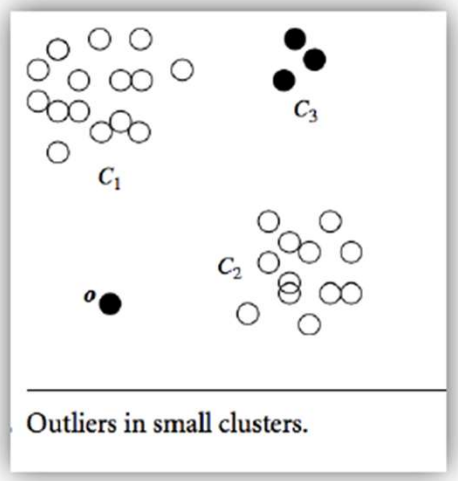
مثال: شناسایی داده های پرت



Object ***a*** is an outlier because it does not belong to any cluster.



Outliers (a, b, c) are far from the clusters to which they are closest



Outliers in small clusters.

الگوریتم های Partitioning

نسخه‌ای ساده و بنیادی از تحلیل خوشه‌بندی هستند.

اشیای یک مجموعه را به چند گروه یا خوشه‌ی مجزا سازمان‌دهی می‌کند.

خوشه‌ها به گونه‌ای تشکیل می‌شوند که یک معیار هدف برای تقسیم‌بندی بهینه شود.

معیار تقسیم‌بندی: تغییرات درون خوشه‌ای Within cluster variation:

تقسیم یک پایگاه داده D شامل N شیء به مجموعه‌ای از k خوشه، به طوری که مجموع فواصل به توان دو کمینه شود (که در آن ci مرکز یا نقطه میانی خوشه Ci است)

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

با توجه به k، یک تقسیم‌بندی از k خوشه پیدا کنید که معیار تقسیم‌بندی انتخاب‌شده را بهینه کند:
بهینه‌ی کلی Global: تمام تقسیم‌بندی‌ها را به صورت کامل بررسی کنید.

روش‌های اکتشافی Heuristic:
الگوریتم‌های K-means و K-medoids

K-means :خوشه توسط مرکز خوشه نمایش داده می‌شود.
K-medoids تقسیم‌بندی حول میانه‌ها: هر خوشه توسط یکی از اشیای موجود در خوشه نمایش داده می‌شود.

K-means الگوریتم

Algorithm: *k*-means. The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

Method:

(1) arbitrarily choose *k* objects from *D* as the initial cluster centers;
(2) **repeat**
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, that is, calculate the mean value of the objects for each cluster;
(5) **until** no change;

مثال الگوریتم K-means

TABLE 10.1 Data points for *k*-means example

A	*b*	*c*	*d*	*E*	*f*	*g*	*h*
(1,3)	(3,3)	(4,3)	(5,3)	(1,2)	(4,2)	(1,1)	(2,1)

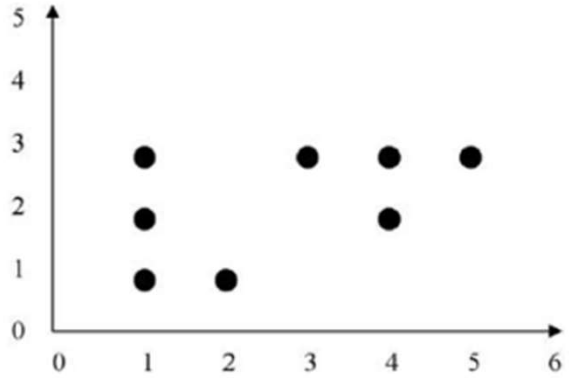


Figure 10.4 How will *k*-means partition these data into $k = 2$ clusters?

حل مثال الگوریتم K-means

TABLE 10.1 Data points for *k*-means example

A	*b*	*c*	*d*	*E*	*f*	*g*	*h*
(1,3)	(3,3)	(4,3)	(5,3)	(1,2)	(4,2)	(1,1)	(2,1)

m1=(1,1)
m2=(2,1)

TABLE 10.2 Finding the nearest cluster center for each record (first pass)

Point	Distance from m_1	Distance from m_2	Cluster Membership
a	2.00	2.24	C_1
b	2.83	2.24	C_2
c	3.61	2.83	C_2
d	4.47	3.61	C_2
e	1.00	1.41	C_1
f	3.16	2.24	C_2
g	0.00	1.00	C_1
h	1.00	0.00	C_2

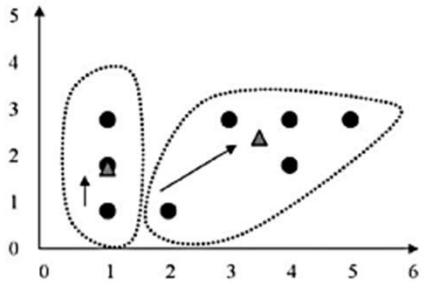


Figure 10.5 Clusters and centroids Δ after first pass through *k*-means algorithm.

ادامه حل مثال الگوریتم K-means

۱

TABLE 10.1 Data points for *k*-means example

A	*b*	*c*	*d*	*E*	*f*	*g*	*h*
(1,3)	(3,3)	(4,3)	(5,3)	(1,2)	(4,2)	(1,1)	(2,1)

m1=(1,2)
m2=(3.6,2.4)

TABLE 10.3 Finding the nearest cluster center for each record (second pass)

Point	Distance from m_1	Distance from m_2	Cluster Membership
a	1.00	2.67	C_1
b	2.24	0.85	C_2
c	3.16	0.72	C_2
d	4.12	1.52	C_2
e	0.00	2.63	C_1
f	3.00	0.57	C_2
g	1.00	2.95	C_1
h	1.41	2.13	C_1

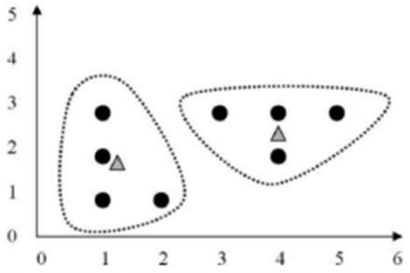


Figure 10.6 Clusters and centroids Δ after second pass through *k*-means algorithm.

