به نام خدا

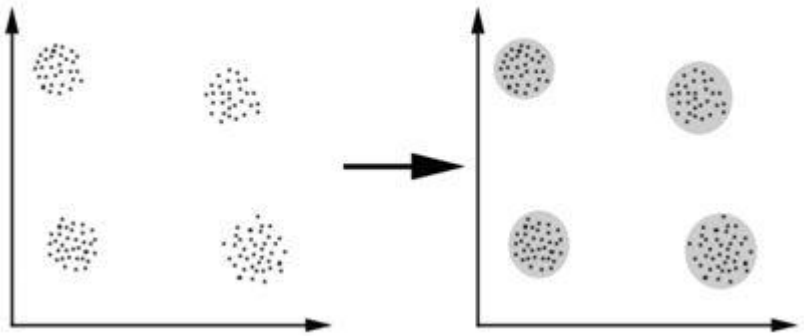
خوشه بندی

CLUSTERING

پاییز ۹۶

مقدمه:

داده و الگو یکی از شاخص‌های بسیار مهم در دنیای اطلاعات هستند. خوشه بندی یکی از بهترین روش‌هایی است که برای کار با داده ها ارائه شده است. قابلیت آن در ورود به فضای داده و تشخیص ساختار انها، خوشه بندی را یکی از ایده ال ترین مکانیزم ها برای کار با دنیای عظیم داده ها کرده است. اولین بار ایده‌ی آن در سال ۱۹۳۵ ارائه شد و امروزه با پیشرفت ها و جهش های عظیمی که در ان پدید امده، خوشه بندی در کاربردها و جنبه های مختلفی حضور یافته است.خوشه بندی یکی از شاخه های یادگیری بدون نظارت می باشد و فرایند خودکاری است که در طی ان، نمونه ها به دسته هایی که اعضای ان مشابه یکدیگر می باشند تقسیم می شوند که به این دسته ها خوشه گفته می‌شود. بنابراین خوشه مجموعه ای از اشیاء می باشد که در ان اشیاء با یکدیگر مشابه بوده و با اشیاء موجود در خوشه های دیگر غیر مشابه می باشند. برای مشابه بودن می توان معیارهای مختلفی را در نظر گرفت مثلا می توان معیار فاصله را برای خوشه بندی مورد استفاده قرار داد و اشیائی را که به یکدیگر نزدیکتر هستند را بعنوان یک خوشه در نظر گرفت که به این نوع خوشه بندی، خوشه بندی مبتنی بر فاصله نیز گفته می شود. بعنوان مثال در شکل ۱ نمونه های ورودی در سمت چپ به چهار خوشه مشابه شکل راست تقسیم می شوند. در این مثال هر یک از نمونه های ورودی به یکی از خوشه ها تعلق دارد و نمونه ای وجود ندارد که متعلق به بیش از یک خوشه باشد.



شکل۱ در این شکل نمونه‌ای از اعمال خوشه‌بندی روی یک مجموعه از داده‌ها مشخص شده است که از معیار فاصله به عنوان عدم شباهت بین داده‌ها استفاده شده است.

What is Clustering?

- **Find K clusters (or a classification that consists of K clusters) so that the objects of one cluster are similar to each other whereas objects of different clusters are dissimilar.**
- **Clustering** is a process of partitioning a set of data (or objects) in a set of meaningful sub-classes, called **clusters**.

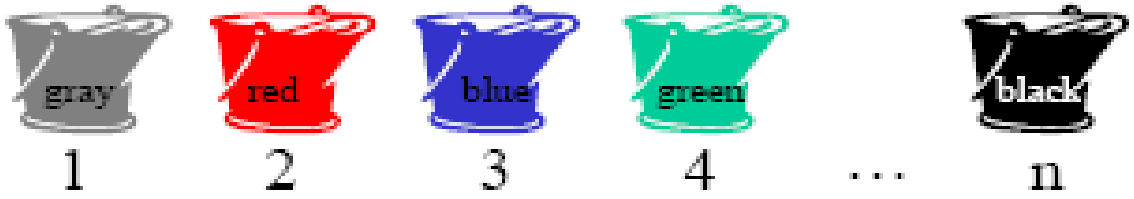
Clustering: unsupervised classification: no predefined classes.

هر کلاستر(خوشه) توسط خصوصیات مشترک اشیائی که درون آن قرار می گیرند تعریف می شود.
کلاستر، یک مجموعه از داده هاست بطوریکه:
۱- داده های موجود در یک کلاستر شبیه یکدیگر هستند.
۲- داده های موجود در کلاسترهای مختلف به یکدیگر شبیه نیستند.

Supervised and Unsupervised

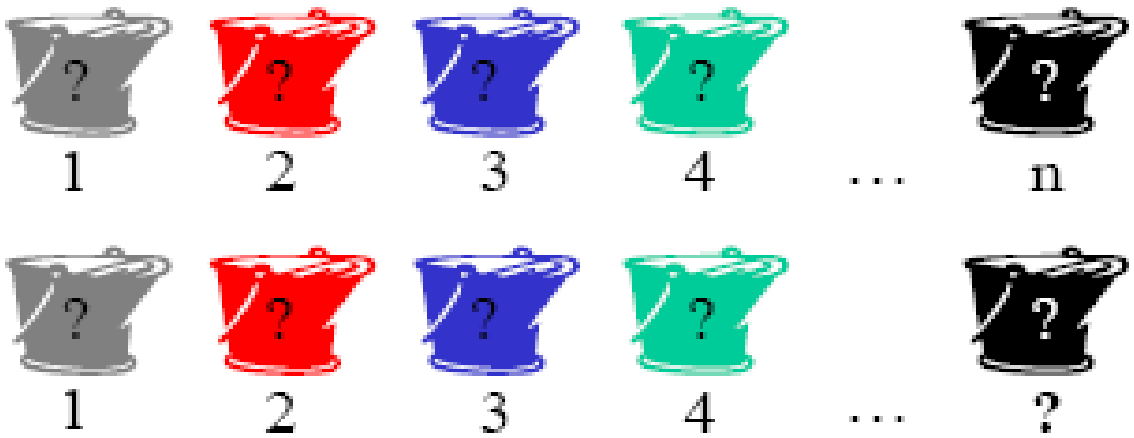
Supervised Classification = Classification

➔ We know the class labels and the number of classes



Unsupervised Classification = Clustering

➔ We do not know the class labels and may not know the number of classes



Types of Features

- continuous range
- Finite discrete set.
- *Nominal*
  sex of an individual
    1 for a male and 0 for a female.
- *Ordinal*
  Values can be *meaningfully ordered*
    Ex. performance of a student in a class
      are 4,3,2,1 = “excellent,” “very good,” “good,” “not good.”

What is a good clustering?

Internal (within the cluster) distances should be small

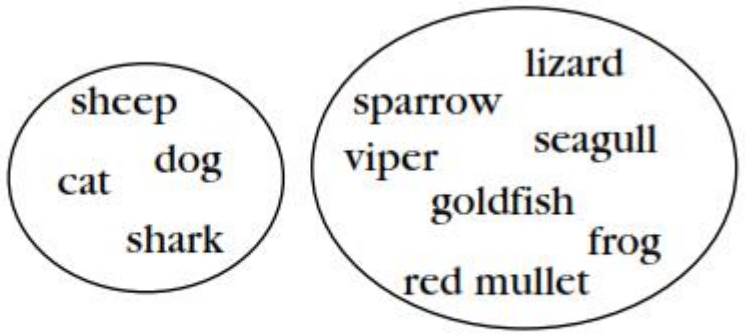
External (intra-cluster) should be large

Clustering is a way to discover new categories (classes)

محاسبه فاصله بین دو داده در خوشه بندی بسیار مهم هست. فاصله که همان معرف عدم تجانس است به ما کمک می کند در فضای داده ای حرکت کنیم و خوشه ها را تشکیل دهیم. با محاسبه فاصله بین دو داده می توان فهمید که چقدر این دو داده به هم نزدیک هستند و بر این اساس آن ها را در یک خوشه قرار داد. توابع ریاضی مختلفی برای محاسبه فاصله وجود دارند. مانند فاصله اقلیدسی، همینگ و...

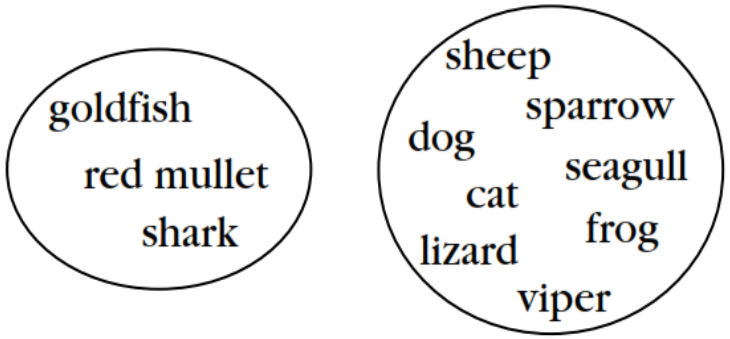
مثال زیر از زیست شناسی بر گرفته شده است و صورت مسئله را برای ما واضح می سازد. به حیوانات زیر توجه کنید: گوسفند، سگ و گربه (پستاندار)، گنجشک و مرغ دریایی (پرنده)، ماهی قرمز، شاه ماهی (ماهی)، افعی و مارمولک(خزنده) و قورباغه(دوزیست). به منظور مرتب کردن این حیوانات در داخل خوشه ها نیاز داریم که یک معیار دسته بندی تعریف کنیم .

اگر ملاک دسته بندی نحوه بدنیا آوردن فرزندان باشد کلاستر (a).



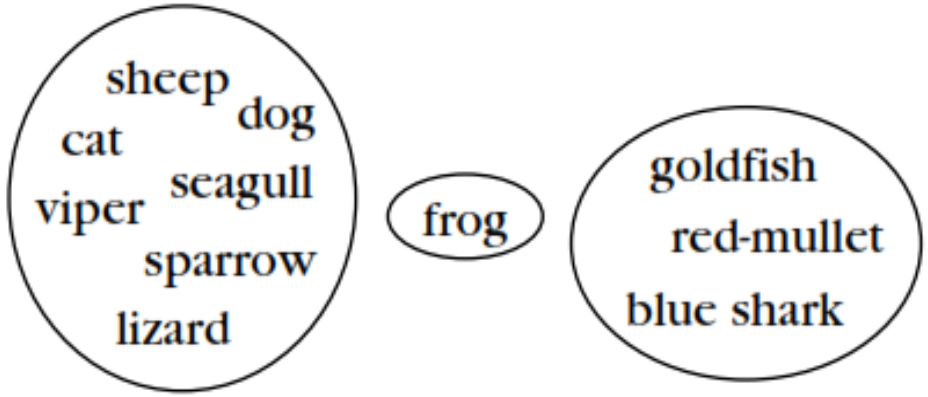
(a)

اگر وجود شش ها را بررسی کنیم، ماهی قرمز، شاه ماهی و کوسه در یک کلاستر و بقیه در یک کلاستر دیگر قرار می گیرند.(b)



(b)

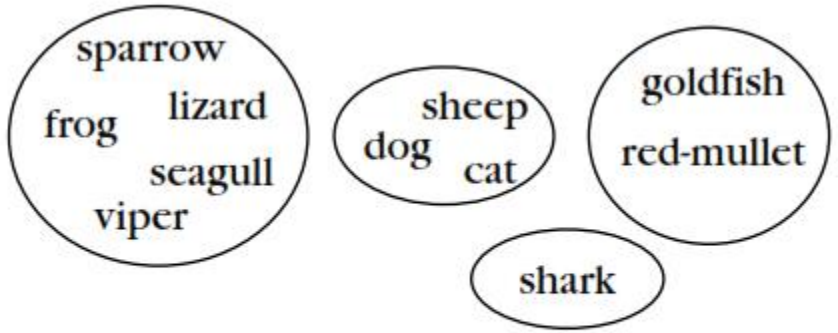
اگر ملاک دسته بندی را محیطی که حیوانات زندگی می کنند قرار دهیم آن گاه گوسفند، سگ، گربه، مرغ دریایی، افعی،گنجشک و مارمولک (حیواناتی که بیرون آب زندگی می کنند) کلاستر اول و ماهی قرمز، شاه ماهی و کوسه(حیواناتی که در آب زندگی می کنند) کلاستر دوم را تشکیل می دهند و قورباغه که می تواند هم در آب و هم در خشکی زندگی کند کلاستر سوم را تشکیل می دهد (C).



