خوشه‌بندی
Clustering

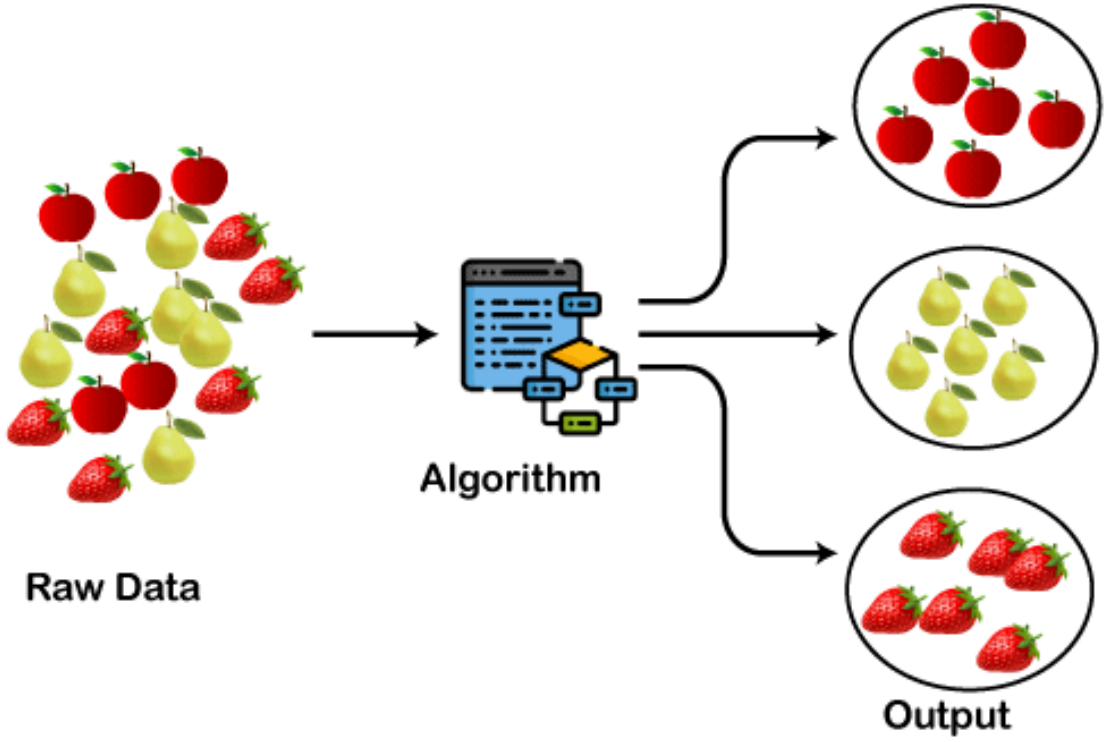
dayche.com . گروه دایکه

دای‌کِه DAYCHE

مقدمه

- یادگیری بدون نظارت – Unsupervised learning

کشف الگوهای ذاتی داده‌ها بر اساس ویژگی‌های مشترک

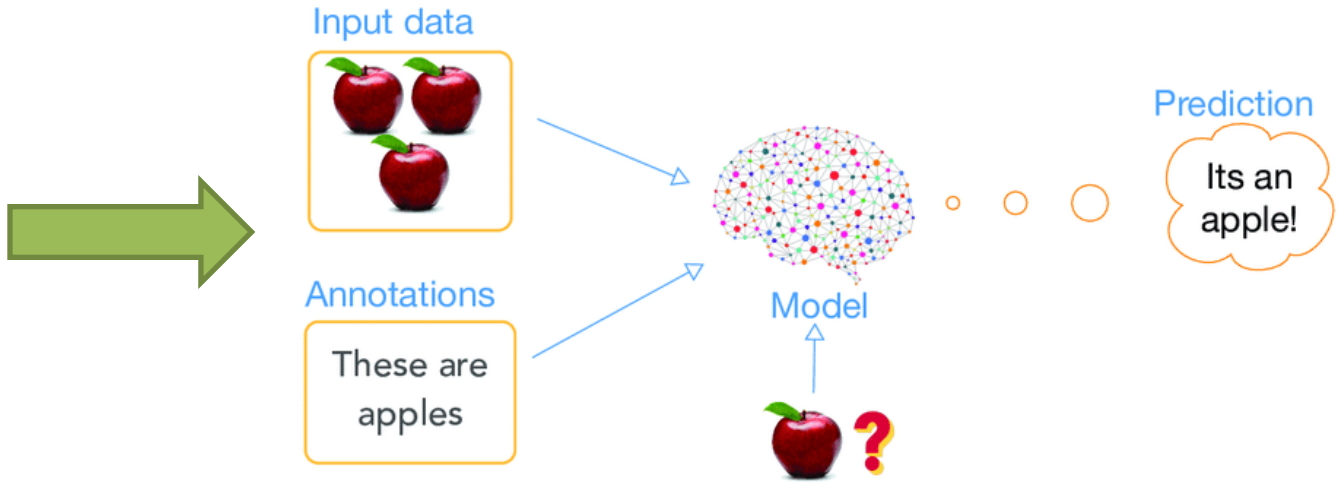


تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

مقدمه – تفاوت یادگیری با نظارت و بدون نظارت

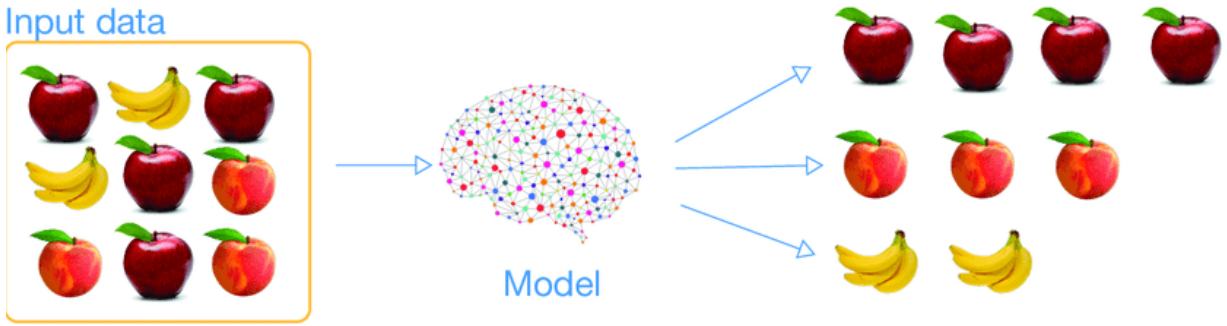
داده‌های برچسب‌گذاری شده

$\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$



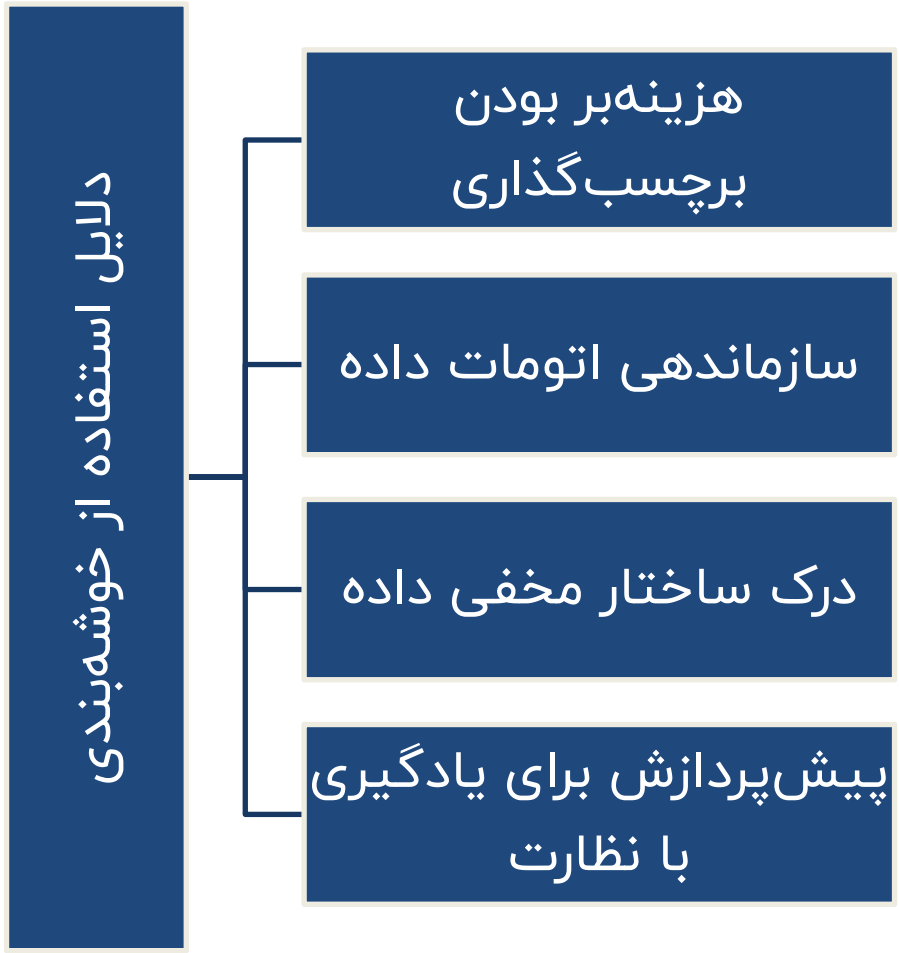
داده‌های بدون برچسب

$\{x_1, x_2, x_3, \dots, x_n\}$