ادامه حل مثال الگوریتم K-means

TABLE 10.1 Data points for *k*-means example

A	*b*	*c*	*d*	*E*	*f*	*g*	*h*
(1,3)	(3,3)	(4,3)	(5,3)	(1,2)	(4,2)	(1,1)	(2,1)

m1=(1.25, 1.75)
m2=(4, 2.75)

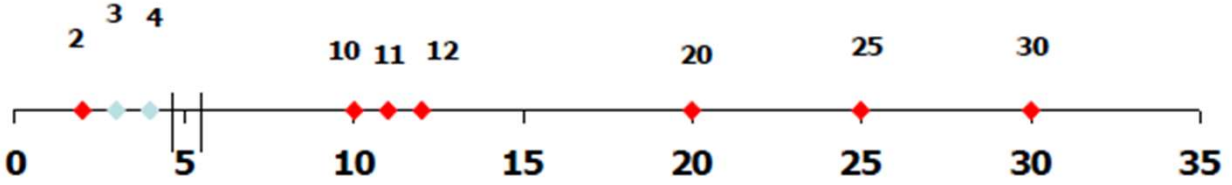
TABLE 10.4 Finding the nearest cluster center for each record (third pass)

Point	Distance from m_1	Distance from m_2	Cluster Membership
a	1.27	3.01	C_1
b	2.15	1.03	C_2
c	3.02	0.25	C_2
d	3.95	1.03	C_2
e	0.35	3.09	C_1
f	2.76	0.75	C_2
g	0.79	3.47	C_1
h	1.06	2.66	C_1

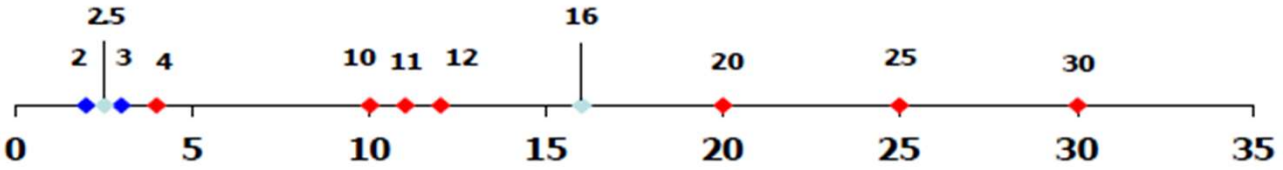
No change in clustering

مثال الگوریتم K-means

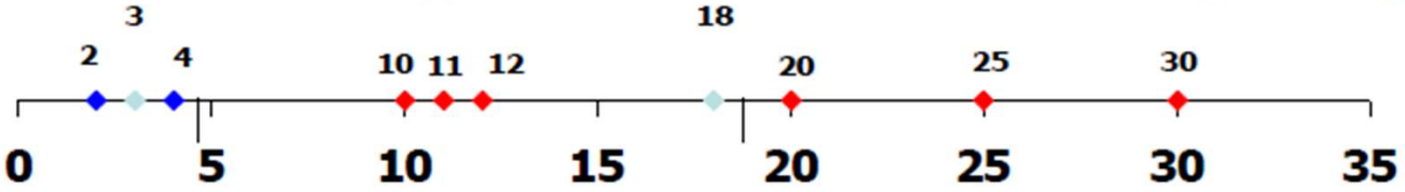
- {2,4,10,12,3,20,30,11,25}, k=2
- $m_1=3, m_2=4$



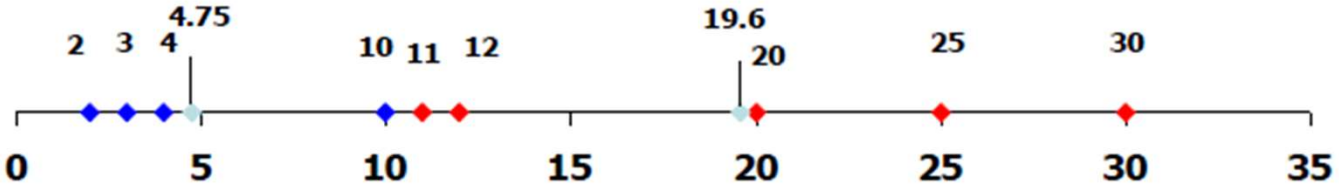
- $K_1=\{2,3\}$, $K_2=\{4,10,12,20,30,11,25\}$, $m_1=2.5, m_2=16$



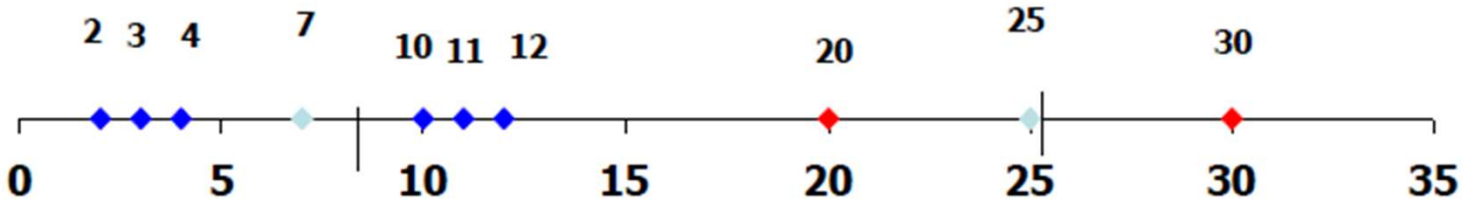
- $K_1=\{2,3,4\}, K_2=\{10,12,20,30,11,25\}$, $m_1=3, m_2=18$



- $K_1=\{2,3,4,10\}, K_2=\{12,20,30,11,25\}$, $m_1=4.75, m_2=19.6$



- $K_1=\{2,3,4,10,11,12\}, K_2=\{20,30,25\}$, $m_1=7, m_2=25$



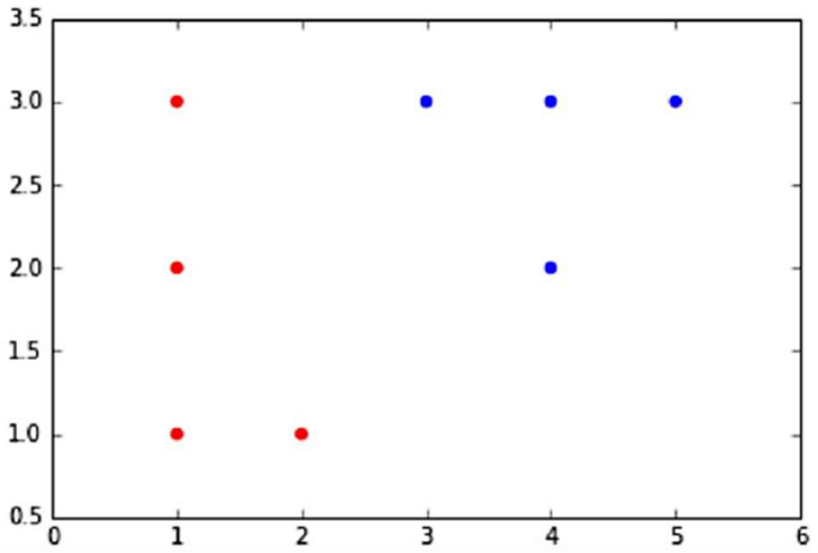
```
from sklearn.cluster import KMeans
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline

X = np.array([[1, 3], [3, 3], [4, 3], [5, 3],
              [1, 2], [4, 2], [1, 1], [2, 1]])

kmeans = KMeans(n_clusters=2).fit(X)
print(kmeans.labels_)

colors = np.array(['r', 'b'])
plt.scatter(X[:,0], X[:,1], color=colors[kmeans.labels_])
```

```
[0 1 1 1 0 1 0 0]
```



K-means

مزایا:

کارآمد و مناسب برای داده های بزرگ: از مرتبه O(tkn)، که در آن n تعداد اشیا، k تعداد خوشه‌ها، و t تعداد تکرارها است. معمولاً $k,t \ll n$

معایب:

اغلب در یک نقطه بهینه محلی متوقف می‌شود.