(c)

اگر وجودستون فقرات ملاک دسته بندی باشد تمام حیوانات در یک دسته قرار می گیرند.

ما می توانیم از ملاک دسته بندی مرکب استفاده کنیم. برای مثال اگر ملاک دسته بندی نحوه بدنیا آوردن فرزندان و وجود شش ها باشد ما چهار نوع کلاستر داریم(d)



(d)

این مثال نشان می دهد که فرایند نسبت دادن اشیا به کلاسترها ممکن است به نتایج بسیار متفاوتی منجر شود. کلاسترینگ یکی از ابتدایی ترین فعالیت های ذهنی است که برای کنترل کردن مقادیر زیاد اطلاعات دریافت شده استفاده می شود. پردازش هر بخش از اطلاعات به عنوان یک موجودیت تک امکان پذیر نیست. بنابراین انسانها به دسته بندی موجودیت ها (حوادث، انسانها، اشیا و غیره) در کلاسترها روی می آورند

Let us now try to give some definitions for "clustering," which, although they may not be universal, give us an idea of what clustering is. Let X be our data set, that is,

$$X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}.$$

We define as an *m-clustering* of X, $\Re$, the partition of X into m sets (*clusters*), $C_1, \ldots, C_m$, so that the following three conditions are met:

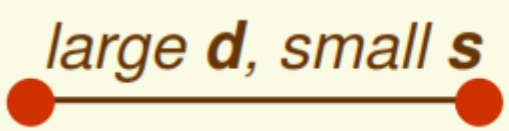
- $C_i \neq \emptyset, i = 1, \ldots, m$
- $\cup_{i=1}^{m} C_i = X$
- $C_i \cap C_j = \emptyset, i \neq j, i,j = 1, \ldots, m$

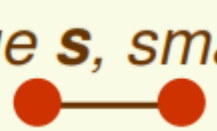
In addition, the vectors contained in a cluster C_i are "more similar" to each other and "less similar" to the feature vectors of the other clusters. Quantifying the terms *similar* and *dissimilar* depends very much on the types of clusters involved.

Proximity measure, either

similarity measure $s(x_i, x_k)$: large if x_i, x_k are similar

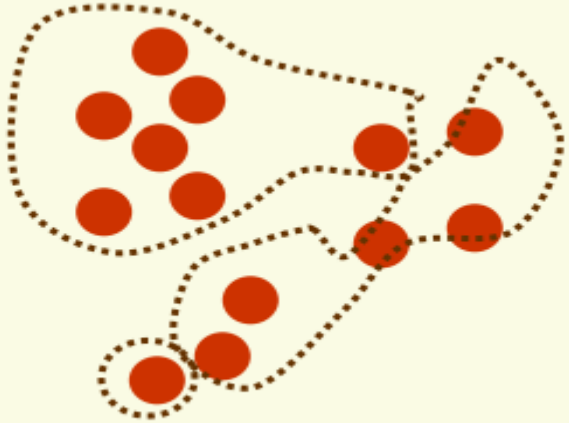
dissimilarity(or distance) measure $d(x_i, x_k)$: small if x_i, x_k are similar





Criterion function to evaluate a clustering

good clustering



bad clustering

PROXIMITY MEASURES

A dissimilarity measure (DM) d on X is a function.

$$d : X \times X \rightarrow R$$

where R is the set of real numbers, such that

$$\exists d_0 \in \mathcal{R} : -\infty < d_0 \leq d(\boldsymbol{x}, \boldsymbol{y}) < +\infty, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in X \tag{11.4}$$

$$d(\boldsymbol{x}, \boldsymbol{x}) = d_0, \quad \forall \boldsymbol{x} \in X \tag{11.5}$$

and

$$d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x}), \quad \forall \boldsymbol{x}, \boldsymbol{y} \in X \tag{11.6}$$

If in addition

$$d(\boldsymbol{x}, \boldsymbol{y}) = d_0 \quad \text{if and only if} \quad \boldsymbol{x} = \boldsymbol{y} \tag{11.7}$$

and

$$d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z}), \quad \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in X \tag{11.8}$$

d is called a *metric DM*. Inequality (11.8) is also known as the *triangular inequality*. Finally, equivalence (11.7) indicates that the minimum possible dissimilarity level value d_0 between any two vectors in X is achieved when they are identical. Sometimes we will refer to the dissimilarity level as distance, where the term is not used in its strict mathematical sense.

A *similarity measure* (SM) s on X is defined as

$$s : X \times X \rightarrow \mathcal{R}$$

such that

$$\exists s_0 \in \mathcal{R} : -\infty < s(\boldsymbol{x}, \boldsymbol{y}) \leq s_0 < +\infty, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in X$$

$$s(\boldsymbol{x}, \boldsymbol{x}) = s_0, \quad \forall \boldsymbol{x} \in X$$

and

$$s(\boldsymbol{x},\boldsymbol{y}) = s(\boldsymbol{y},\boldsymbol{x}), \quad \forall \boldsymbol{x},\boldsymbol{y} \in X \tag{11.11}$$

If in addition

$$s(\boldsymbol{x},\boldsymbol{y}) = s_0 \quad \text{if and only if} \quad \boldsymbol{x} = \boldsymbol{y} \tag{11.12}$$

and

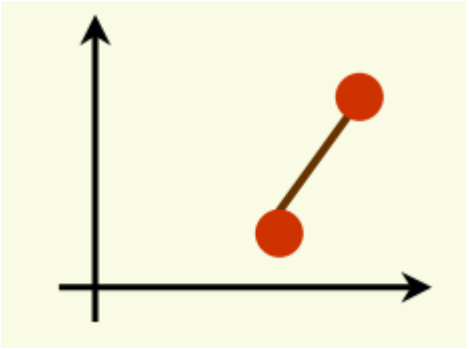
$$s(\boldsymbol{x},\boldsymbol{y})s(\boldsymbol{y},\boldsymbol{z}) \leq [s(\boldsymbol{x},\boldsymbol{y}) + s(\boldsymbol{y},\boldsymbol{z})]s(\boldsymbol{x},\boldsymbol{z}), \quad \forall \boldsymbol{x},\boldsymbol{y},\boldsymbol{z} \in X \tag{11.13}$$

s is called a *metric SM*.

Example 11.2

Let us consider the well-known Euclidean distance, d_2

$$d_2(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{l} (x_i - y_i)^2}$$



where $\boldsymbol{x}, \boldsymbol{y} \in X$ and x_i, y_i are the ith coordinates of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. This is a dissimilarity measure on X, with $d_0 = 0$; that is, the minimum possible distance between two vectors of X is 0. Moreover, the distance of a vector from itself is equal to 0. Also, it is easy to observe that $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$.

In the sequel, we extend the preceding definitions in order to measure "proximity" between subsets of X. Let U be a set containing subsets of X. That is, $D_i \subset X, i = 1, \ldots, k$, and $U = \{D_1, \ldots, D_k\}$. A *proximity measure* $\wp$ *on* U is a function

$$\wp : U \times U \to \mathcal{R}$$

Equations (11.4)–(11.8) for dissimilarity measures and Eqs. (11.9)–(11.13) for similarity measures can now be repeated with D_i, D_j in the place of $\boldsymbol{x}$ and $\boldsymbol{y}$ and U in the place of X.

Usually, the proximity measures between two sets D_i and D_j are defined in terms of proximity measures between elements of D_i and D_j.

Example 11.3

Let $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_6\}$ and $U = \{\{\boldsymbol{x}_1, \boldsymbol{x}_2\}, \{\boldsymbol{x}_1, \boldsymbol{x}_4\}, \{\boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5\}, \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5\}\}$. Let us define the following dissimilarity function:

$$d_{\min}^{ss}(D_i, D_j) = \min_{\boldsymbol{x} \in D_i,\, \boldsymbol{y} \in D_j} d_2(\boldsymbol{x}, \boldsymbol{y})$$

where d_2 is the Euclidean distance between two vectors and $D_i, D_j \in U$.

The minimum possible value of $d_{\min}^{ss}$ is $d_{\min,0}^{ss} = 0$. Also, $d_{\min}^{ss}(D_i, D_i) = 0$, since the Euclidean distance between a vector in D_i and itself is 0. In addition, it is easy to see that the commutative property holds. Thus, this dissimilarity function is a measure. It is not difficult to see that $d_{\min}^{ss}$ is not a metric. Indeed, Eq. (11.7) for subsets of X does not hold in general, since the two sets D_i and D_j may have an element in common. Consider, for example the two sets $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ and $\{\boldsymbol{x}_1, \boldsymbol{x}_4\}$ of U. Although they are different, their distance $d_{\min}^{ss}$ is 0, since they both contain $\boldsymbol{x}_1$.

Proximity Measures between Two Points

Real-Valued Vectors

A. Dissimilarity Measures

The most common DMs between real-valued vectors used in practice are:

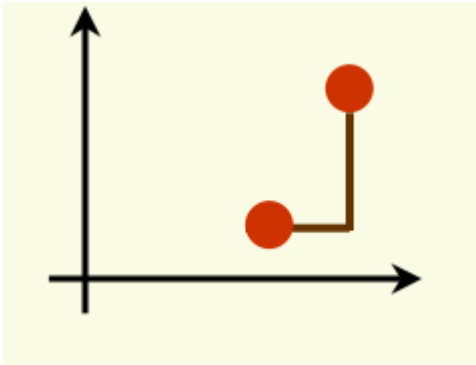
- The *weighted* l_p metric DMs, that is,

$$d_p(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{i=1}^{l} w_i |x_i - y_i|^p \right)^{1/p} \tag{11.14}$$

where x_i, y_i are the ith coordinates of $\boldsymbol{x}$ and $\boldsymbol{y}$, $i = 1, \ldots, l$, and $w_i \geq 0$ is the ith *weight coefficient*. They are used mainly on real-valued vectors. If $w_i = 1, i = 1, \ldots, l$, we obtain the *unweighted* l_p metric DMs.

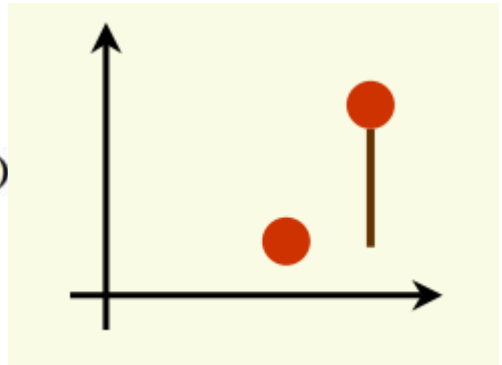
Special l_p metric DMs that are also encountered in practice are the (weighted) l_1 or *Manhattan norm*,

$$d_1(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{l} w_i |x_i - y_i| \tag{11.16}$$



and the (weighted) l_∞ *norm*,

$$d_\infty(\boldsymbol{x}, \boldsymbol{y}) = \max_{1 \le i \le l} w_i |x_i - y_i| \tag{11.17}$$



The l_1 and l_∞ norms may be viewed as overestimation and underestimation of the l_2 norm, respectively. Indeed, it can be shown that $d_\infty(\boldsymbol{x},\boldsymbol{y}) \leq d_2(\boldsymbol{x},\boldsymbol{y}) \leq d_1(\boldsymbol{x},\boldsymbol{y})$ (see Problem 11.6). When $l = 1$ all l_p norms coincide.