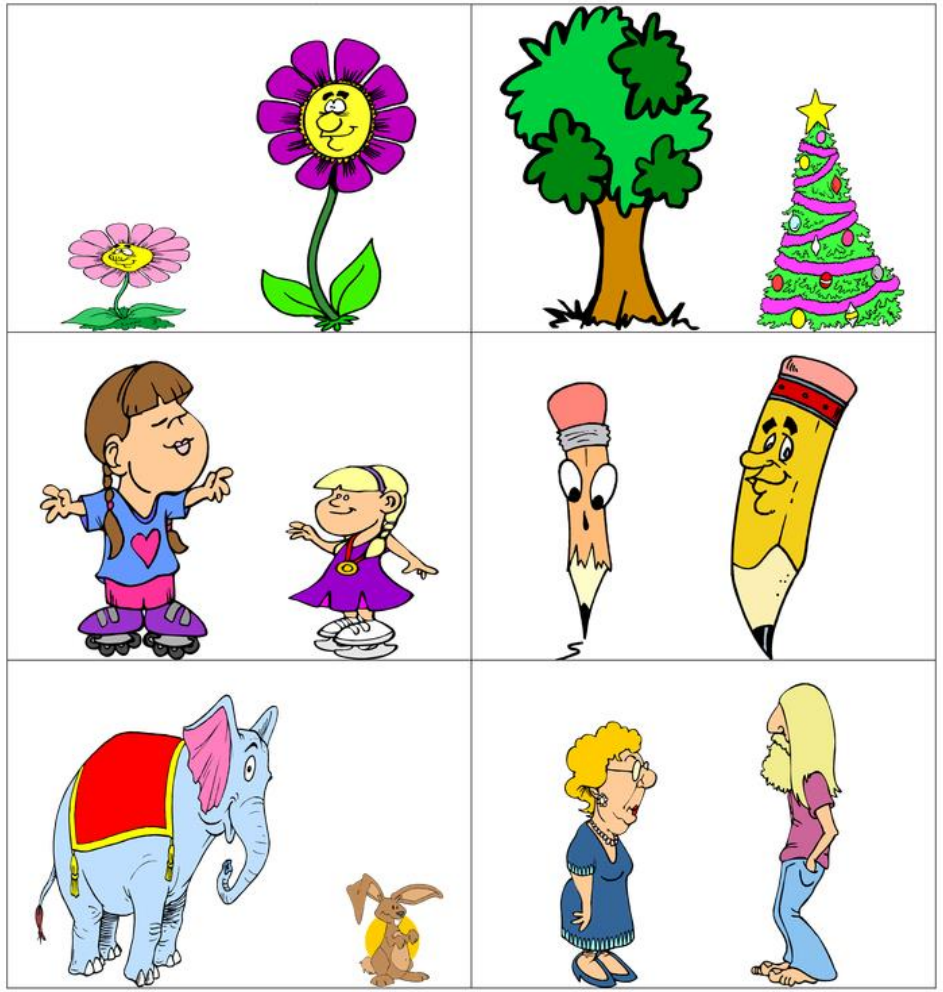
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

انگیزه‌های استفاده از یادگیری بدون نظارت



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

تعریف خوشه‌بندی



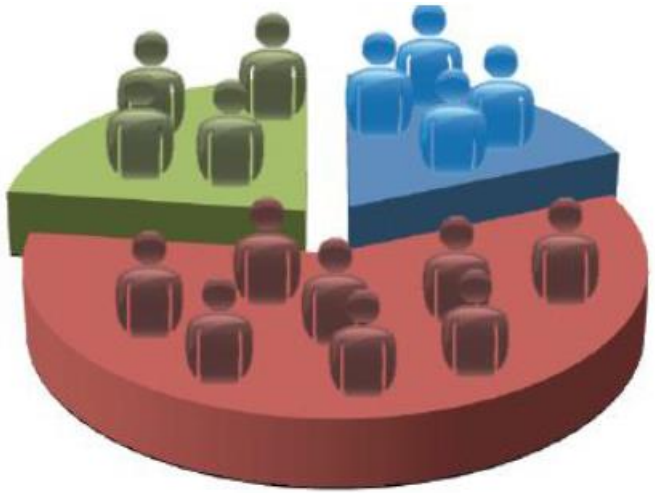
- تعریف
- پیدا کردن کلاس‌ها داده به طوری که:
  - شباهت درون کلاسی بیشینه باشد.
  - شباهت بین کلاسی کمینه باشد.

معیارهای متفاوتی برای شباهت وجود دارد؟
آیا دسته‌بندی دیگر برای این اشیا وجود دارد؟

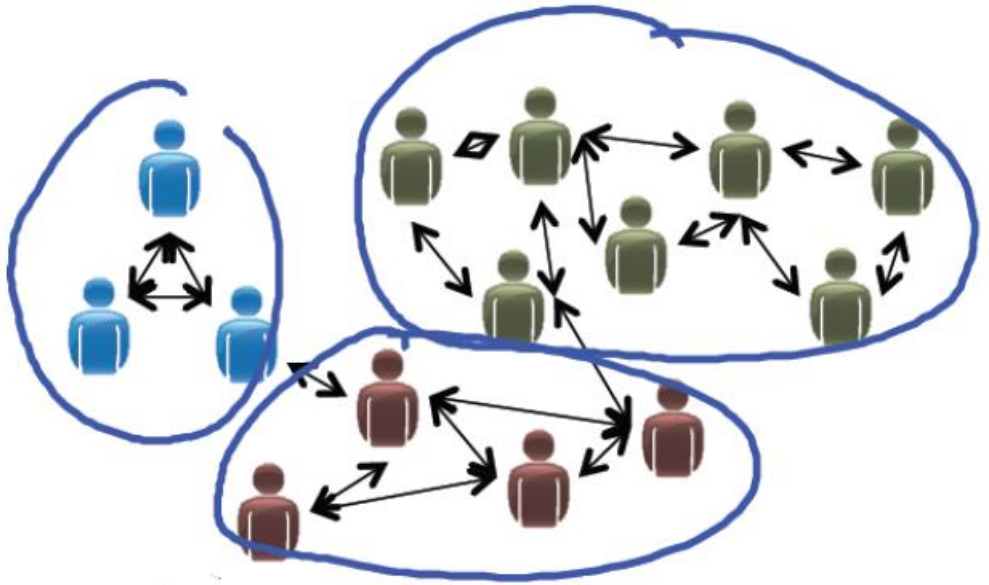
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