فقط برای اشیایی در فضای پیوسته n-بعدی قابل استفاده است.

برای داده‌های دسته‌بندی‌شده باید از روش k-modes استفاده شود.

در مقایسه، میانه‌ها را می‌توان برای طیف گسترده‌ای از داده‌ها اعمال کرد.

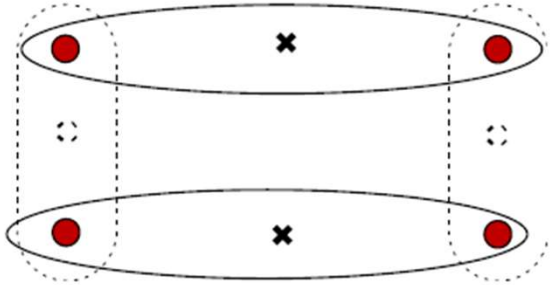
نیاز به مشخص کردن k، تعداد خوشه‌ها، از پیش دارد.

حساس به داده‌های نویزی و نقاط پرت: یک راه‌حل k-medoids است.

حساس به خوشه‌های اولیه: یک راه‌حل k-means++ است.

برای کشف خوشه‌های با اشکال غیر محدب مناسب نیست.

k-means حساس به خوشه‌های اولیه است.



با انتخاب دقیق مراکز اولیه خوشه‌ها، ممکن است بتوانیم نه تنها سرعت همگرایی الگوریتم را افزایش دهیم، بلکه کیفیت خوشه‌بندی نهایی را نیز تضمین کنیم.

الگوریتم k-means++ یک نسخه از k-means است که مراکز اولیه را به شرح زیر انتخاب می‌کند:

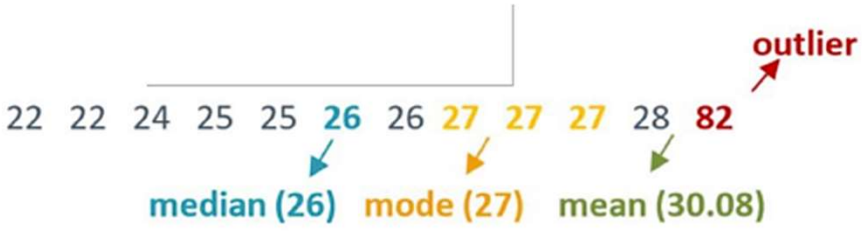
ابتدا، یک مرکز به طور تصادفی از اشیای موجود در مجموعه داده انتخاب می‌شود. مراکز دیگر به طور تصادفی و با احتمال متناسب با dist(p)2 انتخاب می‌شوند، جایی که dist(p) فاصله نقطه p از نزدیک‌ترین مرکزی است که قبلاً انتخاب شده است.

در sklearn.cluster.KMeans، روش اولیه پیش‌فرض، k-means++ است.

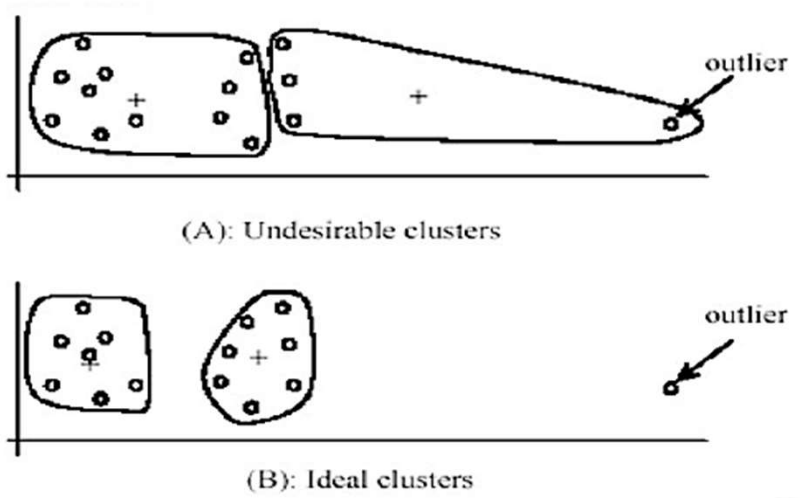
k-medoids

الگوریتم k-means به داده‌های پرت حساس است!

k-medoids: به جای استفاده از میانگین اشیاء در یک خوشه به عنوان نقطه مرجع، می‌توان از میانه‌ها استفاده کرد، که اشیاء با موقعیت مرکزی‌ترین در یک خوشه هستند.



