



فصل هفتم
شناسایی الگو
خوشه‌بندی، مفاهیم پایه

CLUSTERING, BASIC CONCEPTS

محمدجواد فدائی اسلام

خوشه‌بندی (مقدمه)

CLUSTERING (INTRODUCTION)

- هدف در فصل‌های گذشته کلاس‌بندی با ناظر بود.
- در مواردی که ناظر وجود ندارد، برچسبی برای نمونه موجود نیست.
- هدف در اینگونه موارد گروه‌بندی نمونه‌هاست تا شباهت و تفاوت بین آنها استخراج شود و نتایج مفیدی به دست دهد.
- Clustering may be found under different names in different contexts, such as **unsupervised learning** and **learning without a teacher**, **numerical taxonomy**, **typology** and **partition**.

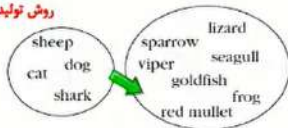
تعریف خوشه‌بندی با مثال

- sheep, dog, cat (mammals),
 - sparrow, seagull (birds),
 - viper, lizard (reptiles),
 - goldfish, red mullet, blue shark (fish),
 - frog (amphibians).
- گوسفند، سگ، گربه (پستانداران)،
 - گنجشک، مرغ دریایی (پرندگان)،
 - افعی، مارمولک (خزندگان)،
 - ماهی قرمز، شاه‌ماهی قرمز، کوسه‌آبی (ماهی)،
 - قورباغه (دوزیستان).



با توجه به معیارهای مختلف، نتایج مختلفی ممکن است از خوشه‌بندی حاصل شود

روش تولید بچه



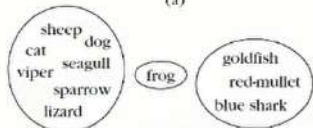
(a)

روش تنفس



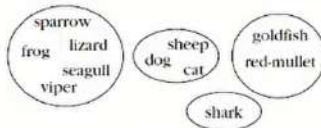
(b)

روش تنفس و به هم زدن



روش زیست

(c)



(d)

FIGURE 11.1

Resulting clusters if the clustering criterion is (a) the way the animals bear their progeny, (b) the existence of lungs, (c) the environment where the animals live, and (d) the way these animals bear their progeny and the existence of lungs.

گام‌های خوشه‌بندی

1. انتخاب ویژگی (Feature selection)
2. معیار مجاورت (Proximity measure)
3. روش خوشه‌بندی (Clustering criterion)
4. اعتبارسنجی نتایج (Validation of the results)
5. تفسیر نتایج (Interpretation of the results)



کاربردهای خوشه‌بندی

- کاهش ویژگی (Data reduction)
- تولید فرضیه (Hypothesis generation)
- آزمایش فرضیه (Hypothesis testing)




انواع ویژگی

- 1. اسمی (Nominal)
- 1. ترتیبی (Ordinal)
- 2. بازه‌ای (Interval)
- 2. نسبی (Ratio)

۱- انواع ویژگی (اسمی و ترتیبی)

○ اسمی

- نام‌هایی که برای ذخیره‌سازی به آنها عدد داده شده است.
- مثال: جنیست، مرد=۱، زن=۲  شماره دانشجویی، شناسه ملی
- هر مقابسه کمی بین آنها بی‌معنی است. تنها عملگر برابری و نابرابری معنی‌دار است.

○ ترتیبی

- ویژگی‌ای که مقدارهای آن را می‌توان مرتب نمود.
- مثال: توصیف کیفی نمره دانش‌آموز کلاس اول: خیلی خوب، خوب، متوسط، نیاز به تلاش بیشتر
- ترتیب در این ویژگی‌ها معنار دار است اما تفاوت کمی بین آنها معنادار نیست.

۱- انواع ویژگی (بازه‌ای و نسبی)



○ بازه‌ای

- تفریق بین دو مقدار معنادار است اما نسبت بین دو مقدار بی معنی است.
- مثال: اندازه‌گیری دما به سانتی‌گراد. دمای سمنان ۳۰ درجه و دمای فیروزکوه ۱۰ درجه است. سمنان ۲۰ درجه گرم‌تر است اما از فیروزکوه ۳ برابر گرم‌تر نیست.