کاربردها

Market segmentation



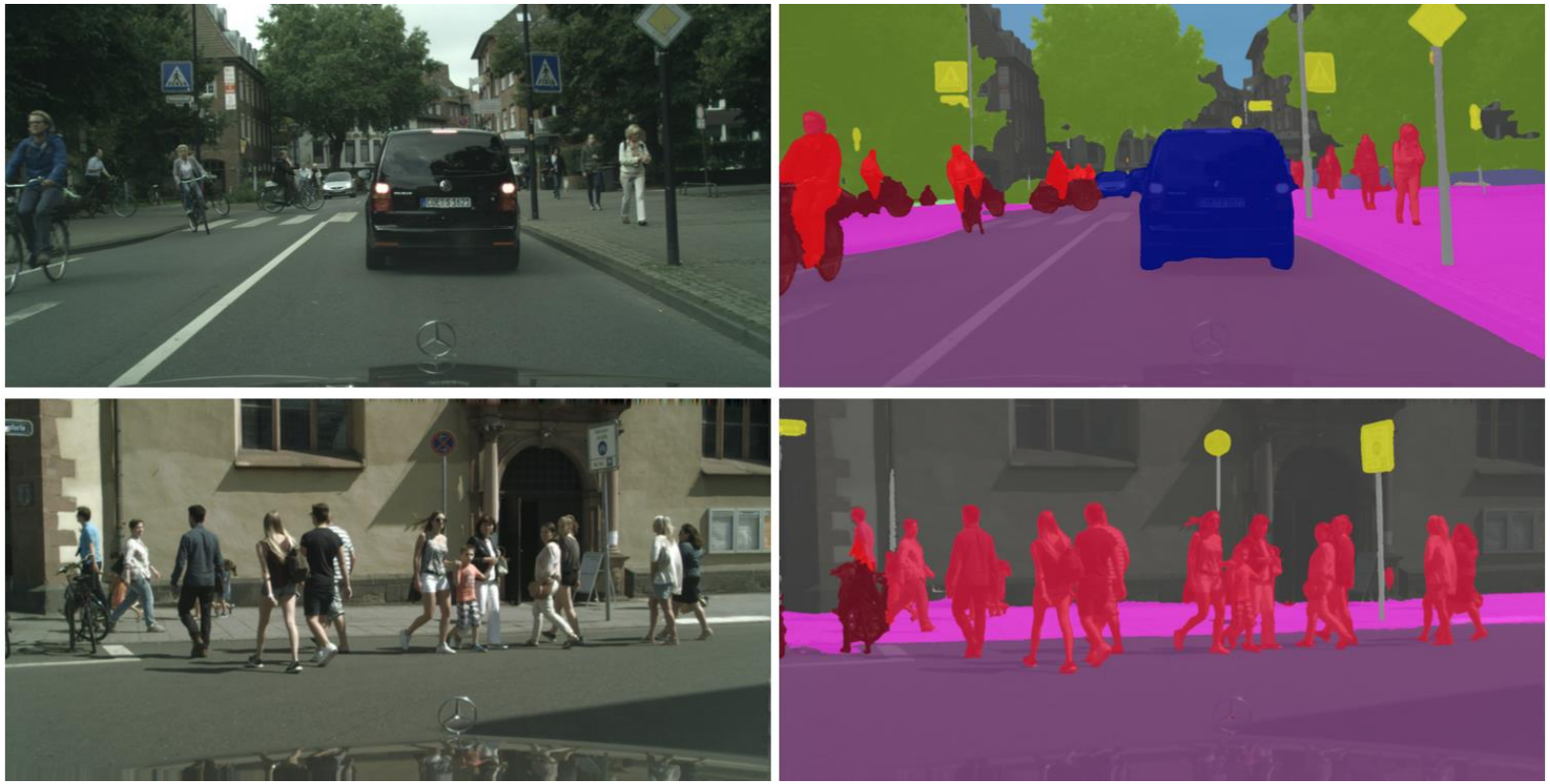
Community detection



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

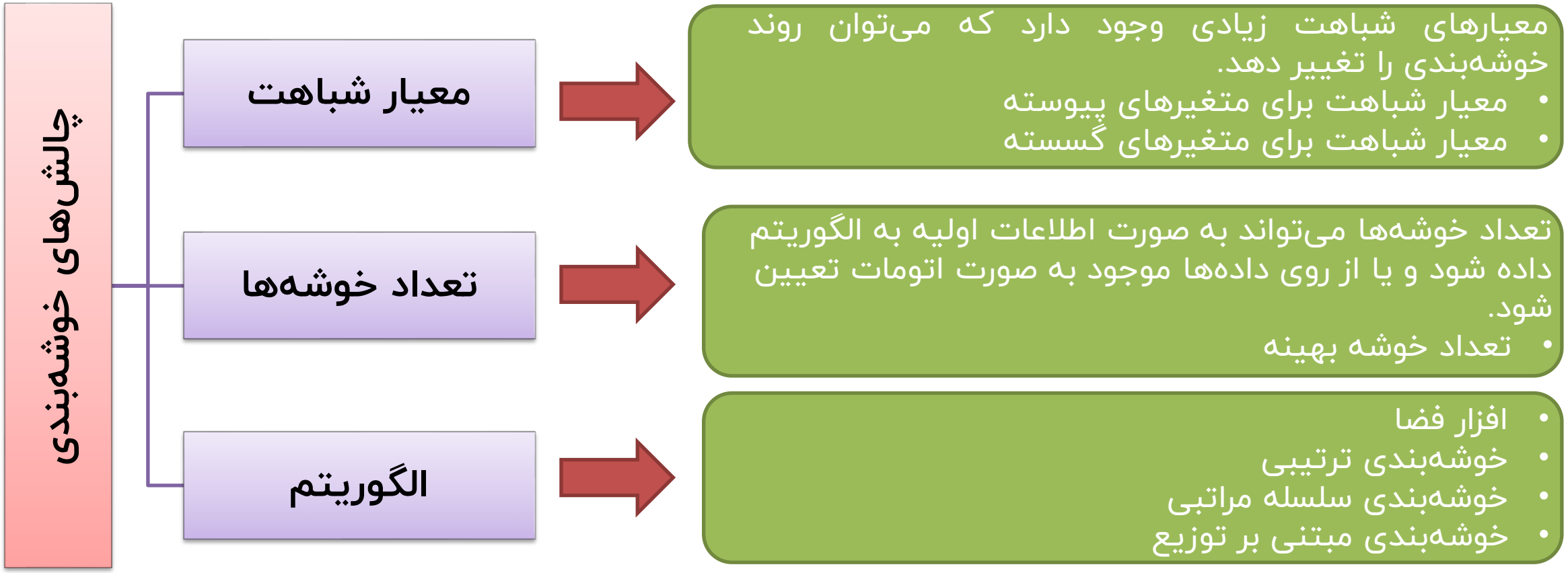
کاربردها

Image Segmentation



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

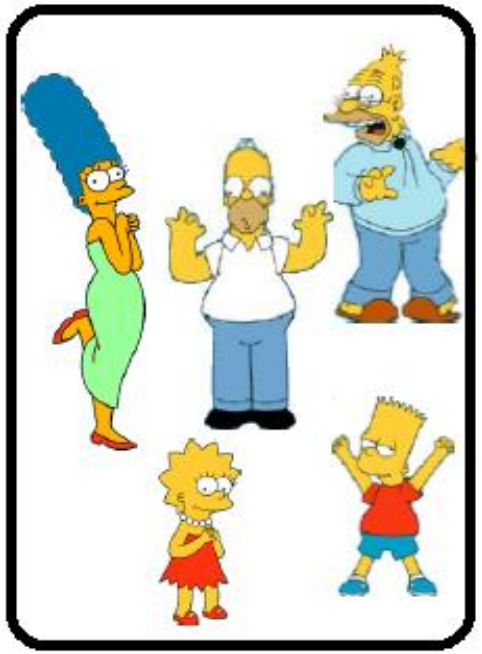
چالش‌های خوشه‌بندی



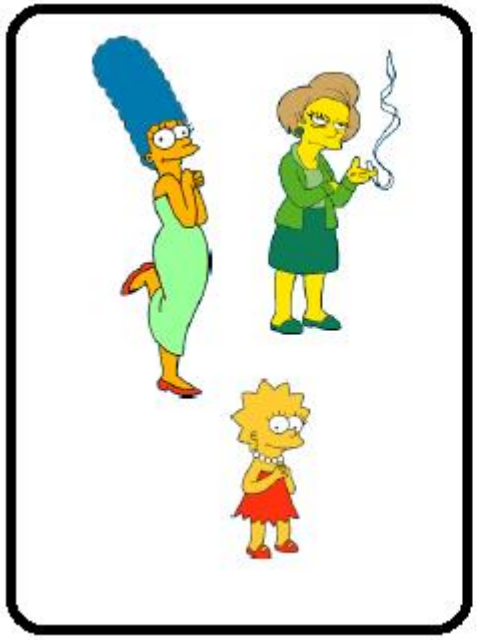