https://www.cese.nsw.gov.au/effective-practices/unit-4-outliers



https://www.slideshare.net/anilyadav5055/15857-cse422-unsupervisedlearning

دانشگاه سمنان
Semnan University
پردیس فرزانگان

خوشه بندی K-Medoids: یافتن اشیاء نماینده (Medoids) در خوشه‌ها

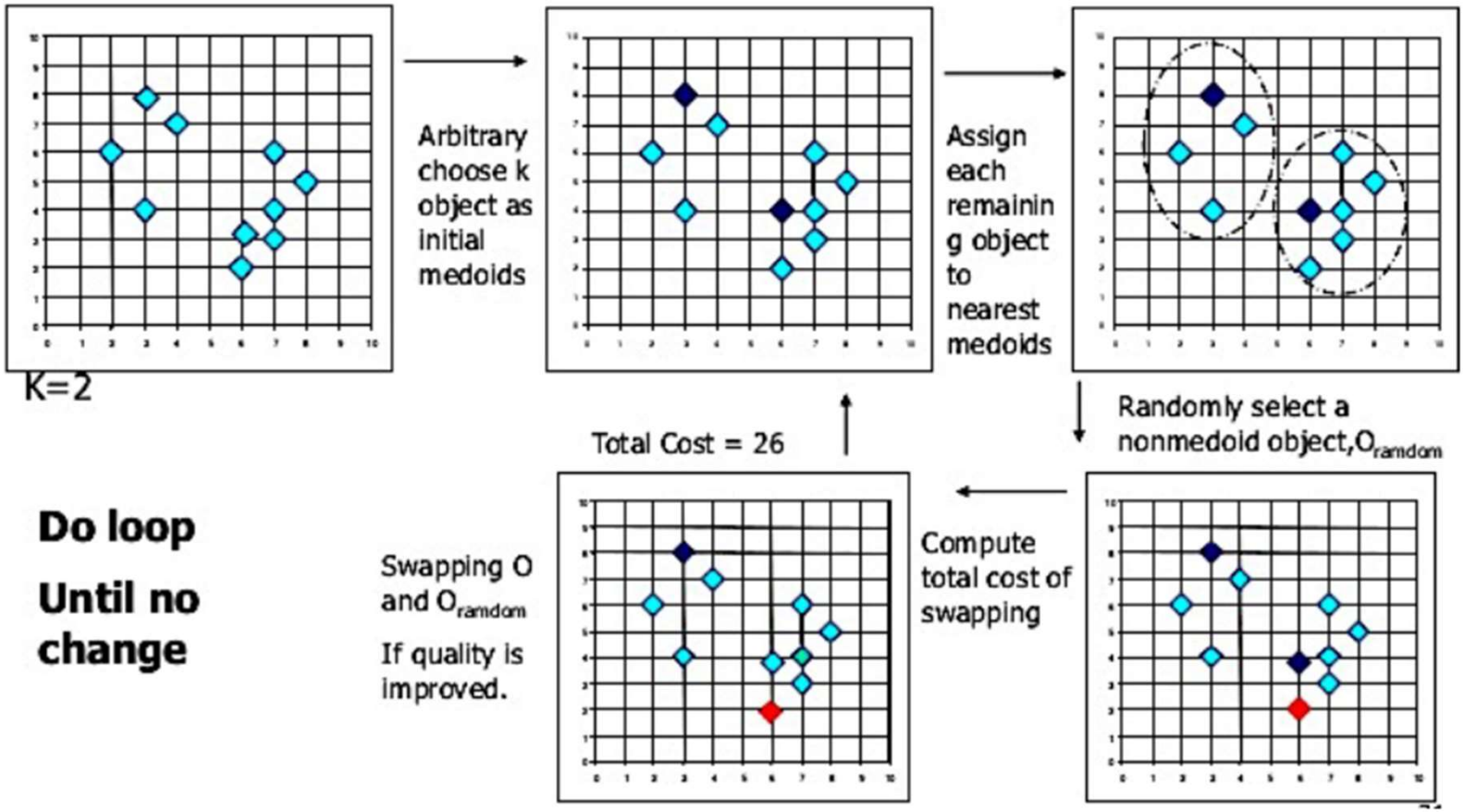
PAM (Partitioning Around Medoids, Kaufmann & Rousseeuw 1987):

از مجموعه اولیه‌ای از میانه‌ها شروع کرده و به صورت تکراری یکی از Medoids را با یکی از non-Medoids جایگزین می‌کند اگر که کل مسافت خوشه‌بندی بهبود یابد.

PAM برای مجموعه‌های داده کوچک مؤثر است، اما برای مجموعه‌های داده بزرگ مقیاس‌پذیری ندارد (به دلیل پیچیدگی محاسباتی)

الگوریتم k-medoids

- شيء به عنوان medoid هاي اوليه به صورت دلخواه اختيار كن.
- تكرار كن تا اينكه هيچ تغييري رخ ندهد.
- هر كدام از اشياء باقيمانده را به خوشه‌اي با نزديكترين medoid تخصيص بده
- بطور تصادفي يك شي غير medoid را انتخاب كن ،.
- هزينة نهايي s را از عوض كردن (medoid آن خوشه) و محاسبه كن
- اگر s<0 آنگاه جای عناصر را عوض كن تا مجموعه K تا medoid جديد شكل بگيرد.


K=2
Arbitrary choose k object as initial medoids
Assign each remainin g object to nearest medoids
Randomly select a nonmedoid object, O_{ramdom}
Compute total cost of swapping
Total Cost = 26
Swapping O and O_{ramdom}
If quality is improved.
Do loop
Until no change

مثال k-medoids

- 1, 2, 6, 7, 8, 10, 15, 17, 20 – break into 3 clusters
  - Cluster = 6 – 1, 2
  - Cluster = 7
  - Cluster = 8 – 10, 15, 17, 20
- Random non-medoid – 15 replace 7 (total cost=-13)
  - Cluster = 6 – 1 (cost 0), 2 (cost 0), 7(1-0=1)
  - Cluster = 8 – 10 (cost 0)
  - New Cluster = 15 – 17 (cost 2-9=-7), 20 (cost 5-12=-7)
- Replace medoid 7 with new medoid (15) and reassign
  - Cluster = 6 – 1, 2, 7
  - Cluster = 8 – 10
  - Cluster = 15 – 17, 20

- Random non-medoid – 1 replaces 6 (total cost= -1)
  - Cluster = 8 – 6 (cost 2-0=2), 7 (cost1-1=0), 10 (cost 0)
  - Cluster = 15 – 17(cost 0), 20(cost 0)
  - New Cluster = 1 – 2 (cost 1-4= -3)
- Replace medoid 6 with new medoid (1) and reassign
  - Cluster = 1 – 2
  - Cluster = 8 – 6, 7, 10
  - Cluster = 15 – 17, 20
- Random non-medoid – 10 replaces 8 (total cost=2) don't replace
  - Cluster = 1– 2(cost 0)
  - Cluster = 15 – 17 (cost 0), 20(cost 0)
  - New Cluster = 10 – 6 (cost 0), 7 (cost 0), 8 (cost 2-0=2)

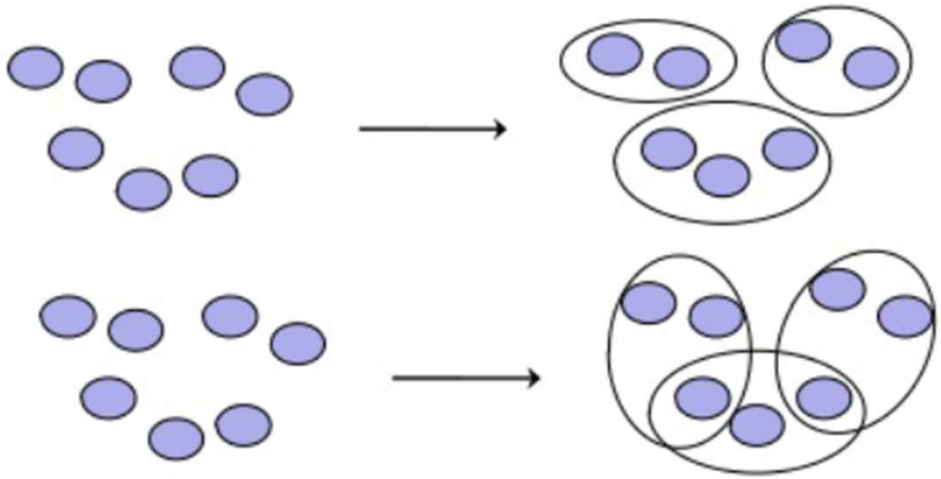
- Random non-medoid – 17 replaces 15 (total cost=0) don't replace
  - Cluster = 1 – 2(cost 0)
  - Cluster = 8 – 6 (cost 0), 7 (cost 0), 10 (cost 0)
  - New Cluster = 17 – 15 (cost 2-0=2), 20(cost 3-5=-2)
- Random non-medoid – 20 replaces 15 (total cost=6) don't replace
  - Cluster = 1 – 2(cost 0)
  - Cluster = 8 – 6 (cost 0), 7 (cost 0), 10 (cost 0)
  - New Cluster = 20 – 15 (cost 5-0=5), 17(cost 3-2=1)
- Other possible changes all have high costs
  - 1 replaces 15, 2 replaces 15, 1 replaces 8, …
- No changes, final clusters
  - Cluster = 1 – 2
  - Cluster = 8 – 6, 7, 10
  - Cluster = 15 – 17, 20