○ نسبی

- نسبت بین دو ویژگی معنا دارد.
- مثال: وزن، شخص ۱۰۰ کیلوپی دو برابر شخص ۵۰ کیلوپی وزن دارد.
- مثال: دما به کلوین

خوشه‌بندی، تعریف

CLUSTERING, DEFINITION

○ اگر X یک مجموعه داده باشد، خوشه‌بندی m تایی، مجموعه X را به m مجموعه C_1, \dots, C_m

مجزا با شرایط زیر افراز می‌کند:

- $C_i \neq \emptyset, i = 1, \dots, m$
- $\bigcup_{i=1}^m C_i = X$
- $C_i \cap C_j = \emptyset, i \neq j, i, j = 1, \dots, m$
- علاوه بر این داده‌های درون خوشه C_i به هم بیشتر شبیه هستند و به داده‌های دیگر خوشه‌ها کمتر شبیه هستند. البته مفهوم شباهت باید با توجه به نوع خوشه‌بندی مشخص شود.



انواع مختلف خوشه‌بندی معیارهای متفاوتی برای تعیین مشابهت لازم دارند

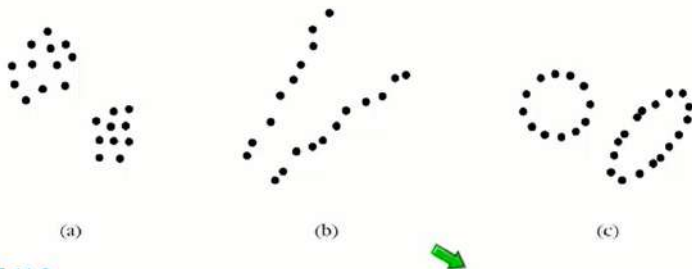


FIGURE 11.3

(a) Compact clusters. (b) Elongated clusters. (c) Spherical and ellipsoidal clusters.

اندازه گیری مجاورت بین دو نقطه (بردار ویژگی حقیقی)-عدم شباهت

The *weighted* l_p metric DMs, that is,

$$d_p(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^l w_i |x_i - y_i|^p \right)^{1/p} \quad w_i \geq 0$$

If $w_i = 1, i = 1, \dots, l$, we obtain the *unweighted* l_p metric DMs. A well-known representative of the latter category of measures is the *Euclidean distance*, by setting $p = 2$.

$w_i = 1 \rightarrow d_p(x, y) = \text{Minkovski distance}$



عدم شباهت بین دو نقطه (بردار ویژگی حقیقی) - ادامه

Special l_p metric DMs that are also encountered in practice are the (weighted) l_1 or *Manhattan norm*,

$$d_1(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^l w_i |x_i - y_i|$$

and the (weighted) l_∞ norm,

$$d_\infty(\mathbf{x}, \mathbf{y}) = \max_{1 \leq i \leq l} w_i |x_i - y_i|$$



The l_1 and l_∞ norms may be viewed as overestimation and underestimation of the l_2 norm, respectively.

$$d_\infty(\mathbf{x}, \mathbf{y}) \leq d_2(\mathbf{x}, \mathbf{y}) \leq d_1(\mathbf{x}, \mathbf{y})$$

Example 11.4

Consider the three-dimensional vectors $\mathbf{x} = [0, 1, 2]^T$, $\mathbf{y} = [4, 3, 2]^T$. Then, assuming that all w_i 's are equal to 1, $d_1(\mathbf{x}, \mathbf{y}) = 6$, $d_2(\mathbf{x}, \mathbf{y}) = 2\sqrt{5}$, and $d_\infty(\mathbf{x}, \mathbf{y}) = 4$. Notice that $d_\infty(\mathbf{x}, \mathbf{y}) < d_2(\mathbf{x}, \mathbf{y}) < d_1(\mathbf{x}, \mathbf{y})$.

اندازه‌گیری مشابهت میان دو نقطه - ضرب داخلی

$$s_{\text{inner}}(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y} = \sum_{i=1}^l x_i y_i.$$

در اغلب موارد ضرب داخلی زمانی استفاده می‌شود که بردار \mathbf{x} و \mathbf{y} نرمال باشند.



اندازه‌گیری مشابهت میان دو نقطه - مشابهت کسینوسی

$$s_{\text{cosine}}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

where $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^l x_i^2}$ and $\|\mathbf{y}\| = \sqrt{\sum_{i=1}^l y_i^2}$ are the lengths of the vectors \mathbf{x} and \mathbf{y} , respectively. This measure is invariant to rotations but not to linear transformations.