تولید محتوا: وحید محمدزاده ایوقی

معیارهای شباهت

کدام خوشه‌بندی درست است؟



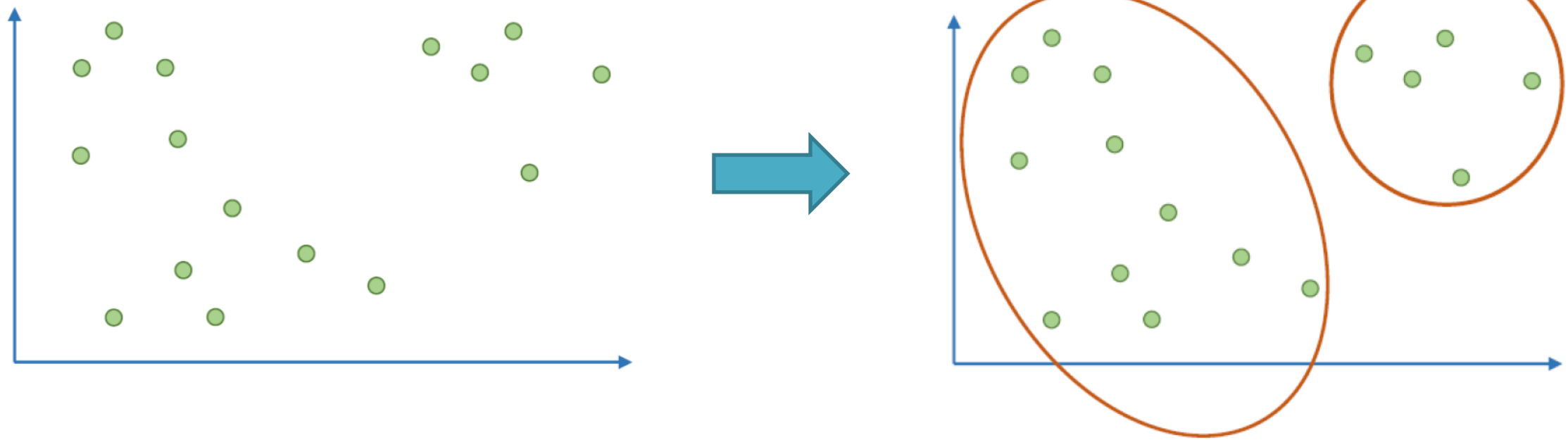
Based on social role



Based on the sex

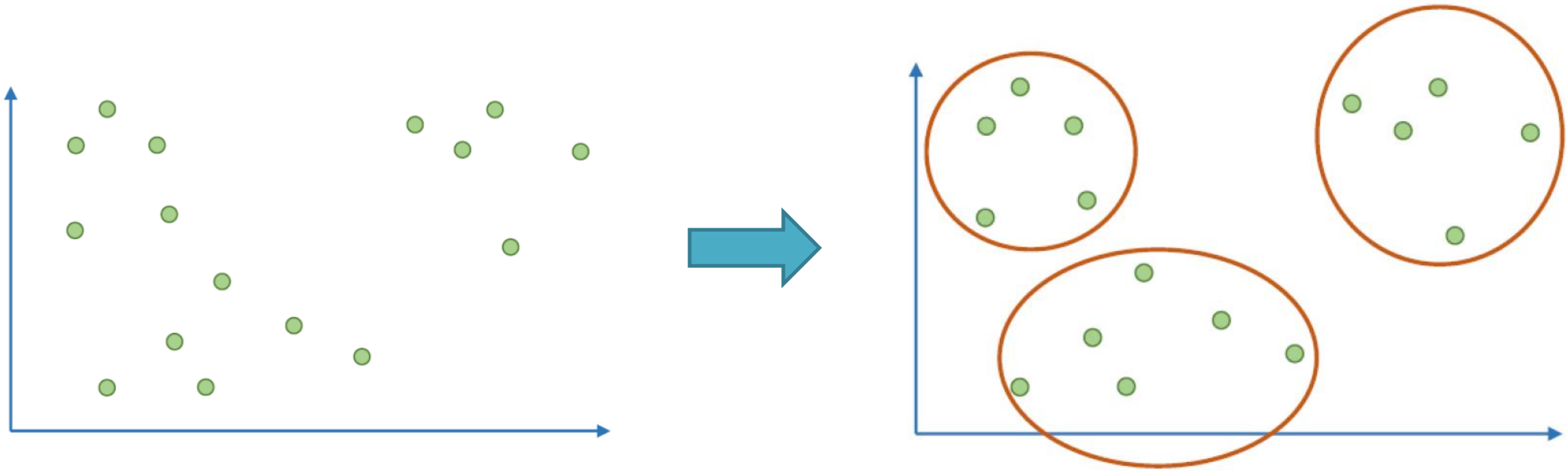
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

معیارهای شباهت



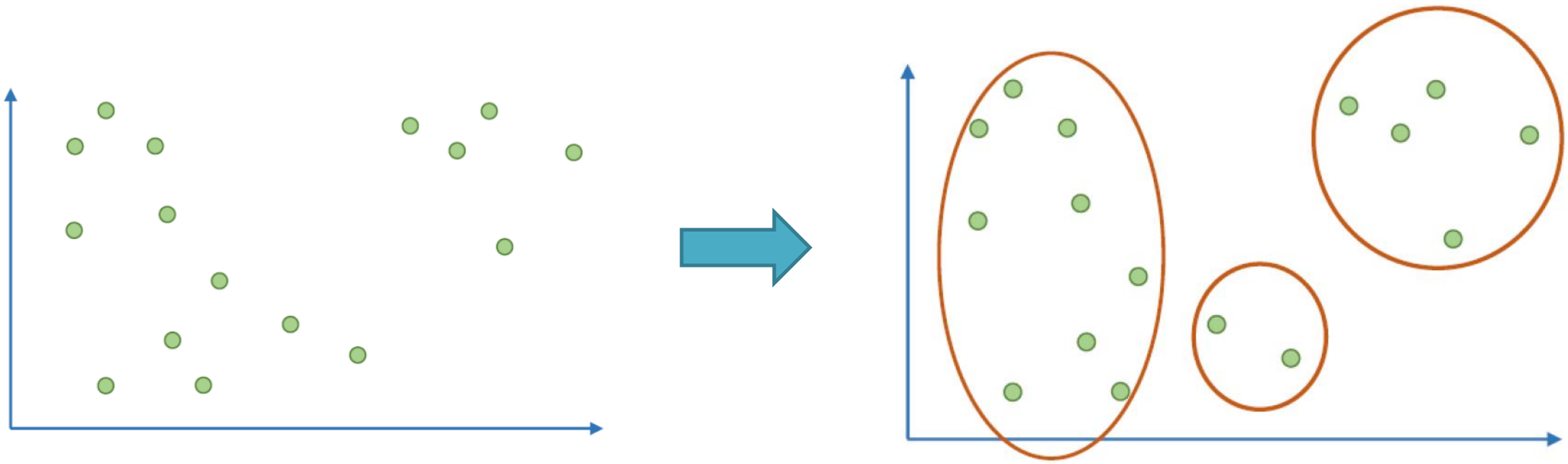
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