Based on these DMs, we can define corresponding SMs as

$$s_p(\boldsymbol{x},\boldsymbol{y}) = b_{\max} - d_p(\boldsymbol{x},\boldsymbol{y}).$$

- Some additional DMs are the following

$$d_G(\boldsymbol{x},\boldsymbol{y}) = -\log_{10}\left(1 - \frac{1}{l}\sum_{j=1}^{l}\frac{|x_j - y_j|}{b_j - a_j}\right) \tag{11.18}$$

where b_j and a_j are the maximum and the minimum values among the jth features of the N vectors of X, respectively. It can easily be shown that $d_G(\boldsymbol{x},\boldsymbol{y})$ is a metric DM. Notice that the value of $d_G(\boldsymbol{x},\boldsymbol{y})$ depends not only on $\boldsymbol{x}$ and $\boldsymbol{y}$ but also on the whole of X. Thus, if $d_G(\boldsymbol{x},\boldsymbol{y})$ is the distance between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ that belong to a set X and $d_G'(\boldsymbol{x},\boldsymbol{y})$ is the distance between the same two vectors when they belong to a different set X', then, in general,

$$d_G(\boldsymbol{x},\boldsymbol{y}) \neq d_G'(\boldsymbol{x},\boldsymbol{y}).$$

Example 11.4

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

Assume now that these vectors belong to a data set X that contains N vectors with maximum values per feature 10, 12, 13 and minimum values per feature 0, 0.5, 1, respectively. Then $d_G(\boldsymbol{x},\boldsymbol{y}) = 0.0922$. If, on the other hand, $\boldsymbol{x}$ and $\boldsymbol{y}$ belong to an X' with the maximum (minimum) values per feature being 20, 22, 23 $(-10, -9.5, -9)$, respectively, then $d_G(\boldsymbol{x},\boldsymbol{y}) = 0.0295$.

Finally, $d_Q(\boldsymbol{x},\boldsymbol{y}) = 0.6455$.

$$d_Q(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\frac{1}{l}\sum_{j=1}^{l}\left(\frac{x_j - y_j}{x_j + y_j}\right)^2}$$

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

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

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

$$d_G(\boldsymbol{x},\boldsymbol{y}) = -\log_{10}\left(1 - \frac{1}{l}\sum_{j=1}^{l}\frac{|x_j - y_j|}{b_j - a_j}\right)$$

B. Similarity Measures

The most common similarity measures for real-valued vectors used in practice are:

The inner product. It is defined as

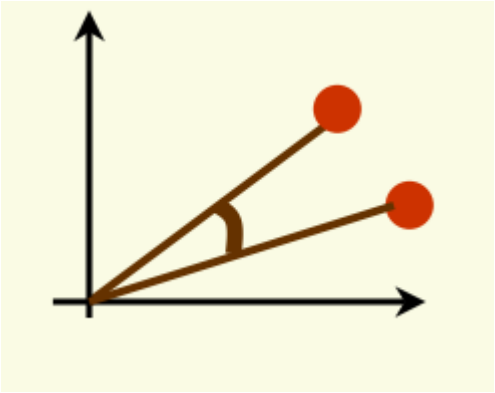
$$s_{\text{inner}}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T \boldsymbol{y} = \sum_{i=1}^{l} x_i y_i.$$

In most cases, the inner product is used when the vectors x and y are normalized.

so that they have the same length a. In these cases, the upper and the lower bounds of s_{inner} are $+a^2$ and $-a^2$, respectively, and $s_{\text{inner}}(\boldsymbol{x}, \boldsymbol{y})$ depends exclusively on the angle between $\boldsymbol{x}$ and $\boldsymbol{y}$.

- cosine similarity measure:

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



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

- Pearson's correlation coefficient.

This measure can be expressed as

$$r_{\text{Pearson}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}_d{}^T \boldsymbol{y}_d}{\|\boldsymbol{x}_d\| \|\boldsymbol{y}_d\|}$$

where $\boldsymbol{x}_d = [x_1 - \bar{x}, \ldots, x_l - \bar{x}]^T$ and $\boldsymbol{y}_d = [y_1 - \bar{y}, \ldots, y_l - \bar{y}]^T$, with x_i, y_i being the ith coordinates of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, and $\bar{x} = \frac{1}{l}\sum_{i=1}^{l} x_i$, $\bar{y} = \frac{1}{l}\sum_{i=1}^{l} y_i$. Usually, $\boldsymbol{x}_d$ and $\boldsymbol{y}_d$ are called difference vectors. Clearly, $r_{\text{Pearson}}(\boldsymbol{x}, \boldsymbol{y})$ takes values between -1 and $+1$.

A related dissimilarity measure can be defined as

$$D(\boldsymbol{x}, \boldsymbol{y}) = \frac{1 - r_{\text{Pearson}}(\boldsymbol{x}, \boldsymbol{y})}{2}$$

- Another commonly used SM is the *Tanimoto measure*, which is also known as Tanimoto distance [Tani 58]. It may be used for real- as well as for discrete-valued vectors. It is defined as

$$s_T(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^T \boldsymbol{y}}{\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - \boldsymbol{x}^T \boldsymbol{y}} \tag{11.23}$$

Proximity Functions between a Point and a Set

In many clustering schemes, a vector x is assigned to a cluster C taking into account the proximity between x and C, $\wp$(x, C).

- The max proximity function:

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

- The min proximity function:

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

- The average proximity function:

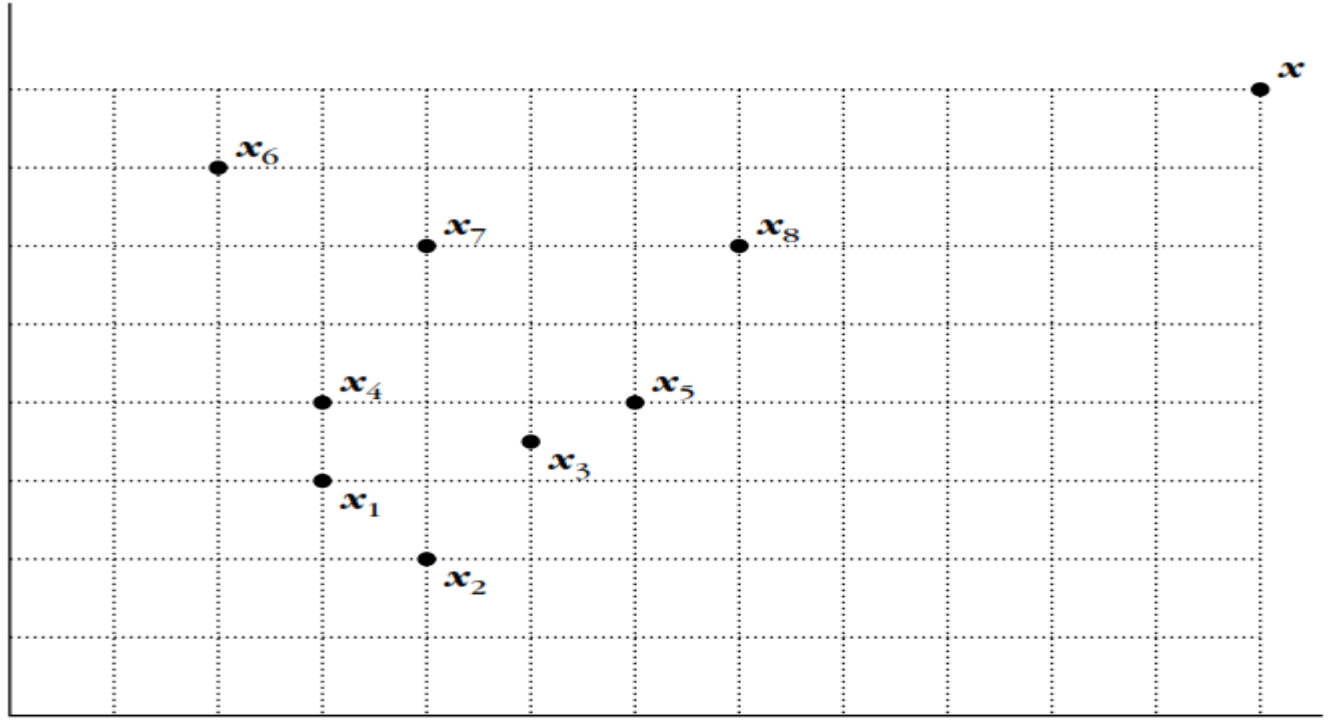
$$\wp^{ps}_{\text{avg}}(\boldsymbol{x}, C) = \frac{1}{n_C} \sum_{\boldsymbol{y} \in C} \wp(\boldsymbol{x}, \boldsymbol{y})$$

where n_C is the cardinality of C.

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

Example 11.9

Let $C = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_6, \boldsymbol{x}_7, \boldsymbol{x}_8\}$, where $\boldsymbol{x}_1 = [1.5, 1.5]^T$, $\boldsymbol{x}_2 = [2, 1]^T$, $\boldsymbol{x}_3 = [2.5, 1.75]^T$, $\boldsymbol{x}_4 = [1.5, 2]^T$, $\boldsymbol{x}_5 = [3, 2]^T$, $\boldsymbol{x}_6 = [1, 3.5]^T$, $\boldsymbol{x}_7 = [2, 3]^T$, $\boldsymbol{x}_8 = [3.5, 3]^T$, and let $\boldsymbol{x} = [6, 4]^T$ (see Figure 11.6). Assume that the Euclidean distance is used to measure the dissimilarity between two points. Then $d_{\max}^{ps}(\boldsymbol{x}, C) = \max_{\boldsymbol{y} \in C} d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{x}, \boldsymbol{x}_1) = 5.15$. For the other two distances we have $d_{\min}^{ps}(\boldsymbol{x}, C) = \min_{\boldsymbol{y} \in C} d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{x}, \boldsymbol{x}_8) = 2.69$ and $d_{\text{avg}}^{ps}(\boldsymbol{x}, C) = \frac{1}{n_C} \sum_{\boldsymbol{y} \in C} d(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{8} \sum_{i=1}^{8} d(\boldsymbol{x}, \boldsymbol{x}_i) = 4.33$.