اندازه‌گیری مشابهت میان دو نقطه - مشابهت کسینوسی

$$s_{\text{cosine}}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

where $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^l x_i^2}$ and $\|\mathbf{y}\| = \sqrt{\sum_{i=1}^l y_i^2}$ are the lengths of the vectors \mathbf{x} and \mathbf{y} , respectively. This measure is invariant to rotations but not to linear transformations.

اندازه‌گیری مجاورت بین یک نقطه و یک مجموعه دو رهیافت

- در رهیافت اول تمام نقاط در یافتن میزان مجاورت به طور مستقیم شرکت می‌کنند.
- در رهیافت دوم از مجموعه نقاط یک نماینده ایجاد می‌شود و فاصله نقطه تا آن محاسبه می‌شود. این نماینده می‌تواند نقطه، خط یا ... باشد



اندازه گیری مجاورت بین یک نقطه و یک مجموعه رهیافت اول

- The *max proximity function*:

$$\wp_{\max}^{ps}(\mathbf{x}, C) = \max_{y \in C} \wp(\mathbf{x}, y)$$

- The *min proximity function*:

$$\wp_{\min}^{ps}(\mathbf{x}, C) = \min_{y \in C} \wp(\mathbf{x}, y)$$

- The *average proximity function*:

$$\wp_{\text{avg}}^{ps}(\mathbf{x}, C) = \frac{1}{n_C} \sum_{y \in C} \wp(\mathbf{x}, y)$$

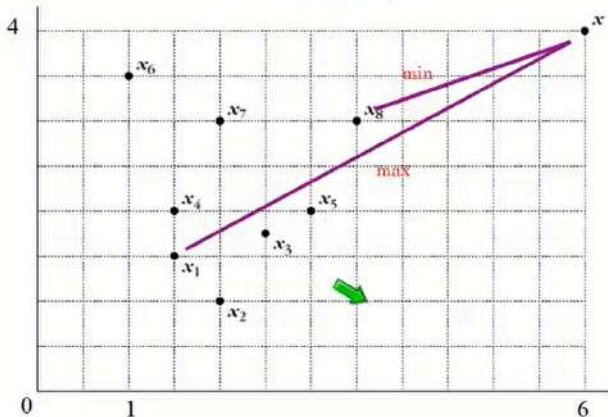
where n_C is the cardinality of C .

In these definitions, $\wp(\mathbf{x}, y)$ may be any proximity measure between two points.

اندازه‌گیری مجاورت بین یک نقطه و یک مجموعه دو رهیافت

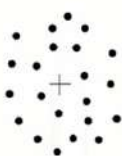
- در رهیافت اول تمام نقاط در یافتن  مجاورت به طور مستقیم شرکت می‌کنند.
- در رهیافت دوم از مجموعه نقاط یک نماینده ایجاد می‌شود و فاصله نقطه تا آن محاسبه می‌شود. این نماینده می‌تواند نقطه، خط یا ... باشد

اندازه گیری مجاورت بین یک نقطه و یک مجموعه رهیافت اول - مثال



اندازه‌گیری مجاورت بین یک نقطه و یک مجموعه
رهیافت دوم

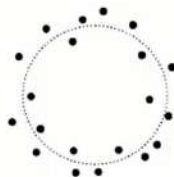
REPRESENTATIVES



(a)



(b)



(c)

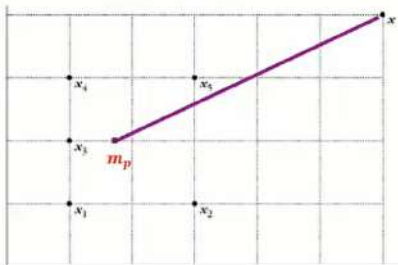
FIGURE 11.7

(a) Compact cluster. (b) Hyperplanar (linear) cluster. (c) Hyperspherical cluster.