معیارهای شباهت



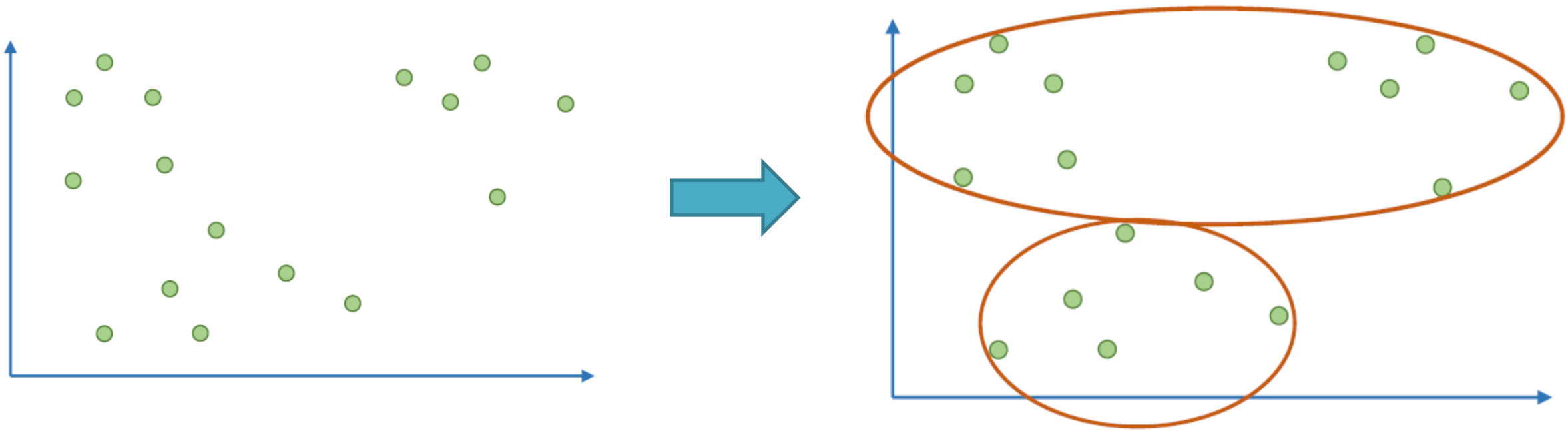
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

معیارهای شباهت



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

معیارهای شباهت



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

معیارهای شباهت

- شباهت بر اساس موقعیت هندسی
  - بر اساس فاصله
    - فضای پیوسته و گسسته
    - کاربرد در تشخیص نابهنجاری، سگمنت کردن تصویر و ...

هر متری که این 4 شرط را برآورده کند می‌تواند یک معیار فاصله باشد

- $D(A,B) = D(B,A)$
- $D(A,A) = 0$
- $D(A,B) = 0 \rightarrow A = B$
- $D(A,B) \leq D(A,C) + D(C,B)$

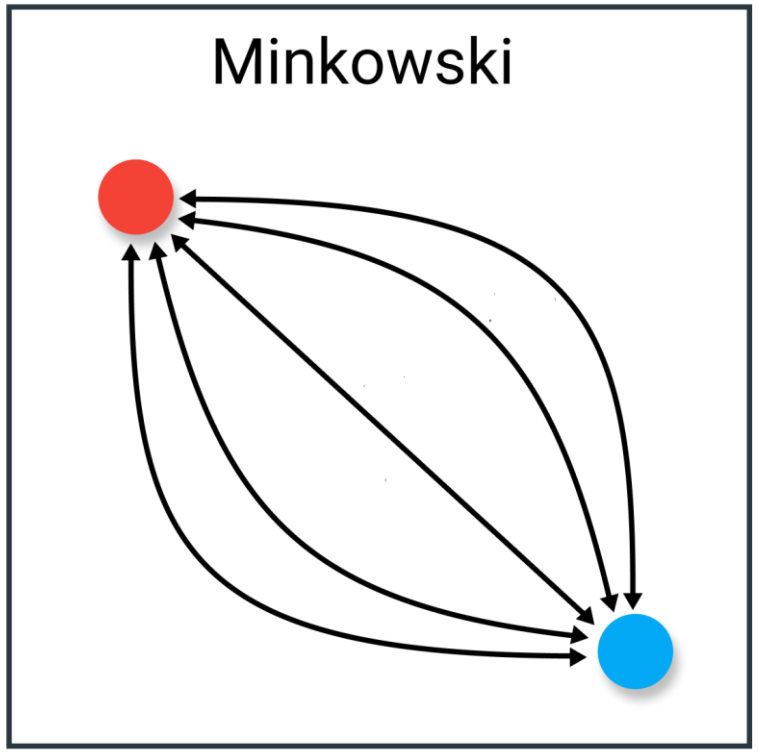
  - بر اساس زاویه
    - کاربرد در پردازش زبان‌های طبیعی

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

معیارهای فاصله

- فاصله Minskowski

$$d(x,y) = \left(\sum_{i=1}^{n} w_i |x_i - y_i|^p \right)^{\frac{1}{p}}$$



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

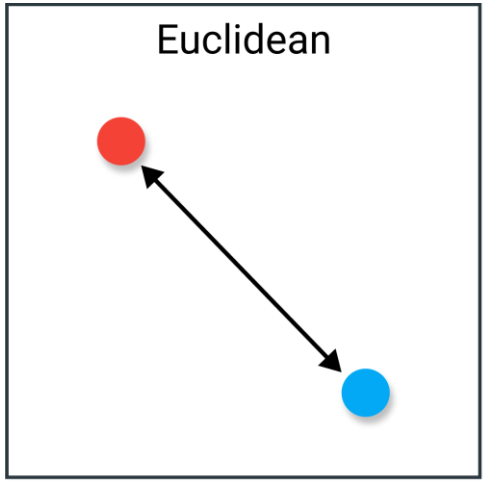
معیارهای فاصله

- فاصله Minskowski

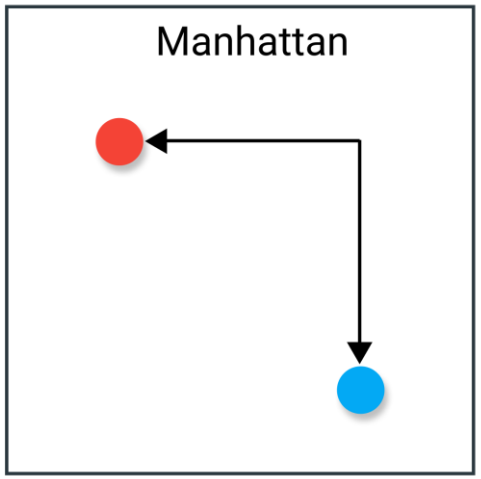
$$d(x,y) = \left(\sum_{i=1}^{n} w_i |x_i - y_i|^p\right)^{\frac{1}{p}}$$