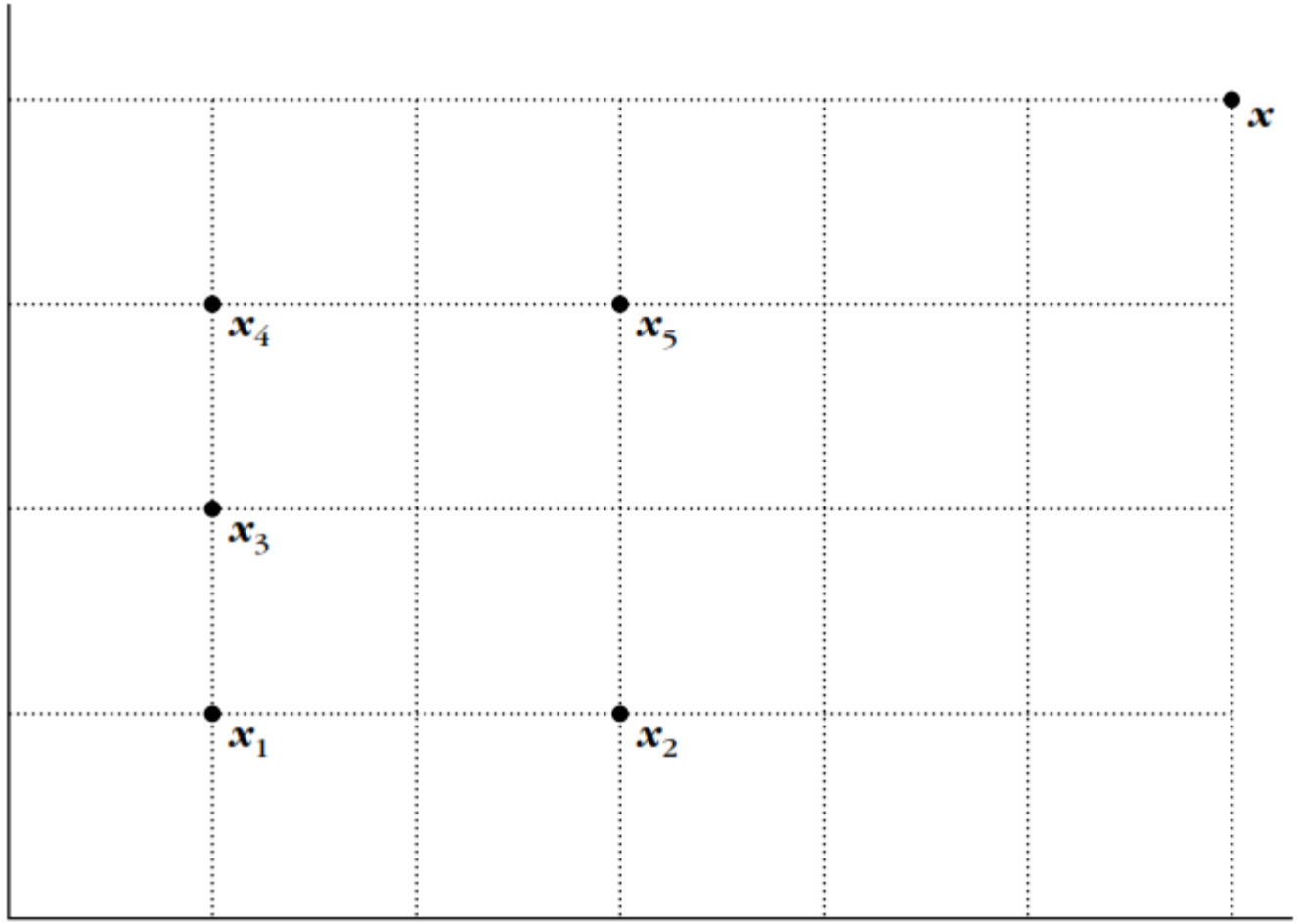
Typical choices for a point representative of a cluster are:

- The mean vector (or mean point)

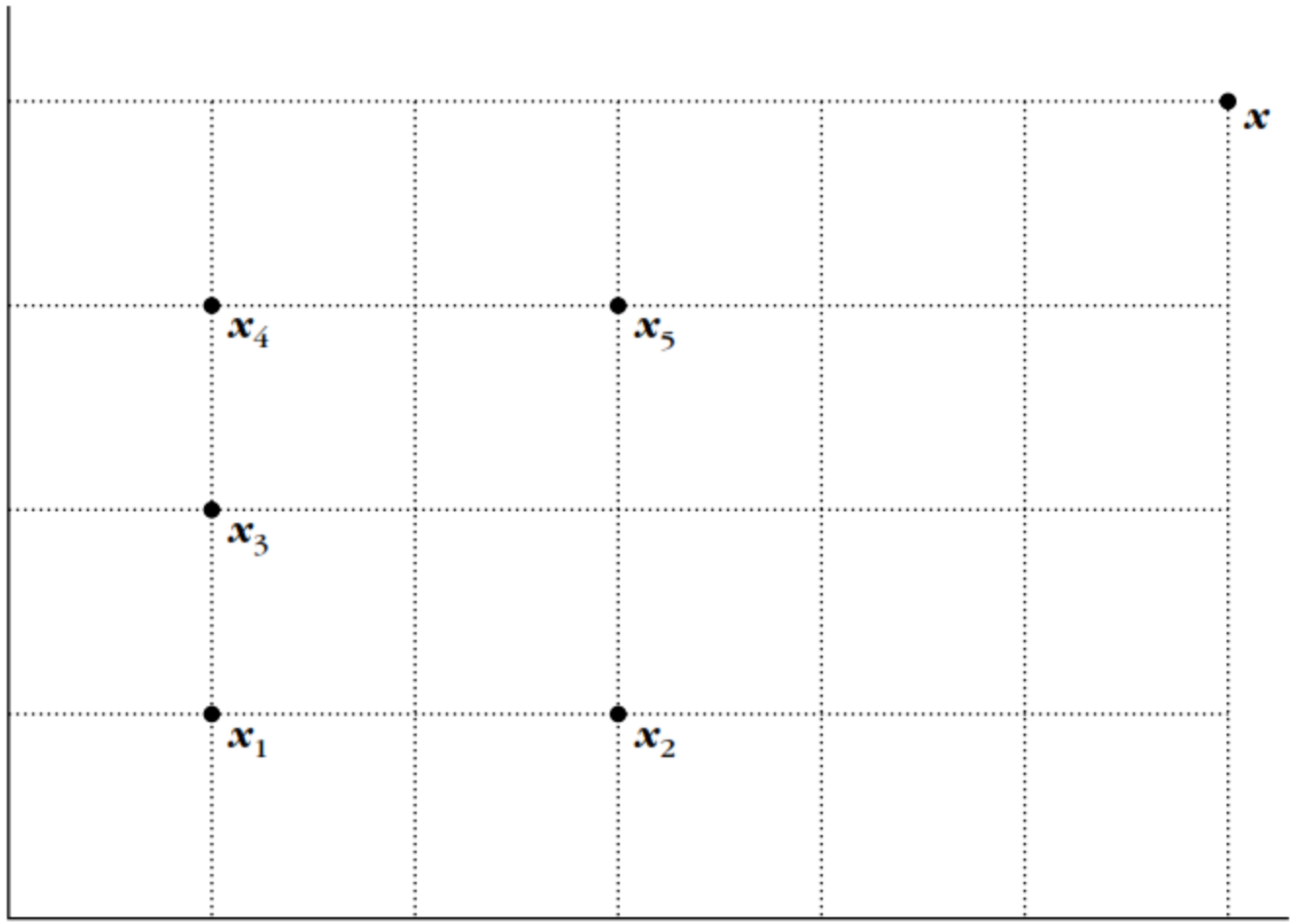
$$\boldsymbol{m}_p = \frac{1}{n_C} \sum_{\boldsymbol{y} \in C} \boldsymbol{y}$$

Example 11.10

Let $C = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5\}$, where $\boldsymbol{x}_1 = [1, 1]^T$, $\boldsymbol{x}_2 = [3, 1]^T$, $\boldsymbol{x}_3 = [1, 2]^T$, $\boldsymbol{x}_4 = [1, 3]^T$, and $\boldsymbol{x}_5 = [3, 3]^T$ (see Figure 11.8). All points lie in the discrete space $\{0, 1, 2, \ldots, 6\}^2$. We use the Euclidean distance to measure the dissimilarity between two vectors in C. The mean point of C is $\boldsymbol{m}_p = [1.8, 2]^T$. It is clear that $\boldsymbol{m}_p$ lies outside the space where the elements of C belong.



To find the mean center $\boldsymbol{m_c}$, we compute, for each point $\boldsymbol{x_i} \in \boldsymbol{C}$, $\boldsymbol{i} = \mathbf{1}, \ldots, \mathbf{5}$, the sum $\boldsymbol{A_i}$ of its distances from all other points of $\boldsymbol{C}$. The resulting values are $\boldsymbol{A_1} = \mathbf{7.83}$, $\boldsymbol{A_2} = \mathbf{9.06}$, $\boldsymbol{A_3} = \mathbf{6.47}$, $\boldsymbol{A_4} = \mathbf{7.83}$, $\boldsymbol{A_5} = \mathbf{9.06}$. The minimum of these values is $\boldsymbol{A_3}$. Thus, $\boldsymbol{x_3}$ is the mean center of $\boldsymbol{C}$.



11.3 PROBLEMS

11.1 Let s be a metric similarity measure on X with $s(\boldsymbol{x}, \boldsymbol{y}) > 0$, $\forall \boldsymbol{x}, \boldsymbol{y} \in X$ and $d(\boldsymbol{x}, \boldsymbol{y}) = a/s(\boldsymbol{x}, \boldsymbol{y})$, with $a > 0$. Prove that d is a metric dissimilarity measure.

11.2 Prove that the Euclidean distance satisfies the triangular inequality.
Hint: Use the Minkowski inequality, which states that for a positive integer p and two vectors $\boldsymbol{x} = [x_1, \ldots, x_l]^T$ and $\boldsymbol{y} = [y_1, \ldots, y_l]^T$ it holds that

$$\left(\sum_{i=1}^{l} |x_i + y_i|^p \right)^{1/p} \leq \left(\sum_{i=1}^{l} |x_i|^p \right)^{1/p} + \left(\sum_{i=1}^{l} |y_i|^p \right)^{1/p}$$

11.6 Prove that

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

for any two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in X.

انواع کلاسترها

کلاستر ها انواع مختلفی دارند که در ادامه به تعدادی از آن ها اشاره خواهد شد:

- ✓ **کلاسترهای بخوبی جدا شده:**

مجموعه نقاط داخل این کلاستر نسبت به نقاط خارج آن به یکدیگر بسیار شبیه ترند.

- ✓ **کلاسترهای مبتنی به مرکز:**

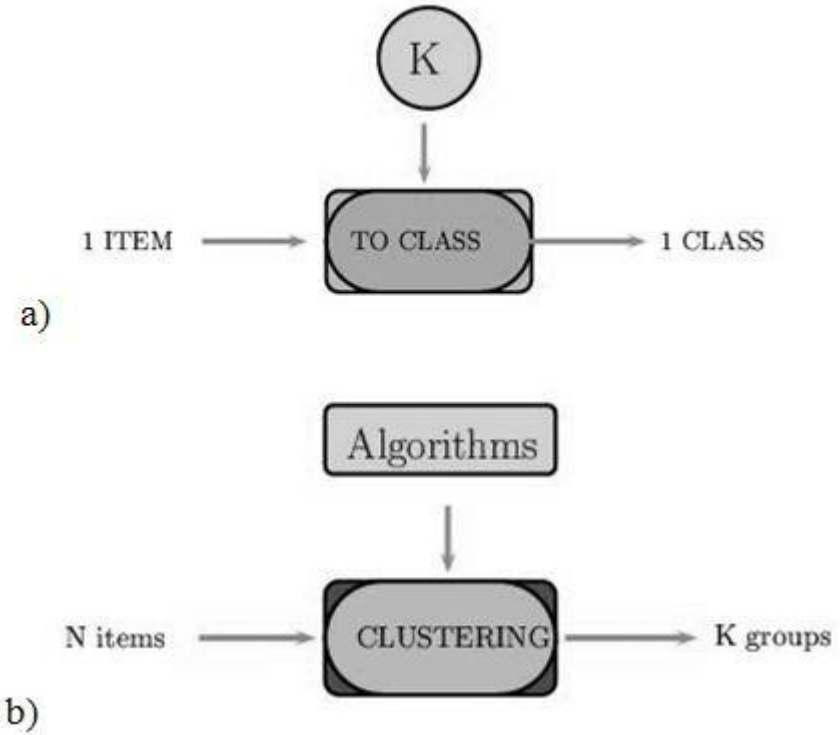
مجموعه نقاط داخل این کلاستر به مرکز کلاستر نسبت به مراکز کلاسترهای دیگر بسیار نزدیکترند.

- **کلاسترهای مبتنی بر مجاورت و نزدیکی:**

مجموعه نقاط داخل این کلاستر به یک یا تعداد بیشتری از نقاط داخل کلاستر نسبت به نقاط خارج آن شبیه اند.