اندازه گیری مجاورت بین یک نقطه و یک مجموعه
 رهیافت دوم: میانگین مجموعه نقاط به عنوان نماینده

The mean vector (or mean point)

$$m_p = \frac{1}{n_C} \sum_{y \in C} y$$

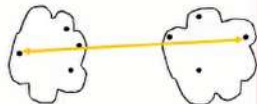


اندازه گیری مجاورت بین دو مجموعه



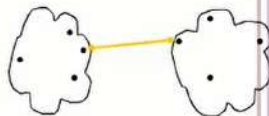
- The *max proximity function*:

$$\wp_{\max}^{ss}(D_I, D_J) = \max_{x \in D_I, y \in D_J} \wp(x, y)$$

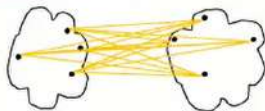


- The *min proximity function*:

$$\wp_{\min}^{ss}(D_I, D_J) = \min_{x \in D_I, y \in D_J} \wp(x, y)$$



اندازه گیری مجاورت بین دو مجموعه (ادامه)



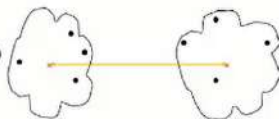
- The *average proximity function*:

$$\wp_{\text{avg}}^{ss}(D_i, D_j) = \frac{1}{n_{D_i} n_{D_j}} \sum_{x \in D_i} \sum_{y \in D_j} \wp(x, y)$$



- The *mean proximity function*:

$$\wp_{\text{mean}}^{ss}(D_i, D_j) = \wp(m_{D_i}, m_{D_j})$$





فصل هشتم
شناسایی الگو
روش‌های خوشه‌بندی

CLUSTERING ALGORITHMS

محمدجواد فدائی اسلام

انواع الگوریتم‌های خوشه‌بندی

- الگوریتم‌های ترتیبی (Sequential algorithms)
- الگوریتم‌های خوشه‌بندی سلسله‌مراتبی (Hierarchical clustering algorithms)
 - تجمیعی (Agglomerative algorithms)
 - تقسیمی (Divisive algorithms)
- خوشه‌بندی متنی بر تابع هزینه (Clustering based on cost function)

الگوریتم‌های ترتیبی

- این الگوریتم‌ها یک نوع خوشه‌بندی ایجاد می‌کنند.
- این الگوریتم‌ها سراسر و سریع هستند.
- در اغلب آنها هر داده تنها یک یا تعداد محدودی بار (کمتر از ۵ یا ۶) در فرآیند خوشه‌بندی شرکت داده می‌شود.
- نتیجه نهایی به ترتیب داده‌ها که در فرآیند خوشه‌بندی شرکت می‌کنند وابسته است.

الگوریتم‌های ترتیبی خوشه‌بندی

BASIC SEQUENTIAL ALGORITHMIC SCHEME (BSAS)

ایده اصلی

وقتی یک داده جدید می‌آید، فاصله آن با خوشه‌های موجود بررسی می‌شود. اگر فاصله آن از خوشه‌های موجود بیشتر از حد آستانه باشد یک خوشه جدید تشکیل می‌شود. در غیر این صورت به نزدیک‌ترین خوشه ملحق می‌شود.

BASIC SEQUENTIAL ALGORITHMIC SCHEME (BSAS)

- $m = 1$
- $C_m = \{\mathbf{x}_1\}$
- For $i = 2$ to N
 - Find C_k : $d(\mathbf{x}_i, C_k) = \min_{1 \leq j \leq m} d(\mathbf{x}_i, C_j)$.
 - If $(d(\mathbf{x}_i, C_k) > \Theta)$ AND $(m < q)$ then
 - $m = m + 1$
 - $C_m = \{\mathbf{x}_i\}$
 - Else
 - $C_k = C_k \cup \{\mathbf{x}_i\}$
 - Where necessary, update representatives²
 - End (if)
- End (For)

الگوریتم‌های ترتیبی خوشه‌بندی-BSAS

- در این الگوریتم یک داده تنها یک‌بار در فرآیند خوشه‌بندی شرکت می‌کند.
- تعداد خوشه‌ها در ابتدا مشخص نیست.
- الگوریتم دو پارامتر دارد که باید توسط کاربر تعیین شود
 - الف) آستانه فاصله
 - ب) حداکثر تعداد خوشه یا q .
- می‌توان حداکثر تعداد خوشه تعیین ننمود و الگوریتم به صورت خودکار و تنها با مقدار آستانه کار کند.
- ترتیب داده‌ها نقش کلیدی دارد.