- فاصله اقلیدسی
- فاصله منهتن
- فاصله چبی‌شف

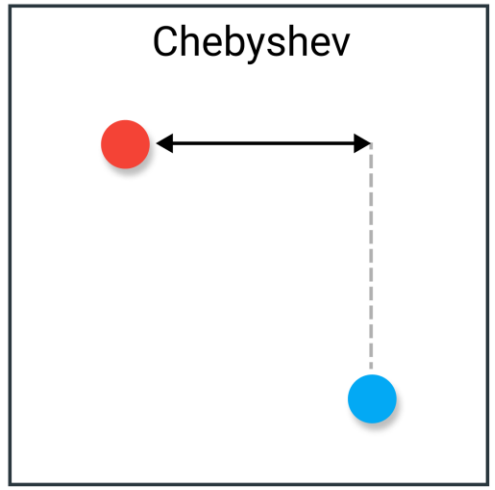
$$d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^2\right)^{\frac{1}{2}}$$



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



$$d(x,y) = \max |x_i - y_i|$$



تولید محتوا: وحید محمدزاده ایوقی

معیارهای شباهت

- شباهت
  - عکس فاصله – در فضای متریک
  - ضرب داخلی
    - معمولا دو بردار را نرمالیزه می‌کنیم تا طول هر بردار واحد باشد.
  - Cosine similarity

$$x^T y = \sum_{i=1}^{n} x_i y_i$$

$$\cos\theta = \frac{a.b}{|a||b|}$$

- زاویه بین دو بردار می‌تواند معیاری برای فاصله دو بردار باشد
  - این فاصله در مقیاس مثلثاتی تعریف می‌شود و نه متریک.

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

معیارهای شباهت

- معیار Tanimoto

$$S(x,y) = \frac{x^T y}{|x|^2 + |y|^2 - x^T y}$$

- معیارهای شباهت برای مقادیر گسسته
  - برای متغیرهای گسسته، معیارهای فاصله در فضای متریک قابل تعریف نیست.
- ماتریس شباهت

$A(x,y) = [a_{ij}]$ ⟵ تعداد محل‌هایی که المان iام بردار اول متناظر با المان jام بردار دوم است.

معیارهای شباهت

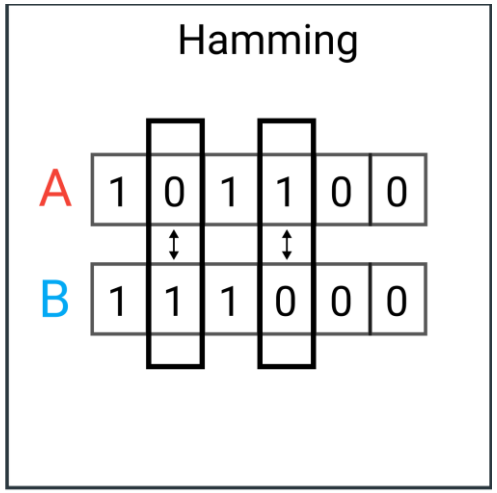
- ماتریس شباهت

$x = [0, 1, 2, 1, 2, 1]^T, y = [1, 0, 2, 1, 0, 1]^T$

تعداد سطوح مختلف = 3 ⟹ ماتریس شباهت یک ماتریس 3 در 3 است

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

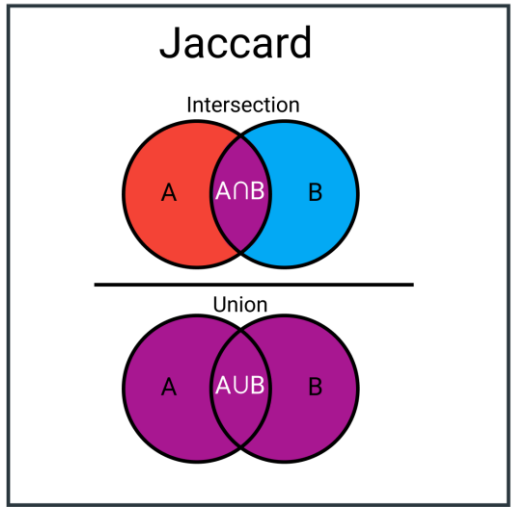
⟹ معیارهای شباهت بر اساس ترکیبی از درایه‌های این ماتریس تعریف می‌شوند

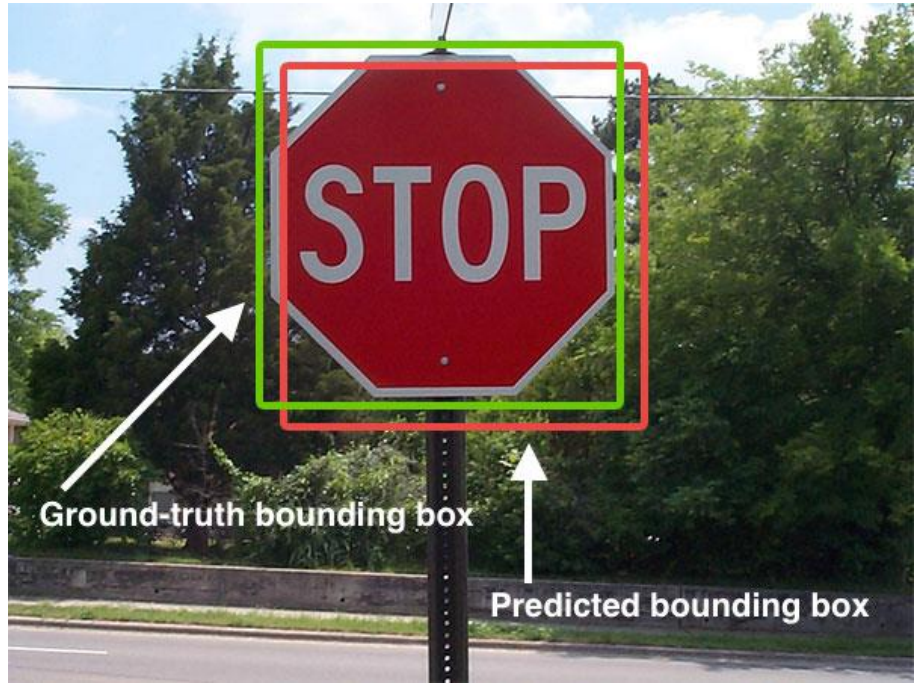


- فاصله Hamming
  - تعداد محل‌های که دو بردار متفاوت هستند
  - جمع عناصر غیرقطری ماتریس شباهت

معیارهای شباهت

- فاصله Jaccard





Template matching

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

معیارهای شباهت

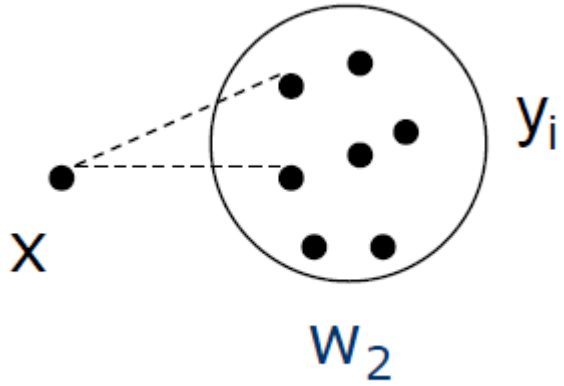
- شباهت بر اساس مشخصات آماری
  - آزمون‌های فرض آماری
  - شاخص‌های آماری – آنتروپی، کشیدگی، ...

مناسب برای سنجش فاصله بین دو بردار با طول متفاوت

- خوشه‌بندی
  - غالبا از روش‌های مبتنی بر تعریف فاصله در فضای متریک استفاده می‌کنند.
  - برای مقادیر گسسته نیاز داریم تا نگاشت پیوسته این مقادیر را تعریف کنیم.

محاسبه فاصله

- فاصله نقطه از نقطه
- فاصله نقطه از مجموعه
  - رویکرد اول



میانگین، میانه،ماکزیمم، مینیمم، ...