خوشه‌بندی در مقابل طبقه‌بندی

در طبقه‌بندی، دسته‌ها(کلاس) از پیش مشخص شده‌اند و هر داده به یک دسته از پیش مشخص شده تخصیص می‌یابد. ولی در خوشه‌بندی هیچ اطلاعی از کلاسهای موجود درون داده‌ها وجود ندارد و به عبارتی خود خوشه‌ها نیز از داده‌ها استخراج می‌شوند. در شکل زیر تفاوت بین خوشه‌بندی و طبقه‌بندی بهتر نشان داده شده است.



a)در طبقه‌بندی با استفاده یک سری اطلاعات اولیه داده‌ها به دسته‌های معلومی نسبت داده می‌شوند

b)در خوشه‌بندی داده‌ها با توجه به الگوریتم انتخاب شده به خوشه‌هایی نسبت داده می‌شوند

کاربردها

یک جستجوی ساده در وب یا حتی در پایگاه داده ی یک کتابخانه، کاربرد شگفت انگیز خوشه بندی را برای ما اشکار می سازد. از انجا که خوشه‌بندی یک روش یادگیری بدون نظارت محسوب می گردد، در موارد بسیاری می‌تواند کاربرد داشته باشد:

- **بازاریابی:** دسته‌بندی مشتری‌ها به دسته‌هایی بر حسب رفتارها و نیازهای آن ها از طریق مجموعه زیادی از ویژگی‌ها و اخرین خریدهای ان ها.
- **زیست‌شناسی:** دسته‌بندی حیوانات و گیاهان از روی ویژگی‌های انها
- **کتابداری :** دسته‌بندی کتابها
- **نقشه‌برداری شهری**: دسته‌بندی خانه‌ها بر اساس نوع و موقعیت جغرافیایی انها.
- **مطالعات زلزله‌نگاری:**تشخیص مناطق حادثه‌خیز بر اساس مشاهدات قبلی.
- **وب**: دسته‌بندی اسناد، دسته‌بندی مشتریان مراجعه کننده به سایتها و ....
- **داده کاوی:** کشف اطلاعات و ساختار جدید از داده‌های موجود
- **در تشخیص گفتار:**در ساخت کتابِ کد از بردارهای ویژگی، در تقسیم کردن گفتار بر حسب گویندگان ان و یا فشرده‌سازی گفتار
- **در تقسیم‌بندی تصاویر:**تقسیم‌بندی تصاویر پزشکی و یا ماهواره‌ای
- **بیمه:**تشخیص افراد متقلب، تشخیص افرادی که بیمه موتور دارند و بیشترین میزان درخواست از بیمه را نیز در سال مشخصی داشته اند.

ایرادات

متاسفانه چندین مسئله در خصوص روش‌های خوشه‌بندی مطرح است که هنوز به شکل کامل پاسخ داده نشده‌اند. و همچنان تلاش‌های بسیاری به منظور حل انها انجام گرفته است:

- روش‌های خوشه‌بندی قادر نیستند تمامی نیازهای مسائل را به طور هم‌زمان برآورده کنند.
- به دلیل پیچیدگی محاسباتی زیاد در برخورد با مجموعه داده‌های بزرگ (تعداد داده زیاد وتعداد ویژگی‌های زیاد برای هر داده) عملی نیستند.
- به دلیل وابستگی شدید به تعریف معیار شباهت بین داده‌ها، در مسائلی که تعریف معیار شباهت مشکل باشد نتایج مطلوبی تولید نمی‌کنند.(در داده‌ها با تعداد ویژگی زیاد)

AGGLOMERATIVE ALGORITHMS

Let $g(C_i, C_j)$ be a function defined for all possible pairs of clusters of X. This function measures the proximity between C_i and C_j. Let t denote the current level of hierarchy. Then, the general agglomerative scheme may be stated as follows:

Generalized Agglomerative Scheme (GAS)

- **Initialization:**
  - Choose $\Re_0 = \{C_i = \{\boldsymbol{x}_i\},\ i = 1, \ldots, N\}$ as the initial clustering.
  - $t = 0$.
- **Repeat:**
  - $t = t + 1$
  - Among all possible pairs of clusters (C_r, C_s) in $\Re_{t-1}$ find the one, say (C_i, C_j), such that

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

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

Example 13.1

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

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

and its corresponding dissimilarity matrix, when the Euclidean distance is in use, is

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

When the Tanimoto measure is used, the similarity matrix of X becomes

$$P'(X) = \begin{bmatrix} 1 & 0.75 & 0.26 & 0.21 & 0.18 \\ 0.75 & 1 & 0.44 & 0.35 & 0.20 \\ 0.26 & 0.44 & 1 & 0.96 & 0.90 \\ 0.21 & 0.35 & 0.96 & 1 & 0.98 \\ 0.18 & 0.20 & 0.90 & 0.98 & 1 \end{bmatrix}$$

Note that in $P(X)$ all diagonal elements are 0, since $d_2(\boldsymbol{x}, \boldsymbol{x}) = 0$, while in $P'(X)$ all diagonal elements are equal to 1, since $s_T(\boldsymbol{x}, \boldsymbol{x}) = 1$.

$$g(C_i, C_j) = d_{\min}^{ss}(C_i, C_j)$$

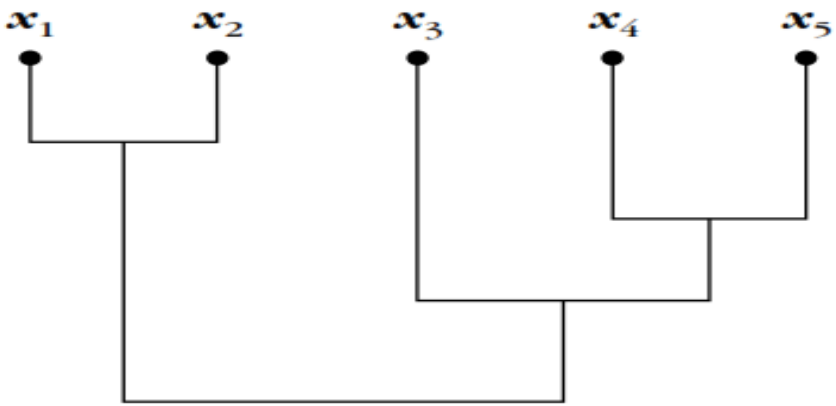
$\{\{x_1\},\{x_2\},\{x_3\},\{x_4\},\{x_5\}\}$

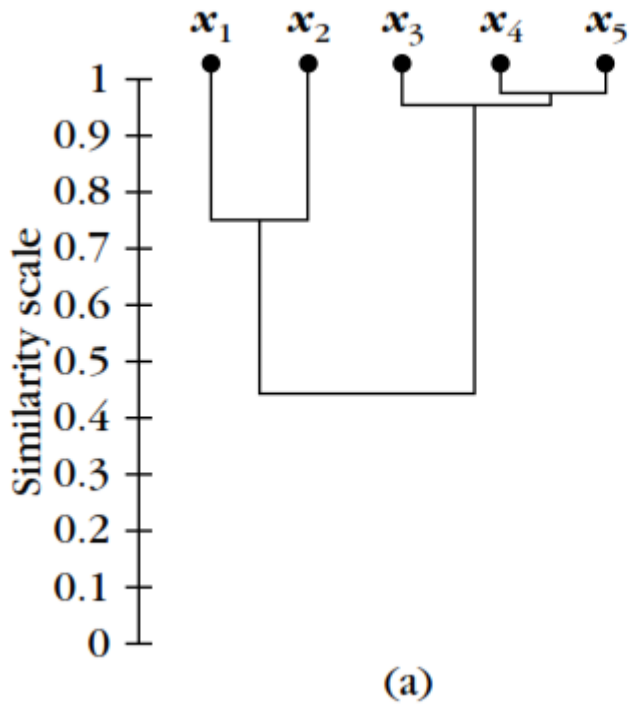
$\{\{x_1, x_2\},\{x_3\},\{x_4\},\{x_5\}\}$

$\{\{x_1, x_2\},\{x_3\},\{x_4, x_5\}\}$

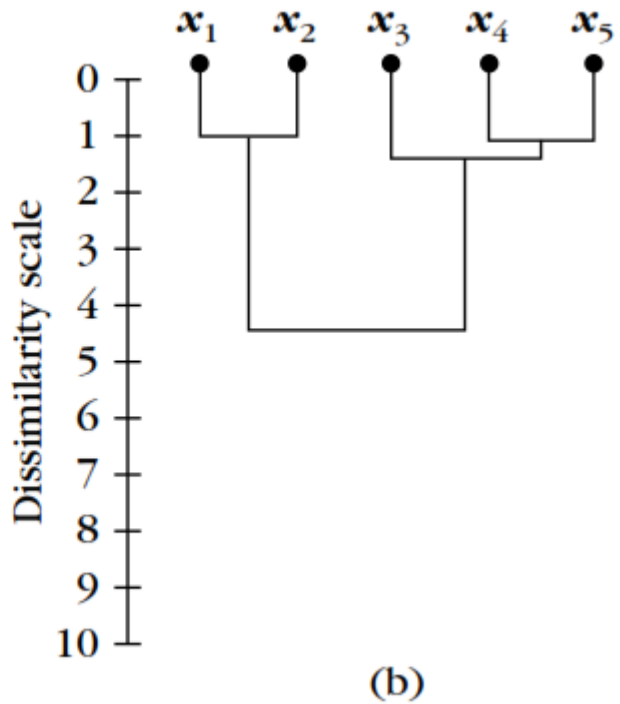
$\{\{x_1, x_2\},\{x_3, x_4, x_5\}\}$

$\{\{x_1, x_2, x_3, x_4, x_5\}\}$



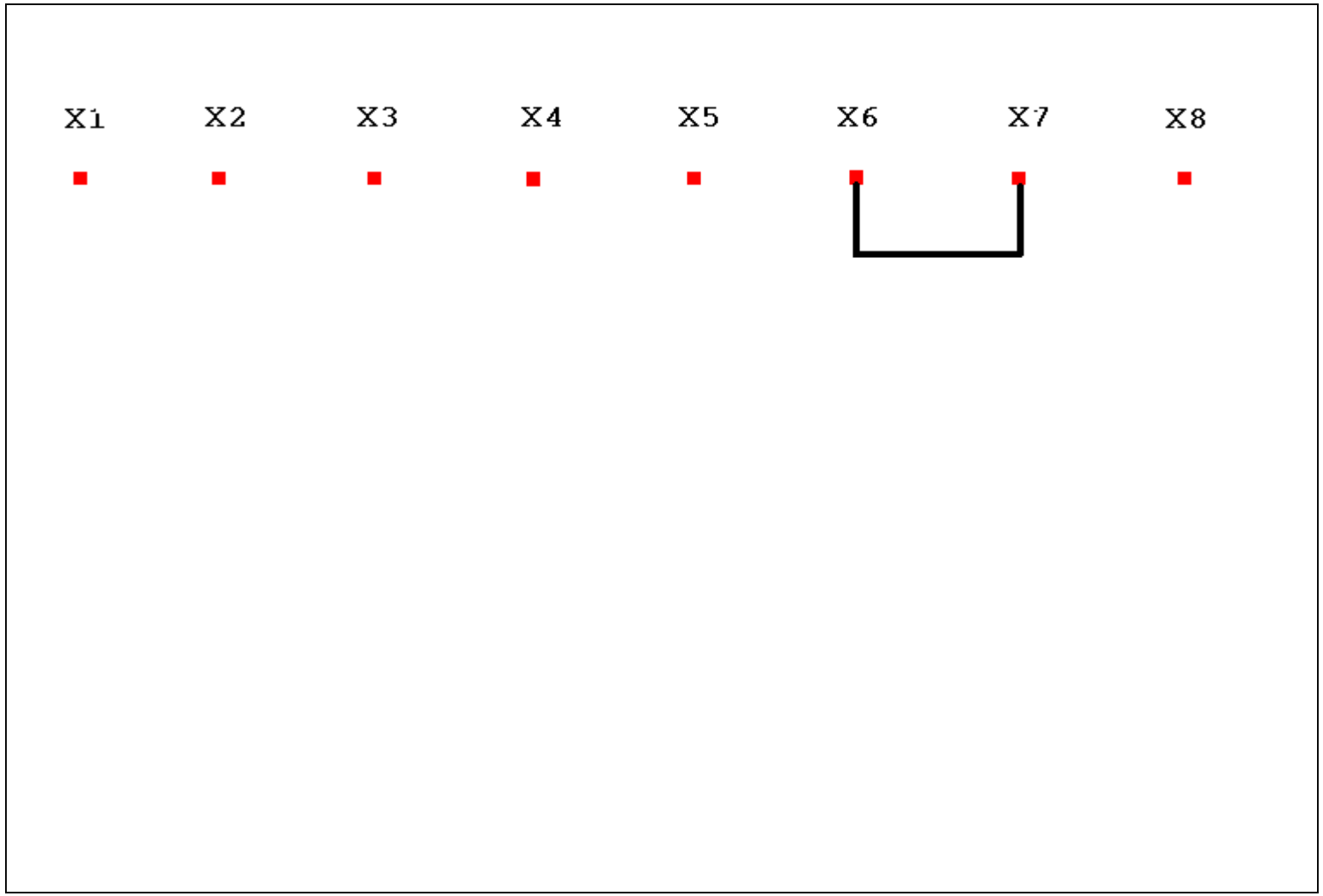


(a)