الگوریتم‌های ترتیبی خوشه‌بندی-MODIFIED BSAS

- در BAS داده X به یک خوشه موجود اضافه می‌شود و یا یک خوشه جدید تشکیل می‌دهد.
- خوشه‌بندی داده X با یکبار دیدن آن تثبیت می‌شود.
- در MBAS این مشکل برطرف شده است و هر داده دو بار در فرآیند خوشه‌بندی شرکت می‌کند.

ایده اصلی

داده‌ها در مرحله اول به فرایند خوشه‌بندی وارد می‌شوند. برخی از آنها خوشه‌بندی می‌شوند. در مرحله دوم داده‌هایی که خوشه‌بندی نشده‌اند. در خوشه مناسب قرار می‌گیرند. در مرحله دوم خوشه جدید ایجاد نمی‌شود.



Cluster Determination

A MODIFICATION OF BSAS (MBSAS)

- $m = 1$
- $C_m = \{\mathbf{x}_1\}$
 - For $i = 2$ to N
 - Find C_k : $d(\mathbf{x}_i, C_k) = \min_{1 \leq j \leq m} d(\mathbf{x}_i, C_j)$.
 - If $(d(\mathbf{x}_i, C_k) > \Theta)$ AND $(m < q)$ then
 - $m = m + 1$
 - $C_m = \{\mathbf{x}_i\}$
 - End {if}
- End {For}



Pattern Classification

- For $i = 1$ to N
- If \mathbf{x}_i has not been assigned to a cluster, then
 - Find C_k : $d(\mathbf{x}_i, C_k) = \min_{1 \leq j \leq m} d(\mathbf{x}_i, C_j)$
 - $C_k = C_k \cup \{\mathbf{x}_i\}$
 - Where necessary, update representatives
- End {if}
- End {For}



THE TWO-THRESHOLD SEQUENTIAL ALGORITHMIC SCHEME (TTSAS)

The Two-Threshold Sequential Algorithmic Scheme (TTSAS)

$m = 0$

$clas(\mathbf{x}) = 0, \quad \forall \mathbf{x} \in X$

$prev_change = 0$

$cur_change = 0$

$exists_change = 0$



TTSAS ...

While (there exists at least one feature vector \mathbf{x} with $\text{clas}(\mathbf{x}) = 0$) do

- For $i = 1$ to N
 - if $\text{clas}(\mathbf{x}_i) = 0$ AND it is the first in the new while loop AND $\text{exists_change} = 0$ then
 - $m = m + 1$
 - $C_m = \{\mathbf{x}_i\}$
 - $\text{clas}(\mathbf{x}_i) = 1$
 - $\text{cur_change} = \text{cur_change} + 1$
 - Else if $\text{clas}(\mathbf{x}_i) = 0$ then
 - Find $d(\mathbf{x}_i, C_k) = \min_{1 \leq j \leq m} d(\mathbf{x}_i, C_j)$
 - if $d(\mathbf{x}_i, C_k) < \Theta_1$ then
 - $C_k = C_k \cup \{\mathbf{x}_i\}$
 - $\text{clas}(\mathbf{x}_i) = 1$
 - $\text{cur_change} = \text{cur_change} + 1$
 - else if $d(\mathbf{x}_i, C_k) > \Theta_2$ then
 - $m = m + 1$
 - $C_m = \{\mathbf{x}_i\}$
 - $\text{clas}(\mathbf{x}_i) = 1$
 - $\text{cur_change} = \text{cur_change} + 1$
 - End {If}
 - Else if $\text{clas}(\mathbf{x}_i) = 1$ then
 - $\text{cur_change} = \text{cur_change} + 1$
 - End {If}
 - End {For}
 - $\text{exists_change} = |\text{cur_change} - \text{prev_change}|$
 - $\text{prev_change} = \text{cur_change}$
 - $\text{cur_change} = 0$
- End {While}

الگوریتم خوشه‌بندی سلسله مراتبی تجمیعی

در ابتدا هر نمونه یک خوشه است. در هر مرحله دو خوشه (یا نمونه) به هم می‌پیوندند و خوشه بزرگتری تشکیل می‌شود.

■ Initialization:

- Choose $\mathfrak{R}_0 = \{C_i = \{\mathbf{x}_i\}, i = 1, \dots, N\}$ as the initial clustering.
- $t = 0$.