دانشگاه سمنان
Semnan University
پردیس فرزانگان

خوشه بندی فازی یا نرم Soft (fuzzy) clustering

خوشه‌بندی سخت Hard clustering: هر شیء می‌تواند دقیقاً به یک خوشه تعلق داشته باشد.
خوشه‌بندی نرم Soft clustering: اشیاء می‌توانند با یک درجه عضویت کسری به چندین خوشه تعلق داشته باشند.
مانند الگوریتم فازی

Fuzzy c-means (FCM)

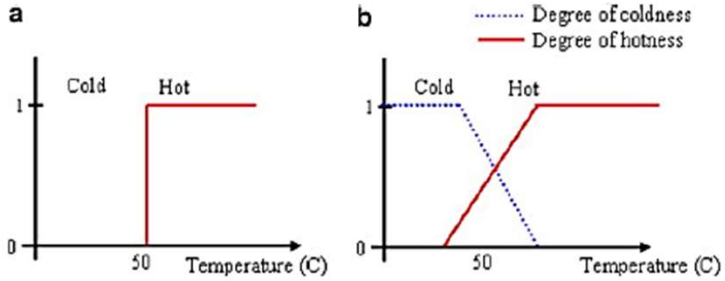


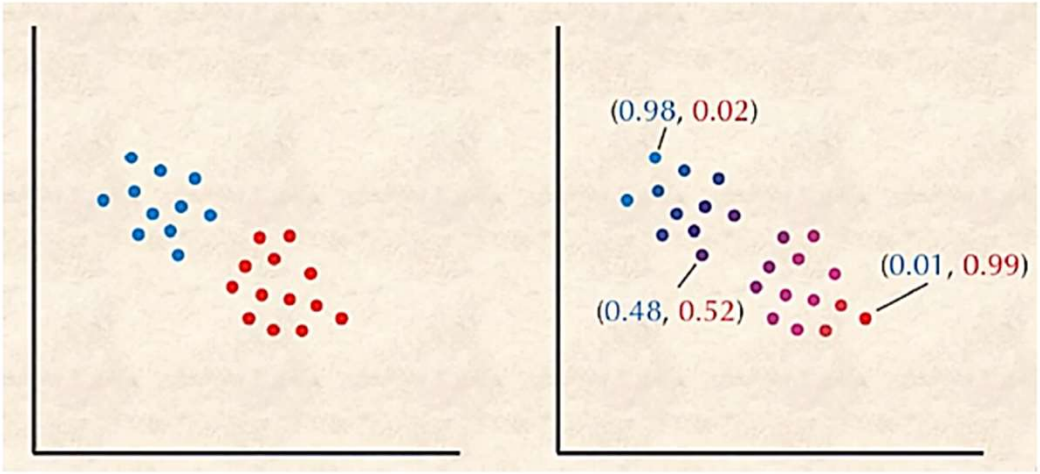
https://www.slideshare.net/kanimozhiu/text-clustering

دانشگاه سمنان
Semnan University
پردیس فرزانگان

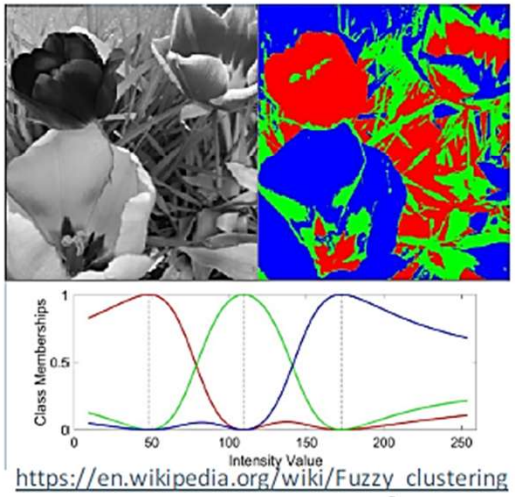
Soft (fuzzy) clustering

An autobiography of Prof. Zadeh:
"My life and work–a retrospective view." *Applied and Computational Mathematics* 10.1 (2011): 4-9.





https://www.youtube.com/watch?v=xtDMHPVDDKk



https://en.wikipedia.org/wiki/Fuzzy_clustering

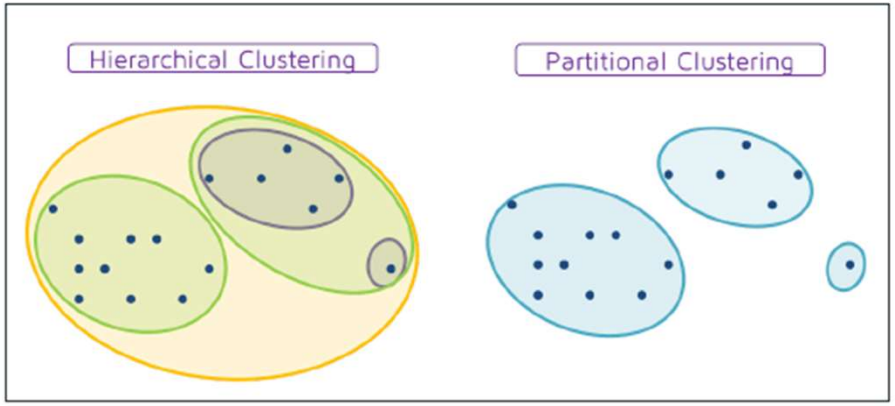
دانشگاه سمنان
Semnan University
پردیس فرزانگان

خوشه بندی سلسله مراتبی

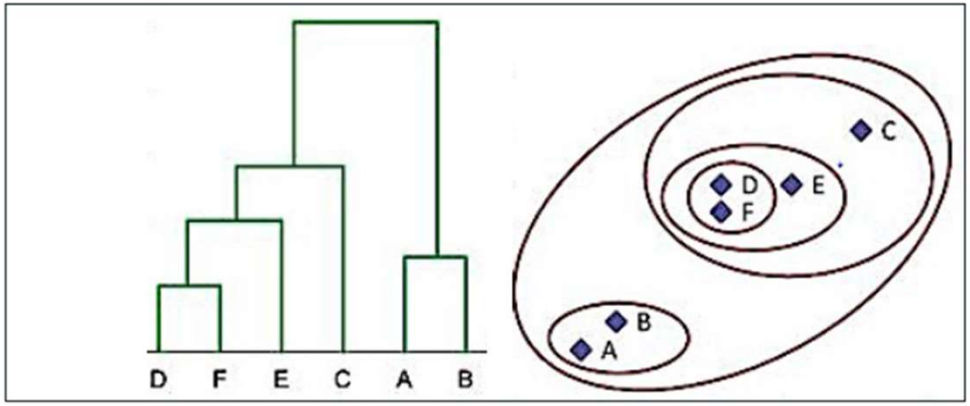
داده‌ها به چندین سطح از تقسیم‌بندی‌های تو در تو (درخت خوشه‌ها) تجزیه شده، که به نمودار درختی تشکیل شده دندروگرام dendrogram گفته می‌شود.