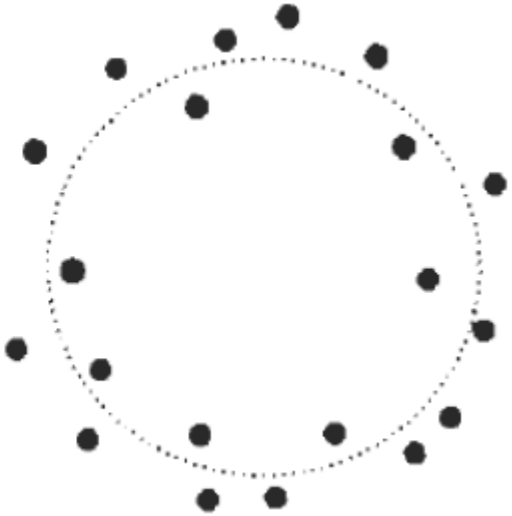
$$d(x, w_2) = T(\{d(x, y_i), \forall y_i \in w_2)$$

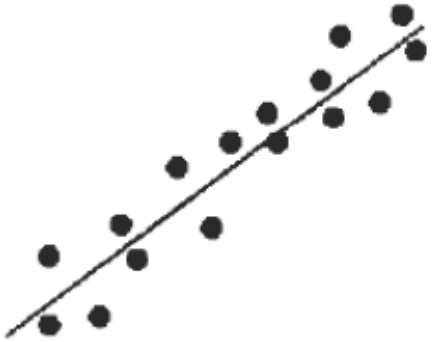
  - رویکرد دوم – تعریف یک prototype برای مجموعه

$$d(x, w_2) = d(x, C)$$

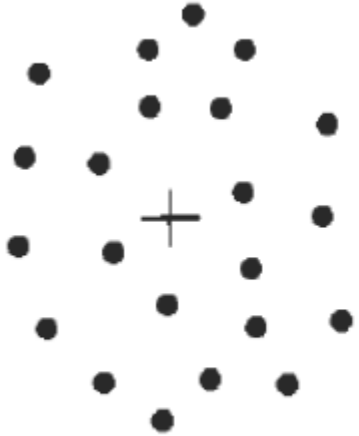
محاسبه فاصله

- انواع Prototype
  - (خط) ابرصفحه، ابرکره، نقطه





میانگین، میانه

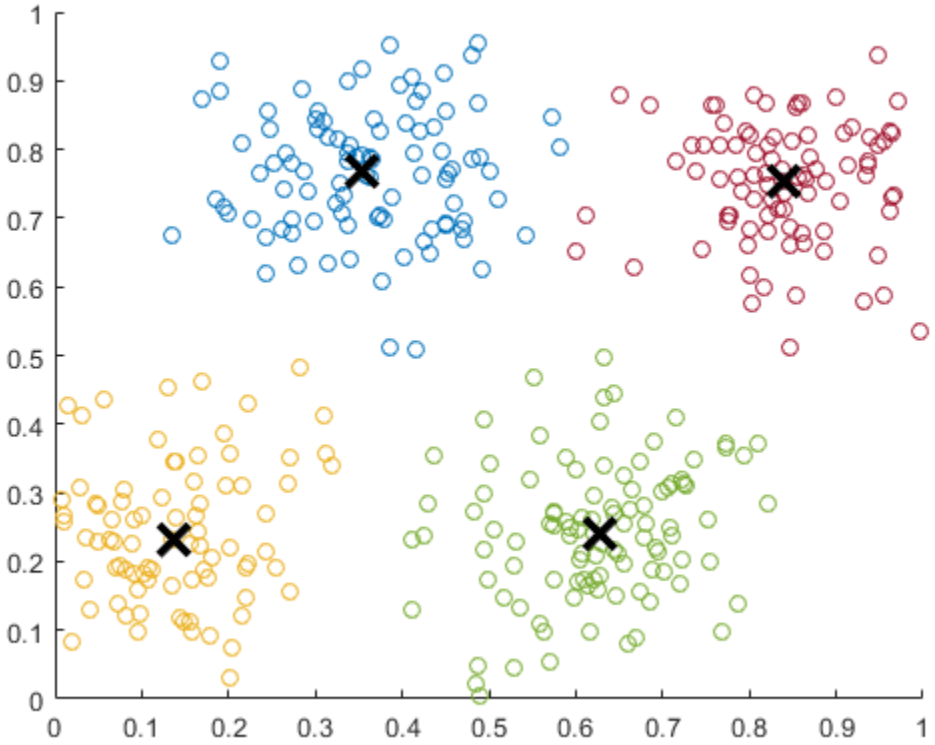


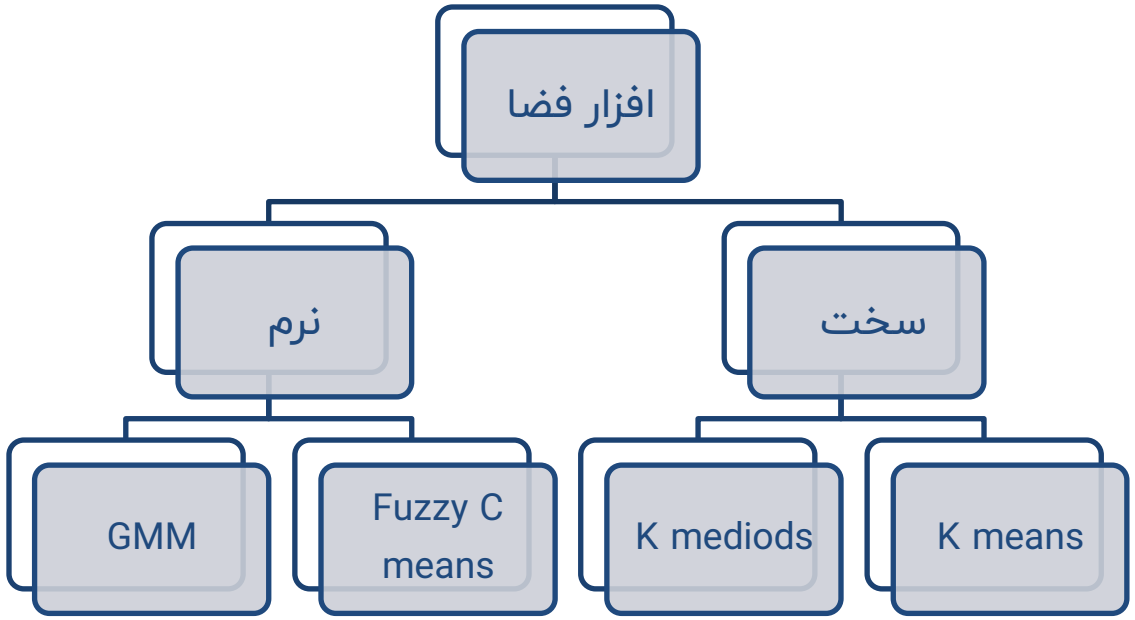
  - تعبیر شباهت – تعبیر فاصله از prototype

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

خوشه‌بندی

- روش افراز فضا
  - تعداد کلاسترها را به صورت پیش‌فرض نیاز دارد.





تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

افراز فضا

- روش K means
  - تعداد خوشه‌ها را به صورت پیش‌فرض نیاز دارد.
  - تعداد بهینه خوشه

داده‌های بدون برچسب به صورت زیر در دست است:

$\{x_1, x_2, x_3, \dots, x_n\}$

Prototypeهای هر خوشه را طوری بیابید که:

- فاصله هر داده از Prototype دسته خود حداقل باشد.
- فاصله هر داده از Prototype دسته‌های دیگر بیشینه باشد.