■ Repeat:


- $t = t + 1$
- Among all possible pairs of clusters (C_r, C_s) in \mathfrak{R}_{t-1} find the one, say (C_i, C_j) , such that

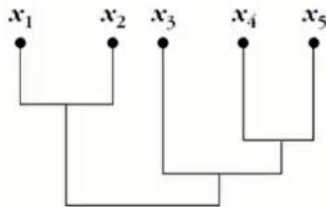
$$g(C_i, C_j) = \begin{cases} \min_{r,s} g(C_r, C_s), & \text{if } g \text{ is a dissimilarity function} \\ \max_{r,s} g(C_r, C_s), & \text{if } g \text{ is a similarity function} \end{cases} \quad (13.1)$$

- Define $C_q = C_i \cup C_j$ and produce the new clustering $\mathfrak{R}_t = (\mathfrak{R}_{t-1} - \{C_i, C_j\}) \cup \{C_q\}$.
- Until all vectors lie in a single cluster.

Example 13.1

Let $X = \{\mathbf{x}_i, i = 1, \dots, 5\}$, with $\mathbf{x}_1 = [1, 1]^T$, $\mathbf{x}_2 = [2, 1]^T$, $\mathbf{x}_3 = [5, 4]^T$, $\mathbf{x}_4 = [6, 5]^T$, and $\mathbf{x}_5 = [6.5, 6]^T$. The pattern matrix of X is

$$D(X) = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 5 & 4 \\ 6 & 5 \\ 6.5 & 6 \end{bmatrix}$$




corresponding dissimilarity matrix, Euclidean distance, is

dendrogram

$$P(X) = \begin{bmatrix} 0 & 1 & 5 & 6.4 & 7.4 \\ 1 & 0 & 4.2 & 5.7 & 6.7 \\ 5 & 4.2 & 0 & 1.4 & 2.5 \\ 6.4 & 5.7 & 1.4 & 0 & 1.1 \\ 7.4 & 6.7 & 2.5 & 1.1 & 0 \end{bmatrix}$$

نقاط قوت الگوریتم تجمیعی

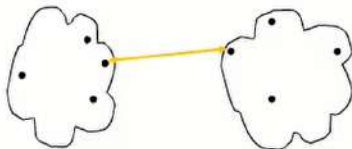
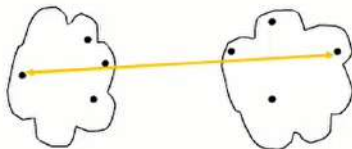
- نیازی به تعیین تعداد خوشه در ابتدای کار نیست. با برش دندروگرام از نقطه دلخواه می‌توان به تعداد دلخواه خوشه داشت.



تعیین فاصله دو خوشه

روش بیشینه - MAX

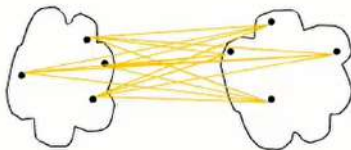
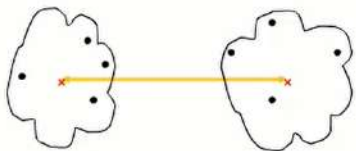
روش کمینه - MIN




تعیین فاصله دو خوشه - ادامه

روش فاصله بین میانگین دو خوشه
Distance Between Centroids

روش میانگین فاصله هر دو نمونه از دو خوشه
Group Average



روش کمینه یا تک اتصال MIN OR SINGLE LINK

مجاورت دو خوشه بر اساس دو نزدیکترین نقطه از هر کدام شکل می گیرد.
بر این اساس فاصله بر اساس یک اتصال محاسبه می شود.
نقطه قوت: می تواند توزیع های غیر بیضوی را به خوبی خوشه بندی کند.
نقطه ضعف: به نویز و نقطه دور افتاده حساس است. 