- نباید هر تعداد خاصی از خوشه را فرض کنیم.

هر تعداد مورد نظر از خوشه را می توان با برش dendrogram در سطح مناسب به دست آورد.



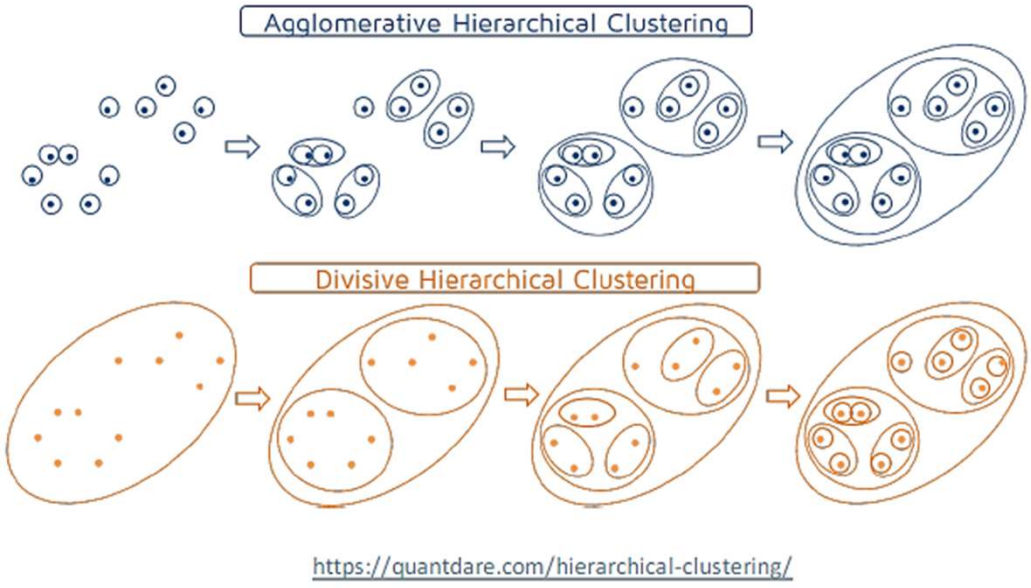
https://quantdare.com/hierarchical-clustering/



دانشگاه سمنان
Semnan University
پردیس فرزانگان

روش های خوشه بندی سلسله مراتبی

تجمیعی Agglomerative (پایین به بالا):
شروع با هر نمونه به‌عنوان یک خوشه منفرد، می باشد. سپس خوشه ها را با ترکیب به خوشه‌های بزرگ و بزرگتر تبدیل می کند تا در نهایت، تمام نمونه‌ها به یک خوشه مشترک تعلق می‌گیرند.



تقسیمی Divisive (بالا به پایین):
شروع با تمام نمونه‌ها که به یک خوشه مشترک تعلق دارند، می باشد. الگوریتم این خوشه را به خوشه های کوچک و کوچک تر تقسیم می کند تا اینکه هر شی در یک خوشه قرار گیرد. در نهایت، هر نمونه یک خوشه جداگانه تشکیل می‌دهد.

این روش به دلیل پیچیدگی محاسباتی بالا کمتر مورد استفاده قرار می گیرد.

فاصله خوشه بندی سلسله مراتبی

الگوریتم‌های سلسله مراتبی سنتی از یک ماتریس شباهت یا فاصله استفاده می‌کنند.
روش محاسبه فاصله مهم است.
معیارهای گوناگونی که در روش سلسله مراتبی برای فاصله بین خوشه ها بکار می رود عبارتند از:

پیوند تک Single Link : کمترین فاصله بین یک عنصر در یک خوشه و یک عنصر در خوشه دیگر.

پیوند کامل Complete Link : بیشترین فاصله بین یک عنصر در یک خوشه و یک عنصر در خوشه دیگر.

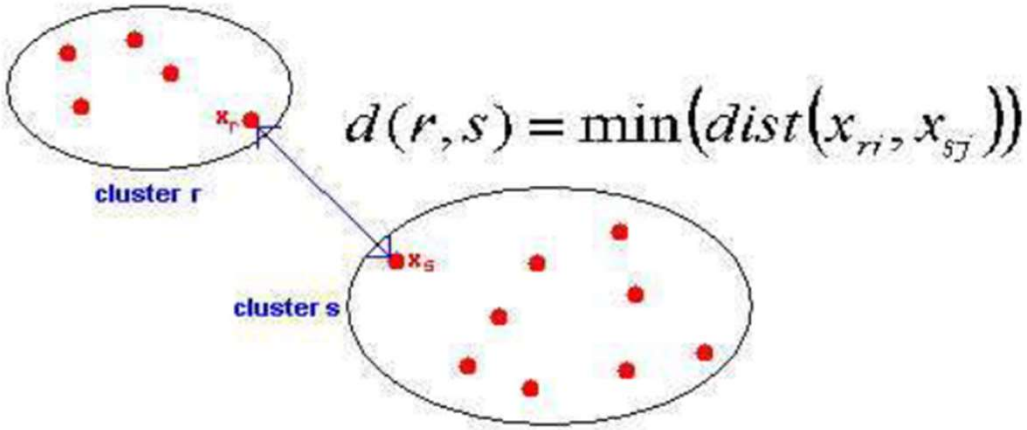
میانگین Average : میانگین فاصله بین یک عنصر در یک خوشه و یک عنصر در خوشه دیگر.

مرکز ثقل Centroid : فاصله بین مراکز ثقل دو خوشه.

مدوید Medoid : فاصله بین مدویدهای دو خوشه

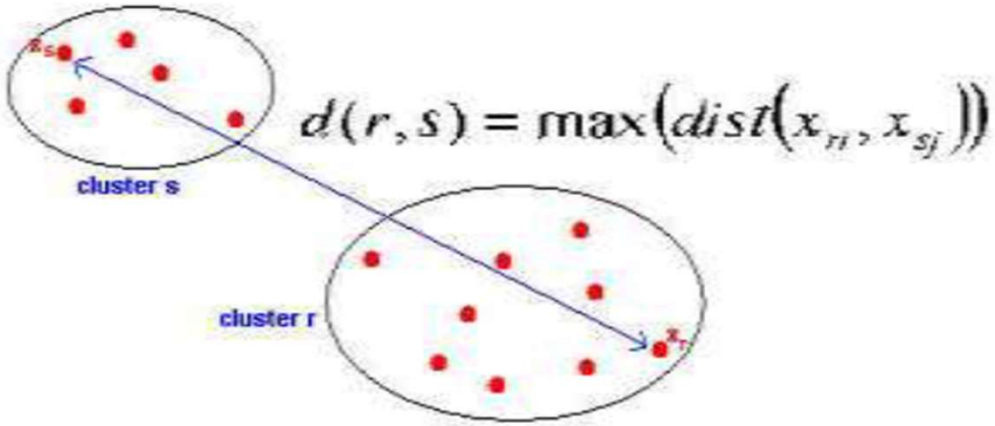
Single linkage

- فاصله بين دسته‌ها بر حسب حداقل فاصله ممكنه بين عناصر آنها محاسبه مي‌شود.
- كليه فاصله بين زوجهاي عناصر دو دسته محاسبه شده و حداقل آنها فاصله بين دو دسته را تعيين مي‌كند.