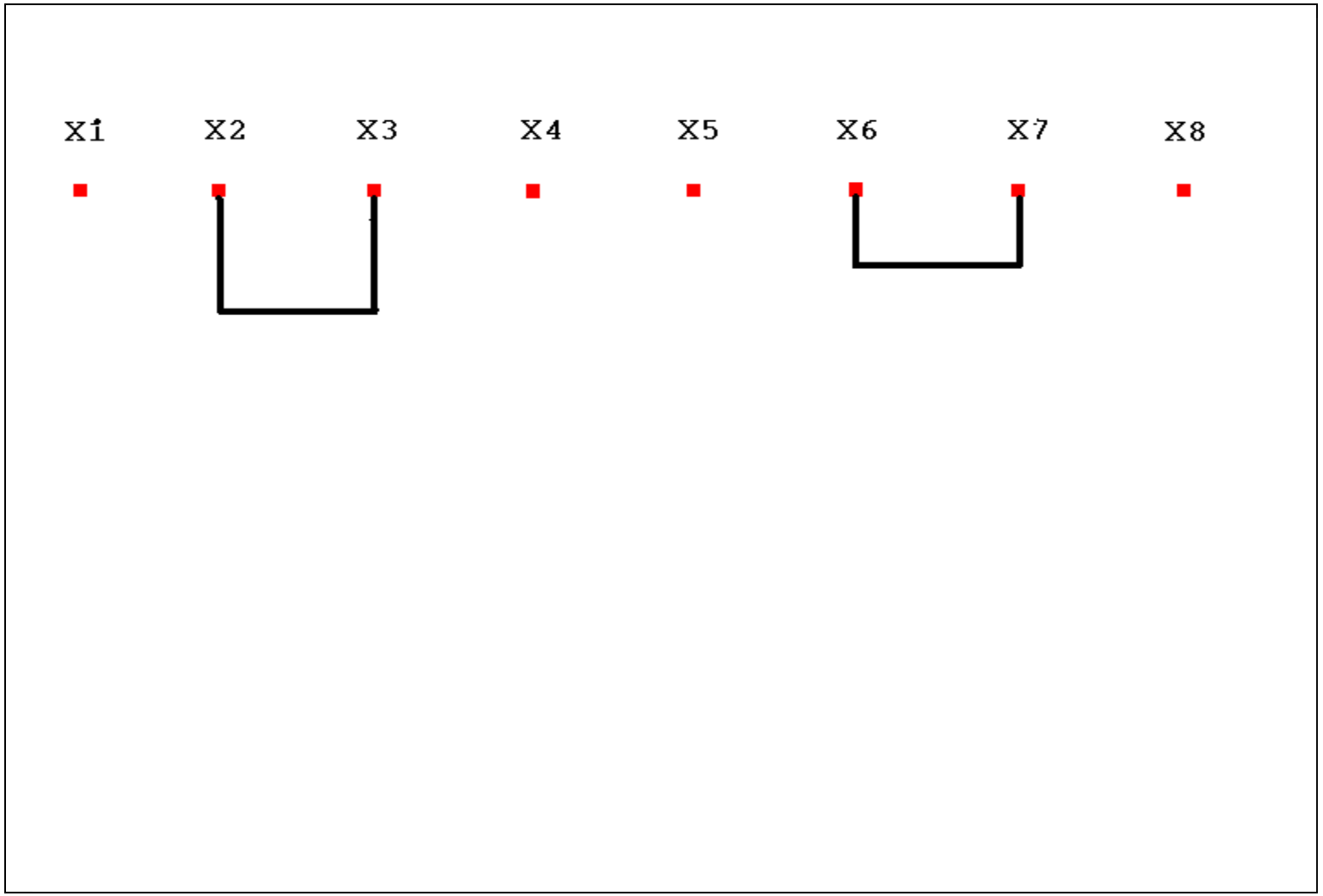


(b)

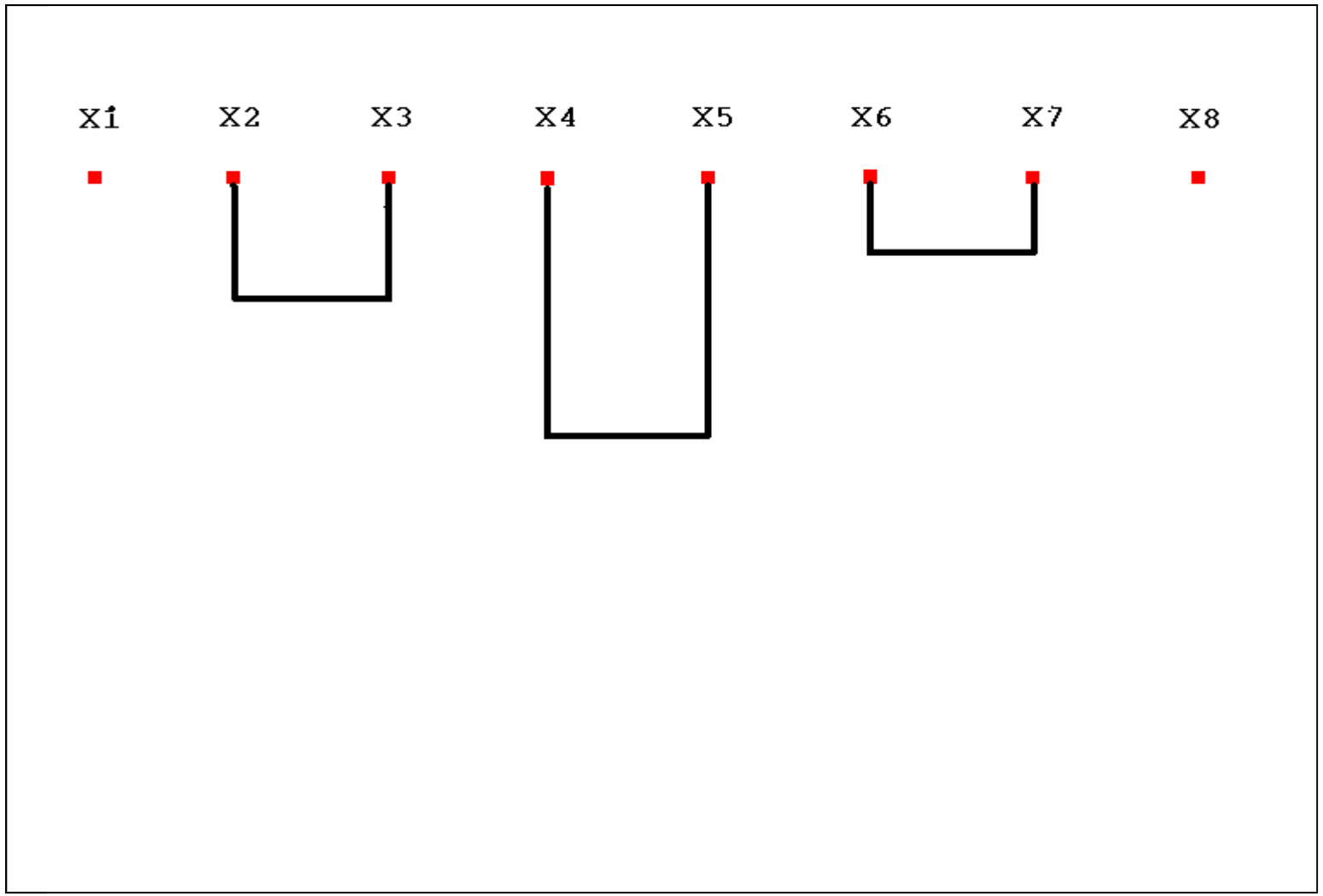
Nearest Neighbor, Level 2, k = 7 clusters.



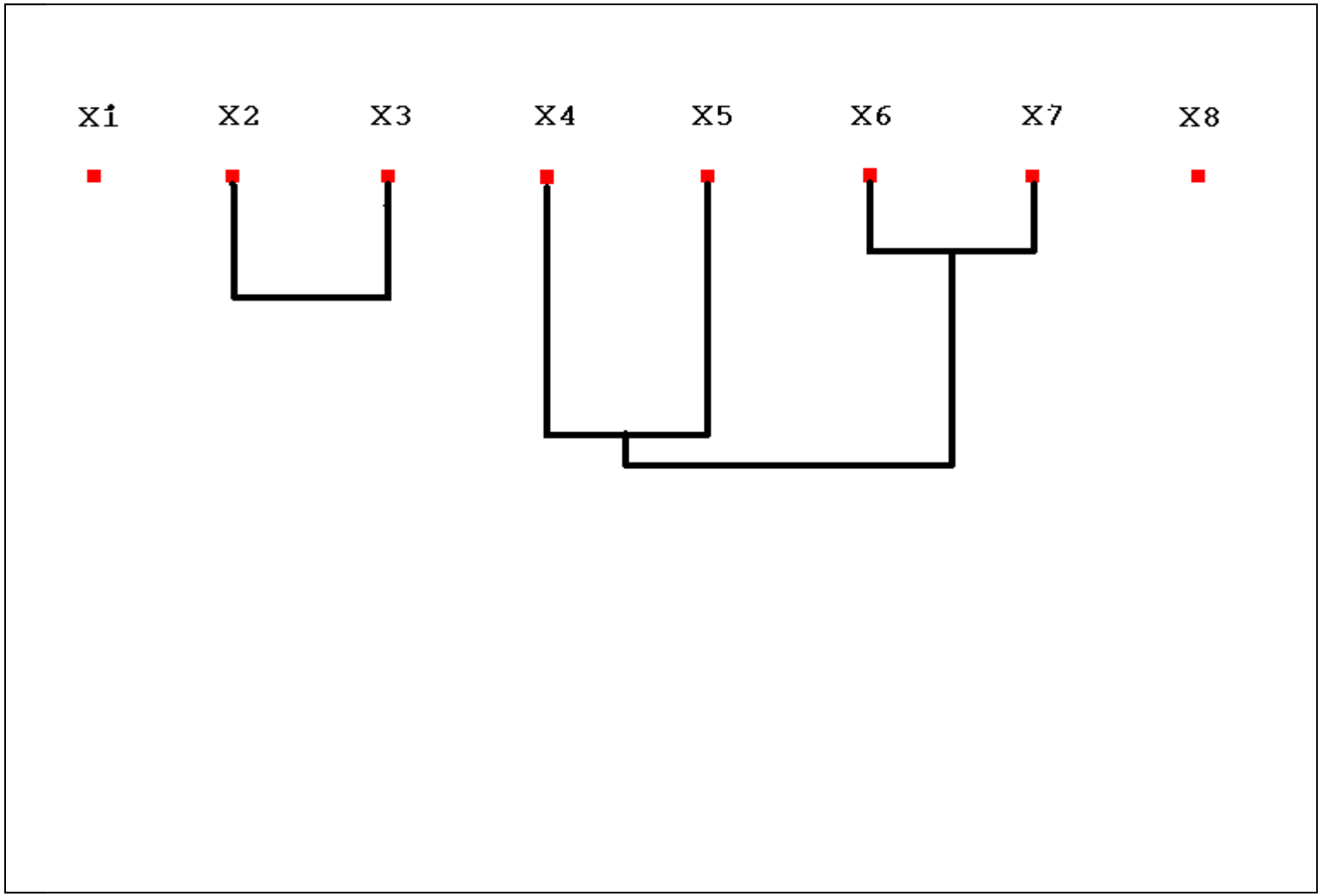
Nearest Neighbor, Level 3, k = 6 clusters.