نقطه قوت روش کمینه

می تواند توزیع های غیر بیضوی را به خوبی خوشه بندی کند



Original Points

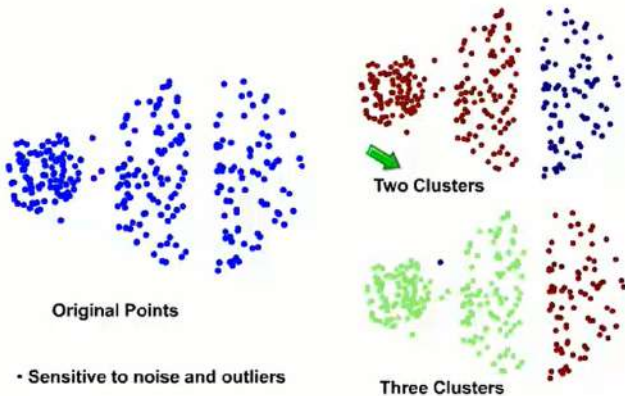


Six Clusters



- Can handle non-elliptical shapes

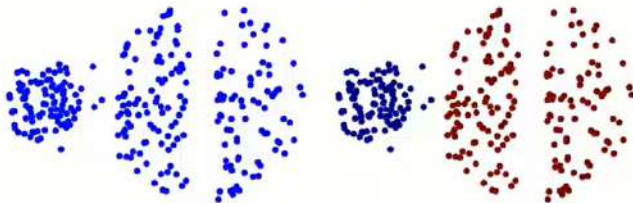
نقطه ضعف روش کمینه: حساسیت به نویز و نقطه دور افتاده



• Sensitive to noise and outliers

روش بیشینه یا اتصال کامل
MAX OR COMPLETE LINKAGE

مجاورت دو خوشه بر اساس دو دورترین نقطه از هر کدام شکل می گیرد.
نقطه قوت روش بیشینه: حساسیت کمتر به نویز و نقطه دور افتاده



Original Points

Two Clusters

روش میانگین فاصله هر دو نمونه از دو خوشه

GROUP AVERAGE

- محاسبه همسایگی یا مجاورت میان دو خوشه با استفاده از میانگین فاصله هر دو نمونه از دو خوشه.
- اگر میانگین گرفته نشود معیار مجاورت به سمت خوشه‌های بزرگتر گرایش پیدا می‌کند.
- این روش در برابر نویز مقاومت بیشتری دارد.

$$\text{proximity}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\sum_{\substack{p_i \in \text{Cluster}_i \\ p_j \in \text{Cluster}_j}} \text{proximity}(p_i, p_j)}{|\text{Cluster}_i| \times |\text{Cluster}_j|}$$

الگوریتم خوشه‌بندی سلسله مراتبی تقسیمی

DIVISIVE ALGORITHMS

- این روش در مقابل روش تجمیعی قرار دارد.
- در هر مرحله یکی از خوشه‌ها به دو خوشه تقسیم می‌شود.

خوشه‌بندی به عنوان مساله بهینه‌سازی و بر اساس تابع هزینه CLUSTERING BASED ON COST FUNCTION OPTIMIZATION

- در این مسایل یک تابع هزینه به نام J تعریف می‌شود.
- معمولا تعداد خوشه‌ها ثابت نگه‌داشته می‌شود.
- از مفاهیم حساب دیفرانسیل استفاده می‌شود و در حالی که سعی دارند تابع J را کمینه نمایند، خوشه‌بندی را تولید می‌نمایند.
- معمولا در نقطه بهینه محلی تابع J متوقف می‌شوند.
- معمولا روش‌های تکرار شونده‌ای هستند.


انواع روش‌های خوشه‌بندی مبتنی بر تابع هزینه



چهار روش عمده وجود دارد:

- ۱- روش خوشه‌بندی سخت (Hard Clustering)
- ۲- روش خوشه‌بندی فازی (Fuzzy Clustering)
- ۳- روش خوشه‌بندی امکانی (possibilistic clustering)
- ۴- روش تجزیه مخلوط (The mixture decomposition clustering)

The Isodata or k-Means or c-Means Algorithm

- Choose arbitrary initial estimates $\theta_j(0)$ for the θ_j 's, $j = 1, \dots, m$.
- Repeat 
 - For $i = 1$ to N
 - Determine the closest representative, say θ_j , for \mathbf{x}_i .
 - Set $b(i) = j$.
 - End {For}
 - For $j = 1$ to m
 - Parameter updating: Determine θ_j as the mean of the vectors $\mathbf{x}_i \in X$ with $b(i) = j$.
 - End {For}.
- Until no change in θ_j 's occurs between two successive iterations.

K-means Algorithm

- Also known as **Lloyd's algorithm**.
- K-means is sometimes synonymous with this algorithm

1: Select K points as the initial centroids.

2: **repeat**

3: Form K clusters by assigning all points to the closest centroid.

4: Recompute the centroid of each cluster.