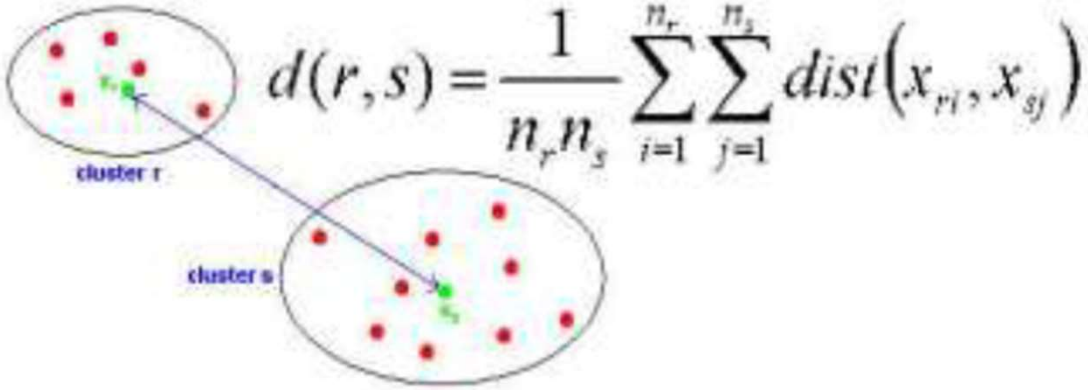
Complete linkage

- فاصله بین دسته‌ها بر حسب دورترین فاصله ممکنه بین عناصر آنها محاسبه مي‌شود.



Average linkage

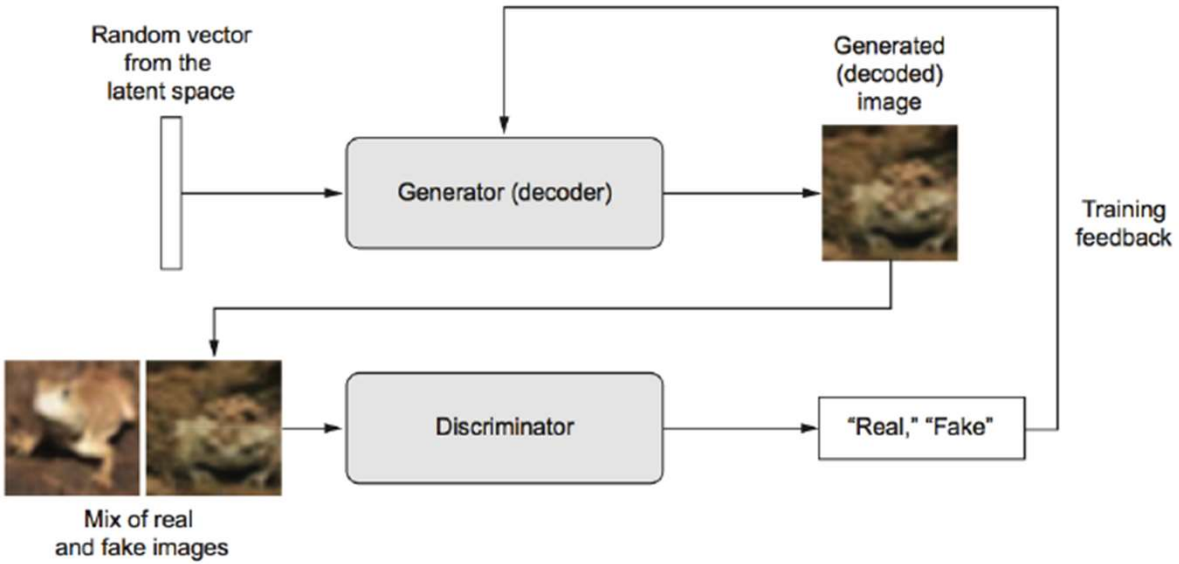
- فاصله بين دو دسته مساوي مقادير متوسط كليه فاصله‌هاي ممكنه بين عناصر دو دسته است.



$$d(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} dist\left(x_{ri}, x_{sj}\right)$$

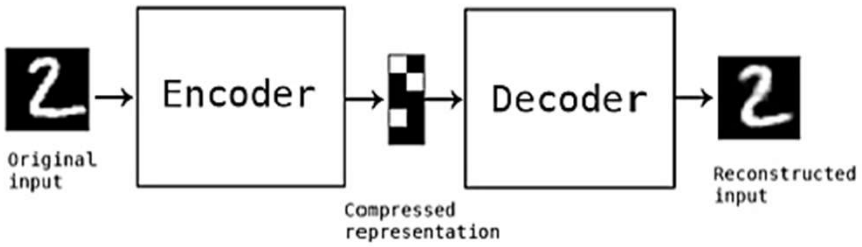
یادگیری بدون نظارت GAN

- Generative adversarial networks (GANs), introduced in 2014 by Goodfellow et al.

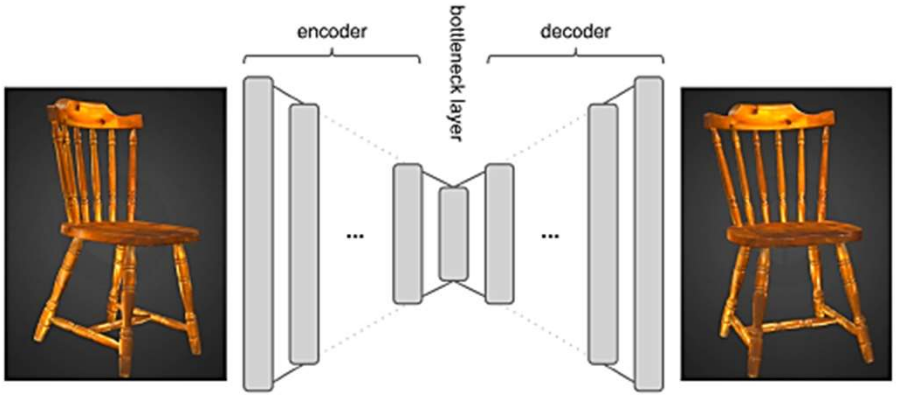


www.miketyka.com

یادگیری بدون نظارت Auto encoder



https://blog.keras.io/building-autoencoders-in-keras.html



http://www.inference.vc/stereovision-autoencoder/

ارزیابی کیفیت خوشه بندی

معیارهای ارزیابی خوشه‌بندی:

۱. ارزیابی خارجی Extrinsic: وجود داده‌های واقعی Ground Truth

در این روش، خوشه‌بندی با استفاده از داده‌های واقعی (Ground Truth) مقایسه می‌شود.