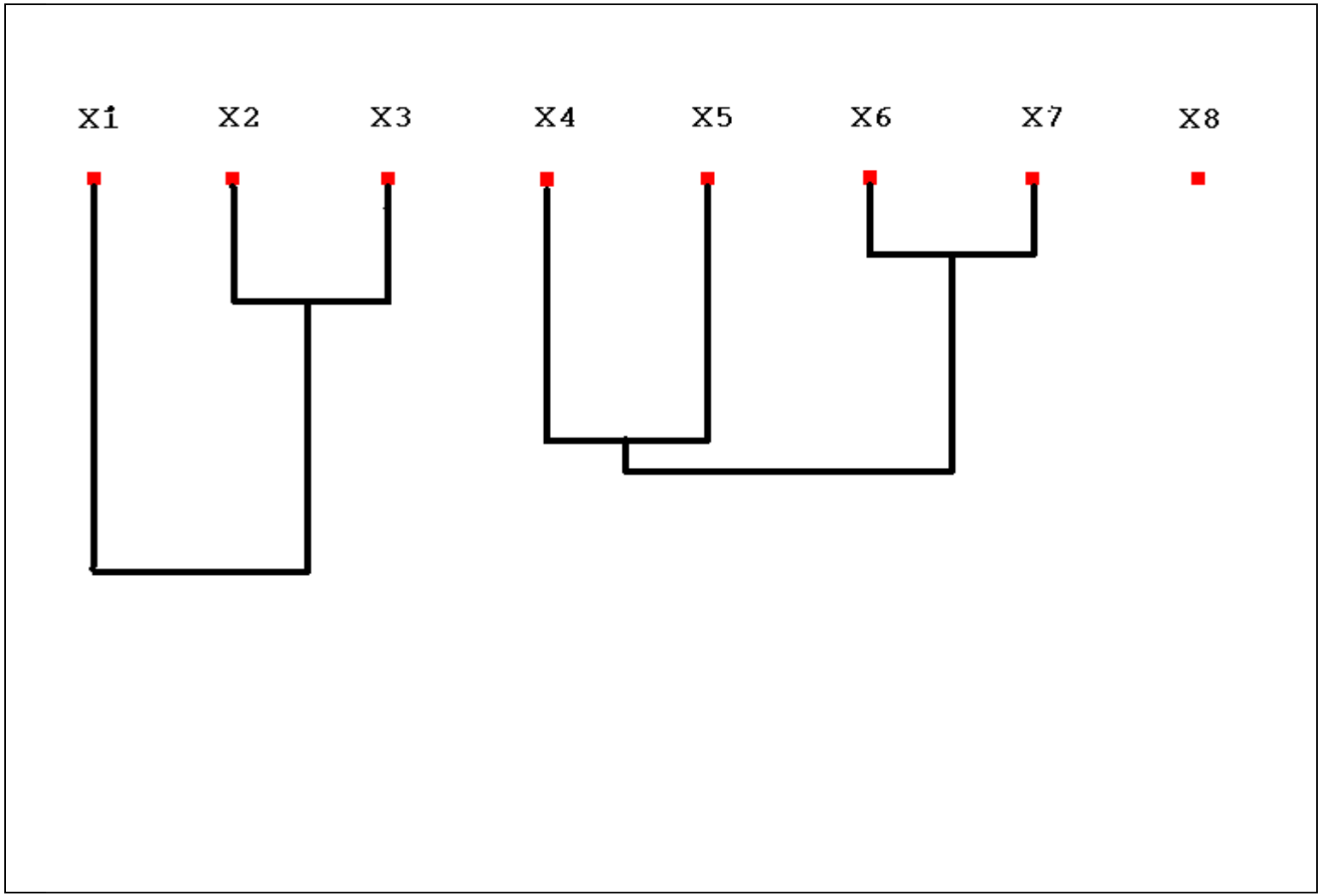
Nearest Neighbor, Level 4, k = 5 clusters.



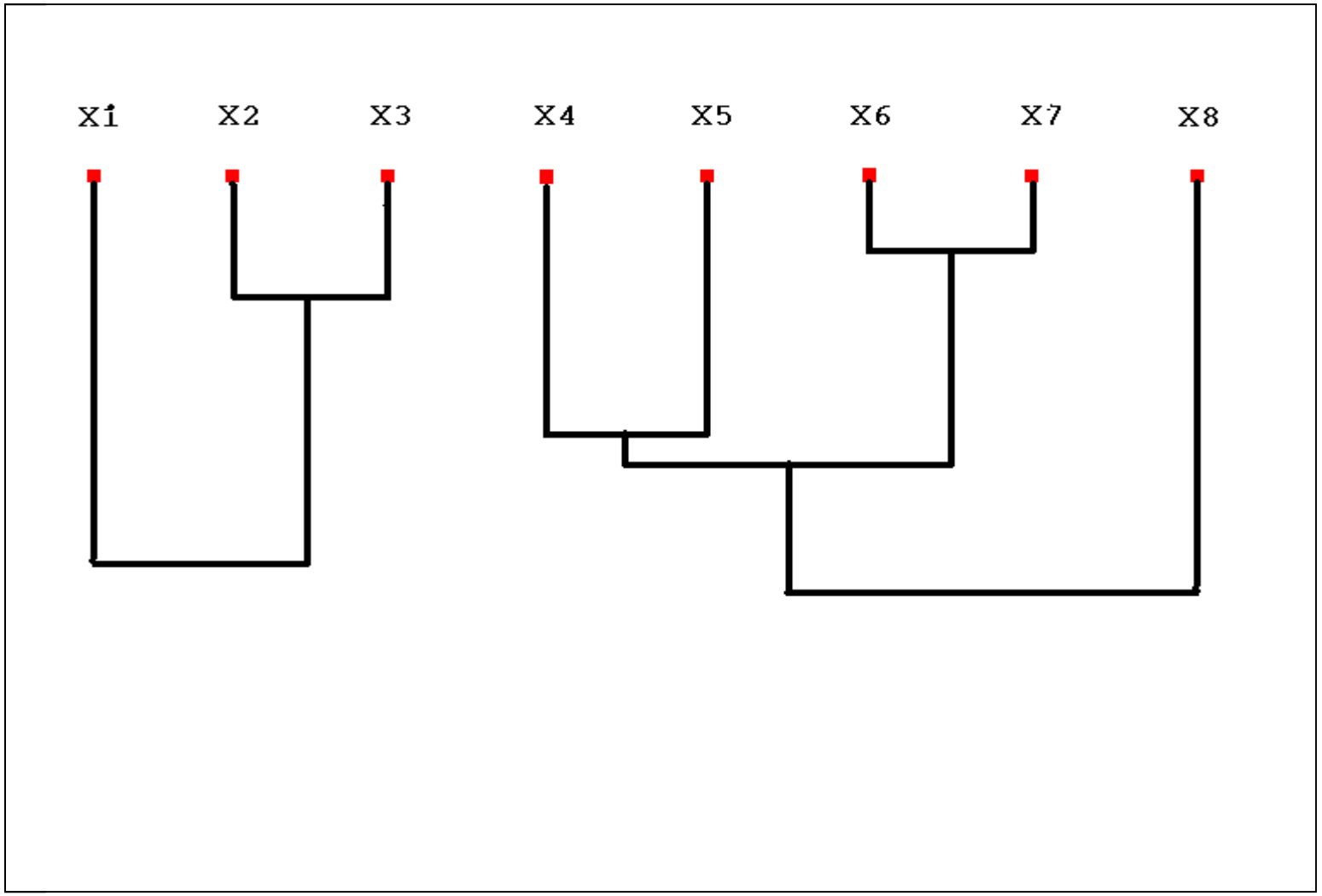
***Nearest Neighbor, Level 5, k = 4 clusters*.**



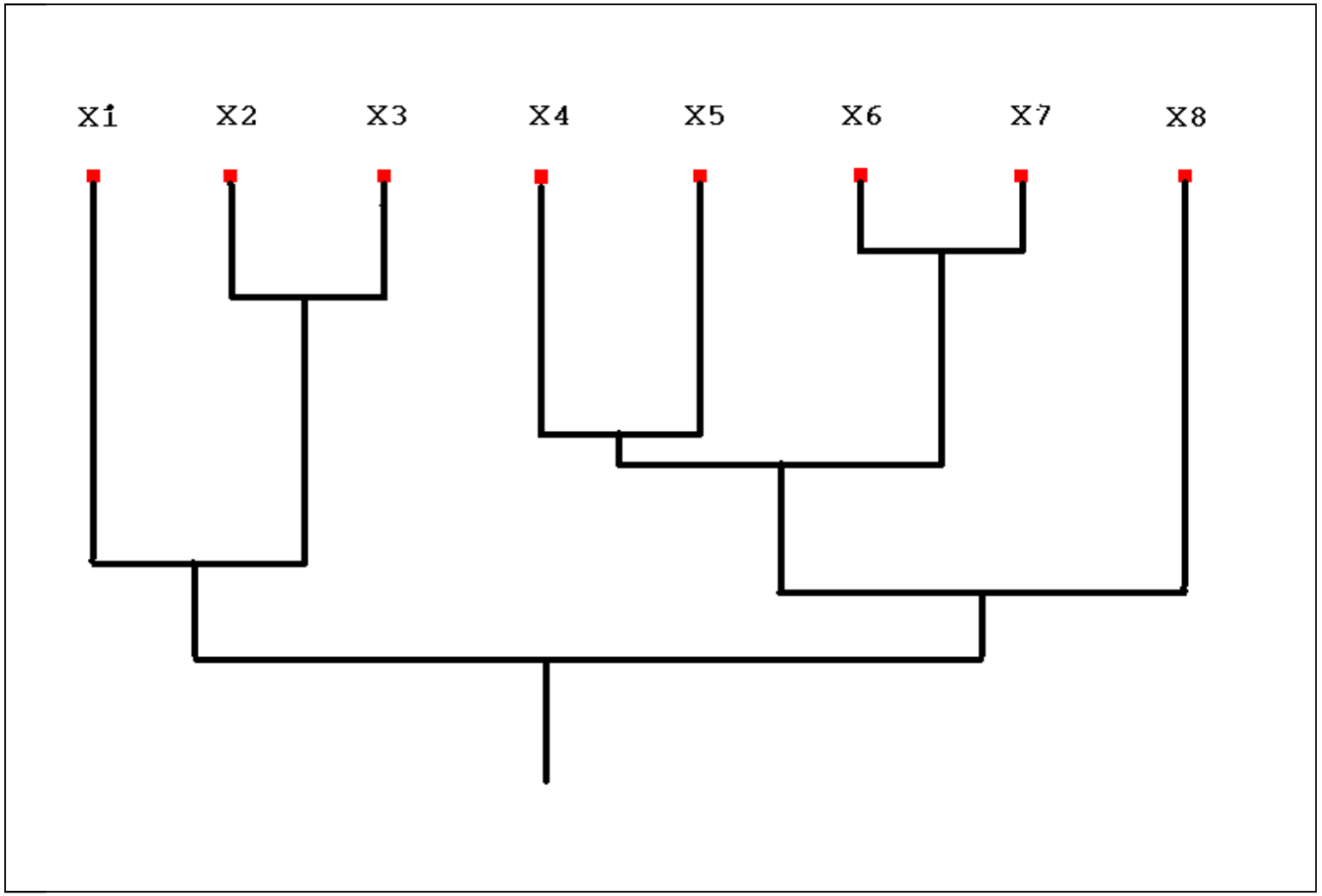
***Nearest Neighbor, Level 6, k = 3 clusters*.**