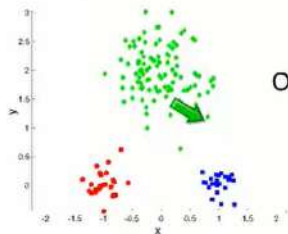
5: **until** The centroids don't change

K-means Algorithm – Initialization

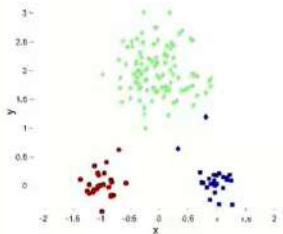
- Initial centroids are often chosen **randomly**.
 - Clusters produced vary from one run to another.



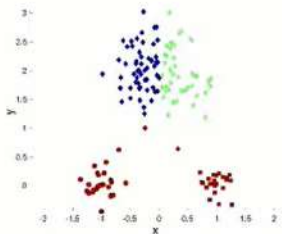
Two different K-means Clusterings



Original Points

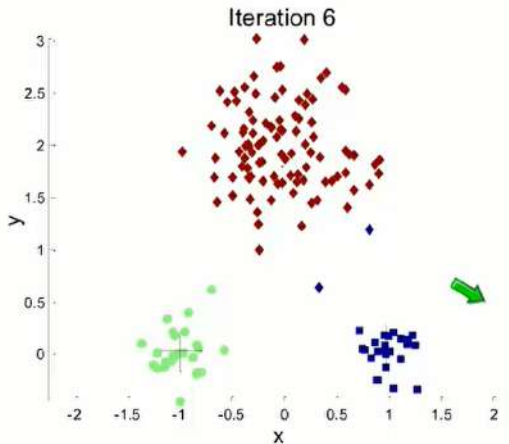


Optimal Clustering

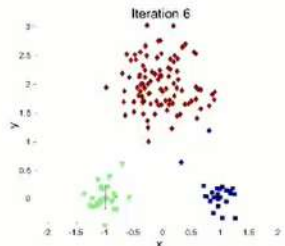
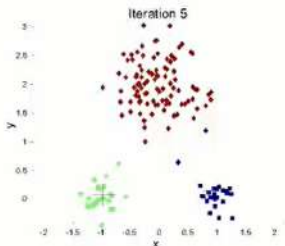
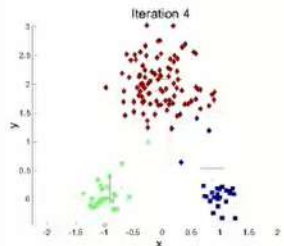
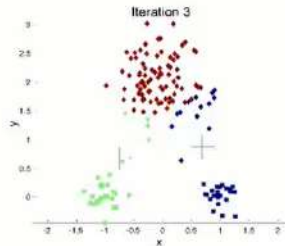
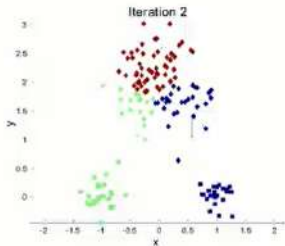
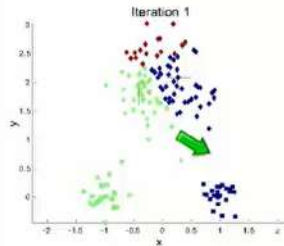


Sub-optimal Clustering

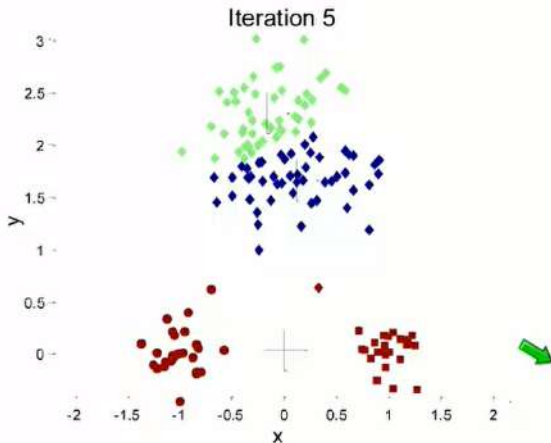
Importance of Choosing Initial Centroids



Importance of Choosing Initial Centroids



Importance of Choosing Initial Centroids



Importance of Choosing Initial Centroids ...