از معیارهای خاصی برای ارزیابی کیفیت خوشه‌بندی استفاده می‌شود.

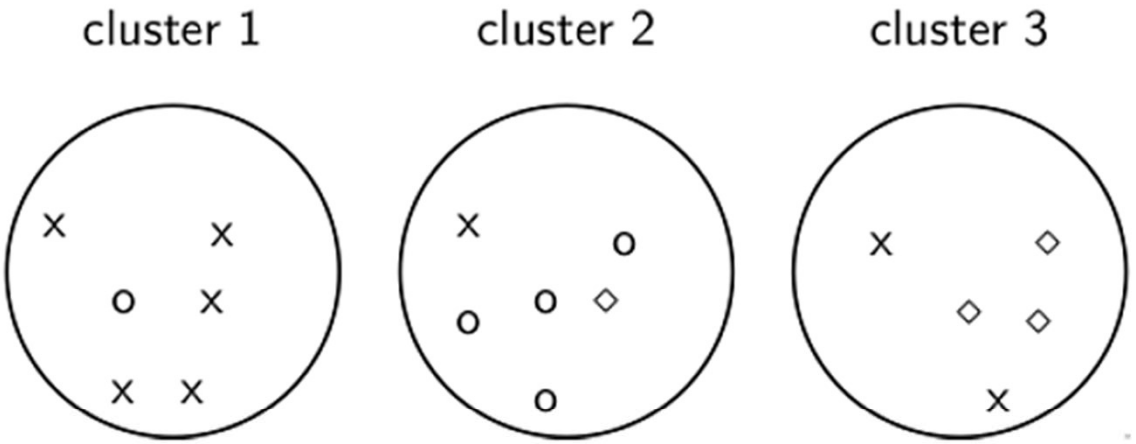
۲. ارزیابی داخلی Intrinsic : عدم وجود داده‌های واقعی Ground Truth

در این روش، بدون استفاده از داده‌های واقعی، کیفیت خوشه‌بندی بر اساس معیارهایی مانند جدایی خوشه‌ها (Separability) و تراکم یا فشردگی خوشه‌ها (Compactness) بررسی می‌شود

ارزیابی خارجی: خلوص Purity

$$\text{purity}(\Omega, C) = \frac{1}{N}\sum_{k}\max_{j}\left|\omega_k \cap c_j\right|$$

- $\Omega = \{\omega_1, \ldots, \omega_K\}$ is the set of clusters and $C = \{c_1, \ldots, c_J\}$ is the set of classes.
- For each cluster ω_k : find class c_j with most members n_{kj} in ω_k.
- Sum all n_{kj} and divide by total number of points.



$5 = \max_j |\omega_1 \cap c_j|$ (class x, cluster 1);

$4 = \max_j |\omega_2 \cap c_j|$ (class o, cluster 2);

$3 = \max_j |\omega_3 \cap c_j|$ (class ◊, cluster 3).

Purity is $(1/17) \times (5 + 4 + 3) \approx 0.71$.

ارزیابی خارجی: Rand index

بر اساس جدول تداخل 2x2 برای تمام جفت‌های نمونه‌ها:

در این روش، ارزیابی خوشه‌بندی بر اساس یک جدول تداخل Contingency Table 2x2 انجام می‌شود. این جدول ارتباط بین جفت‌های نمونه‌ها را با توجه به برچسب‌های واقعی Ground Truth و خوشه‌بندی انجام‌شده بررسی می‌کند.

چهار حالت ممکن برای هر جفت نمونه در جدول:

True Positive (TP): جفت نمونه‌ها در داده‌های واقعی و خوشه‌بندی یکسان هستند.

False Positive (FP): جفت نمونه‌ها در خوشه‌بندی یکسان هستند اما در داده‌های واقعی متفاوت‌اند.

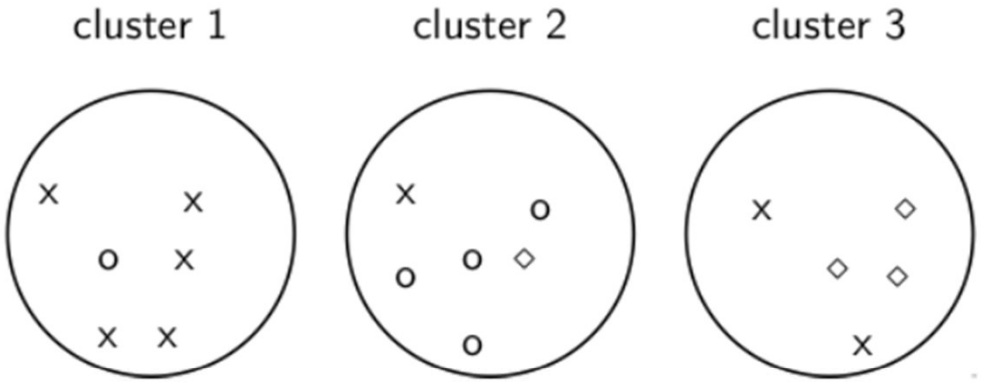
False Negative (FN): جفت نمونه‌ها در داده‌های واقعی یکسان هستند اما در خوشه‌بندی متفاوت‌اند.

True Negative (TN): جفت نمونه‌ها در داده‌های واقعی و خوشه‌بندی متفاوت‌اند.

	same cluster	different clusters
same class	true positives (TP)	false negatives (FN)
different classes	false positives (FP)	true negatives (TN)

- TP+FN+FP+TN is the total number of pairs: $\binom{n}{2}$ for n samples.

Definition: $\mathrm{RI} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$



	same cluster	different clusters
same class	TP = 20	FN = 24
different classes	FP = 20	TN = 72

RI = (20 + 72)/(20 + 20 + 24 + 72) ≈ 0.68

ارزیابی داخلی: مجموع مربعات خطا Sum of Squared Error

SSE مجموع مربعات فاصله‌ها بین نقاط داده و مرکز خوشه Centroid مربوطه است.
فرمول کلی SSE به صورت زیر است:

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

که در آن:

k: تعداد خوشه‌ها

Ci: مجموعه نقاط در خوشه i

mi: مرکز خوشه i

x یک نمونه داده در خوشه Ci

SSE کمتر نشان‌دهنده خوشه‌بندی بهتر است، زیرا نقاط داده به مرکز خوشه نزدیک‌تر هستند (تراکم بیشتر).

با افزایش تعداد خوشه‌ها، SSE معمولاً کاهش می‌یابد، زیرا خوشه‌های کوچک‌تر نقاط را دقیق‌تر توصیف می‌کنند.