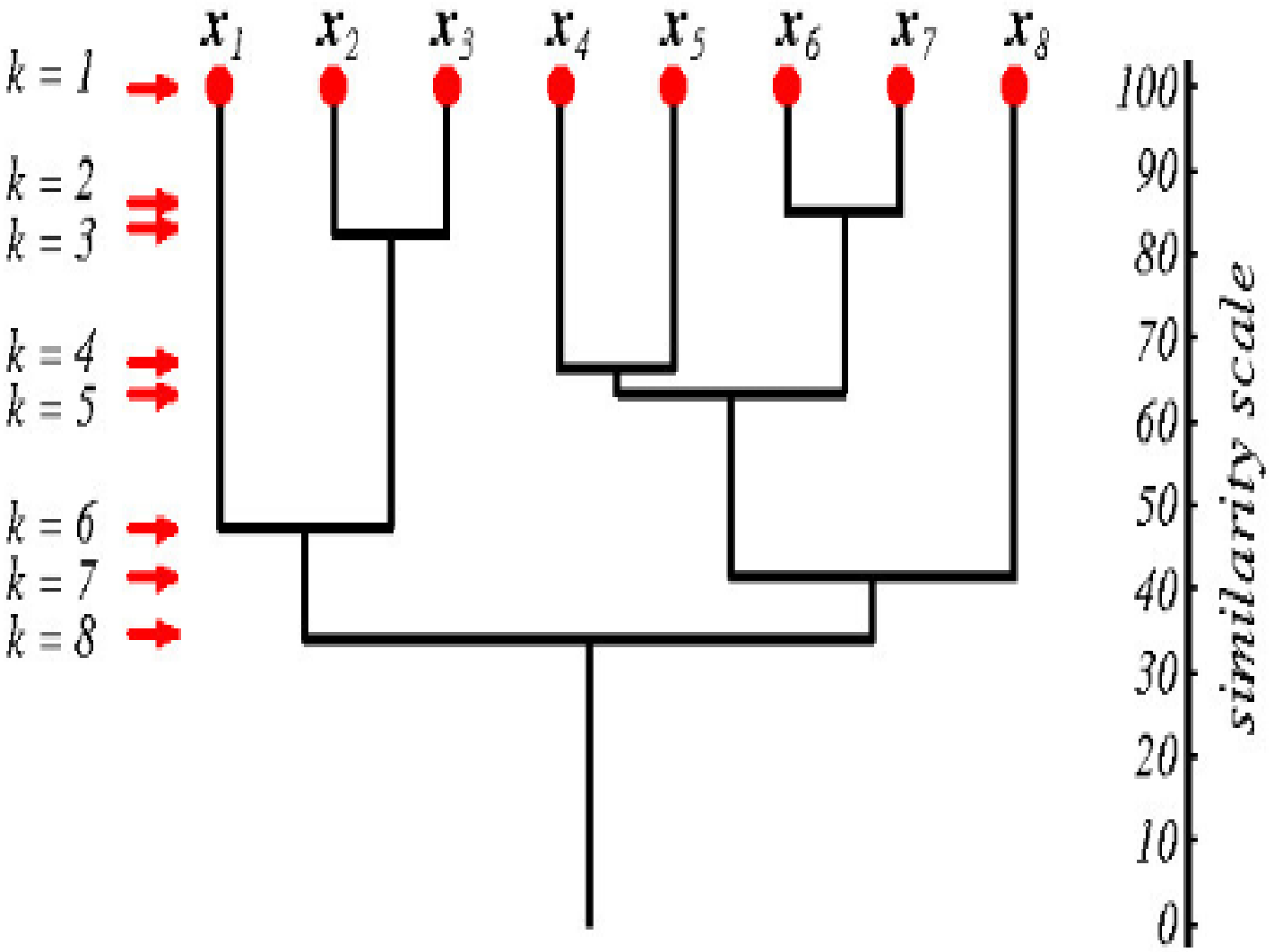


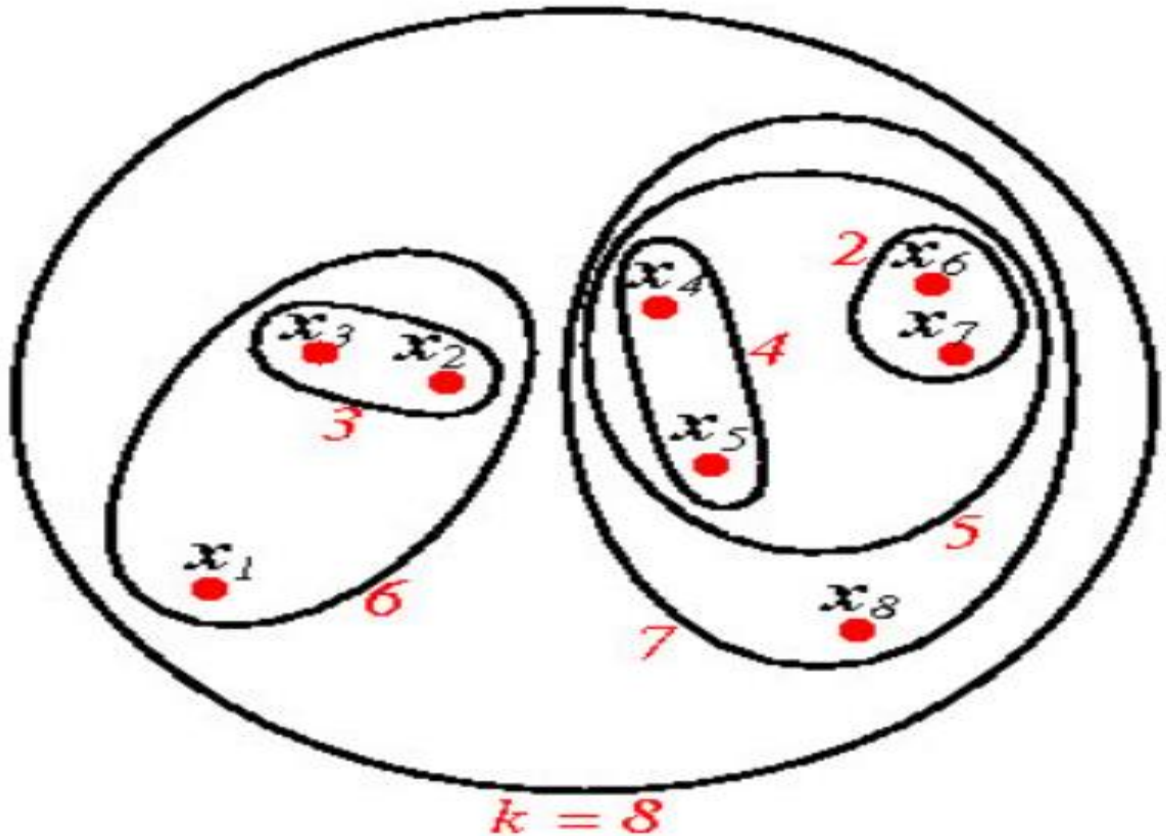
***Nearest Neighbor, Level 7, k = 2 clusters*.**



Nearest Neighbor, Level 8, k = 1 cluster.




x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8
$k = 1$
$k = 2$
$k = 3$
$k = 4$
$k = 5$
$k = 6$
$k = 7$
$k = 8$
100
90
80
70
60
50
40
30
20
10
0
similarity scale


x_1
x_2
x_3
x_4
x_5
x_6
x_7
x_8
2
3
4
5
6
7
$k = 8$

K- Means

این روش علی‌رغم سادگی آن یک روش پایه برای بسیاری از روش‌های خوشه‌بندی دیگر (مانند خوشه‌بندی فازی) محسوب می‌شود. این روش روشی انحصاری و مسطح محسوب می‌شود. برای این الگوریتم شکلهای مختلفی بیان شده است. ولی همهٔ آنها دارای روالی تکراری هستند که برای تعدادی ثابت از خوشه‌ها سعی در تخمین موارد زیر دارند:

- بدست آوردن نقاطی به عنوان مراکز خوشه‌ها. این نقاط در واقع همان میانگین نقاط متعلق به هر خوشه هستند.
- نسبت دادن هر نمونه داده به یک خوشه که آن داده کمترین فاصله تا مرکز آن خوشه را دارا باشد.

در نوع ساده‌ای از این روش ابتدا به تعداد خوشه‌های مورد نیاز نقاطی به صورت تصادفی انتخاب می‌شود. سپس در داده‌ها با توجه با میزان نزدیکی (شباهت) به یکی از این خوشه‌ها نسبت داده می‌شوند و بدین ترتیب خوشه‌های جدیدی حاصل می‌شود. با تکرار همین روال می‌توان در هر تکرار با میانگین‌گیری از داده‌ها مراکز جدیدی برای آنها محاسبه کرد و مجدداً داده‌ها را به خوشه‌های جدید نسبت داد. این روند تا زمانی ادامه پیدا می‌کند که دیگر تغییری در داده‌ها حاصل نشود. تابع زیر به عنوان تابع هدف مطرح است.

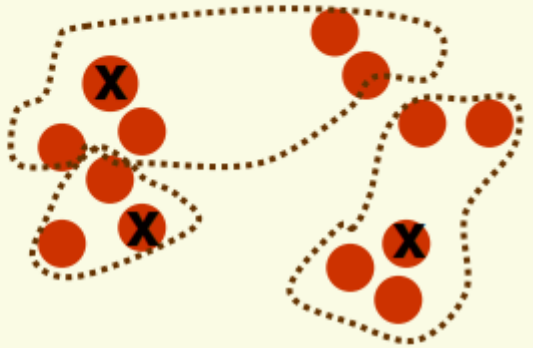
$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

که $\| \|$ معیار فاصلهٔ بین نقاط و c_j مرکز خوشهٔ jام است.

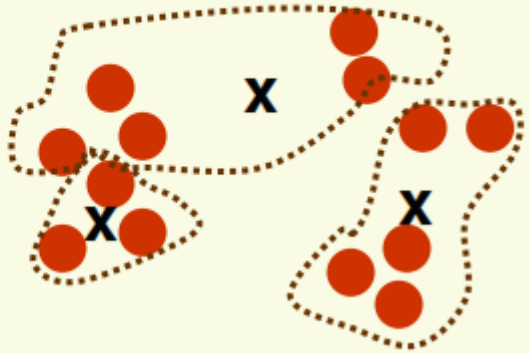
K-means Clustering

k = 3

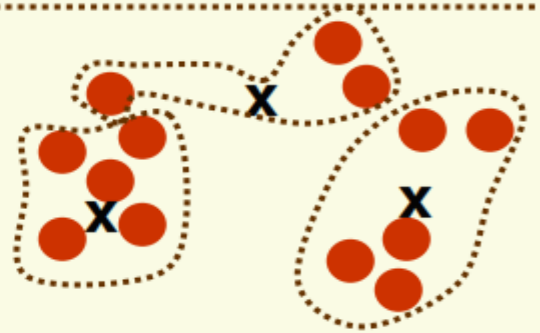
1. Initialize
   - pick ***k*** cluster centers arbitrary
   - assign each example to closest center



2. compute sample means for each cluster



3. reassign all samples to the closest mean



4. if clusters changed at step 3, go to step 2