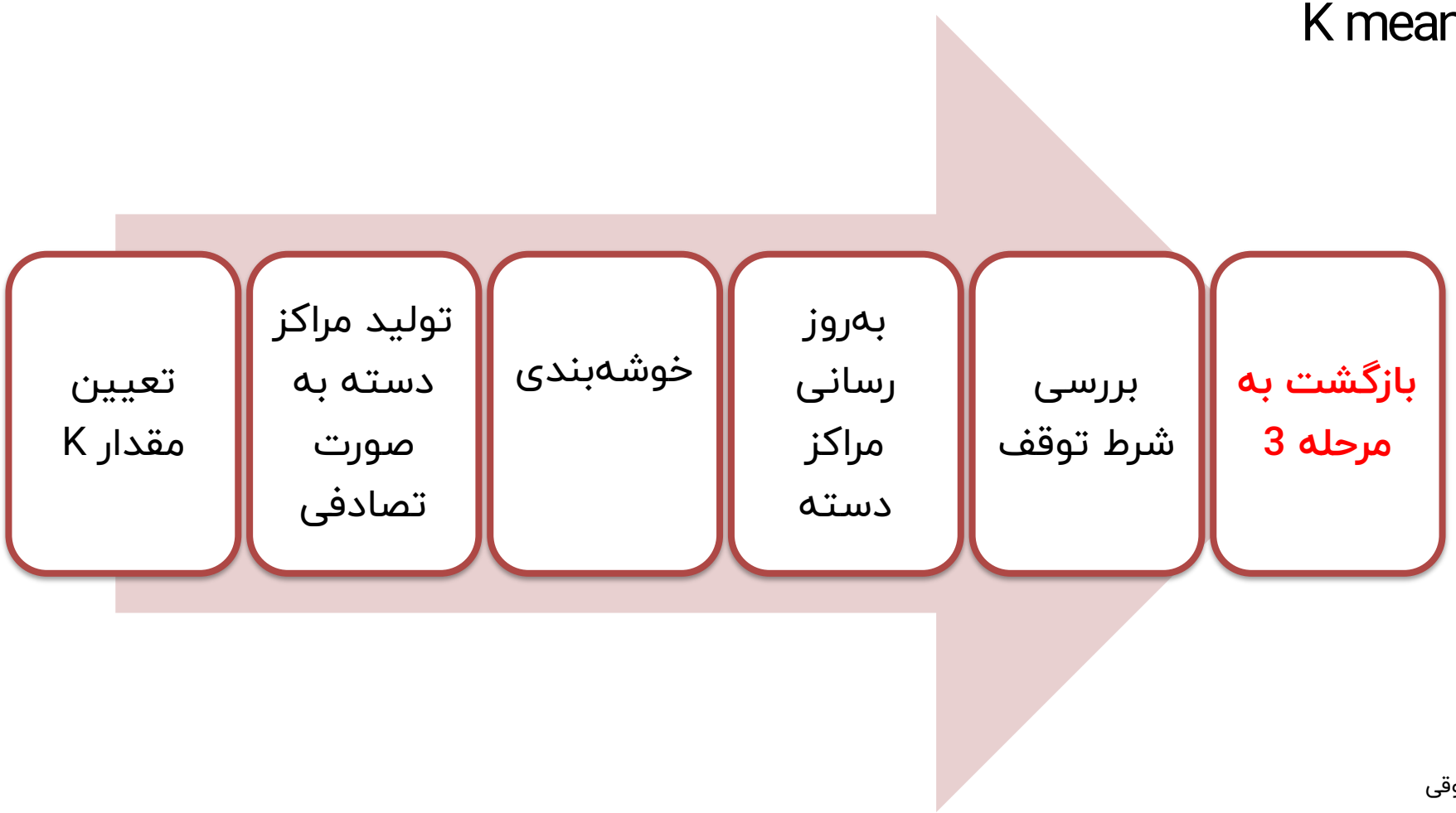
$$c_i^* = \arg\min |x - c_i|^2$$

$$c_i^* = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- فاصله بین دو کلاستر چرا لحاظ نشده است؟

افراز فضا

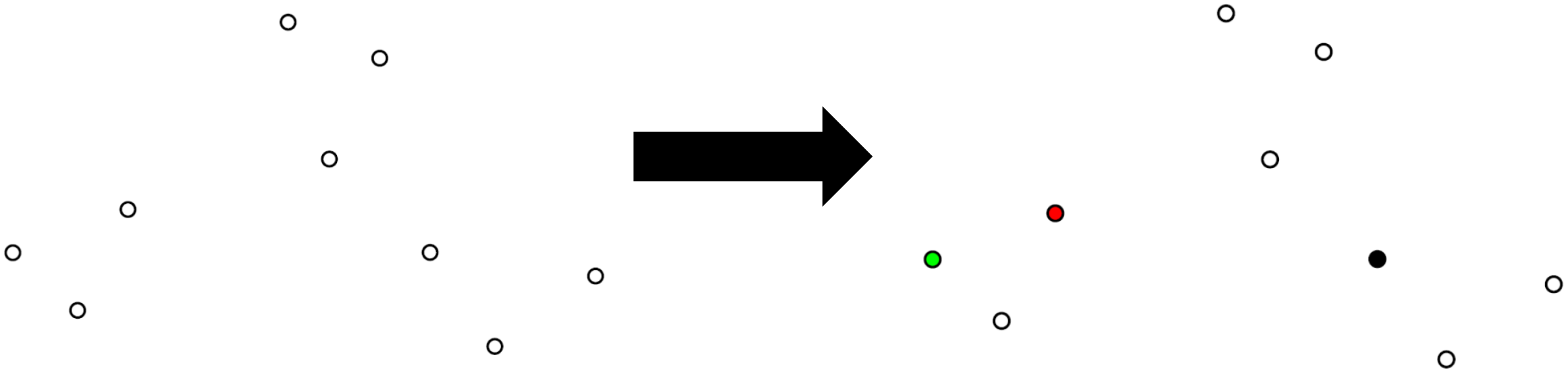
- روش K means



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

افراز فضا

- روش K means

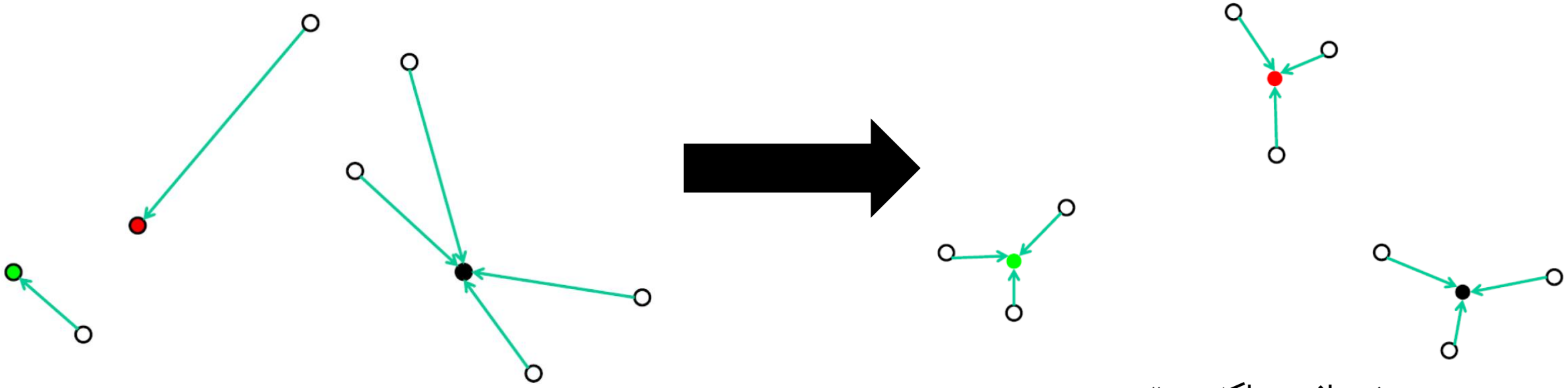


انتخاب K مرکز تصادفی

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

افراز فضا

- روش K means



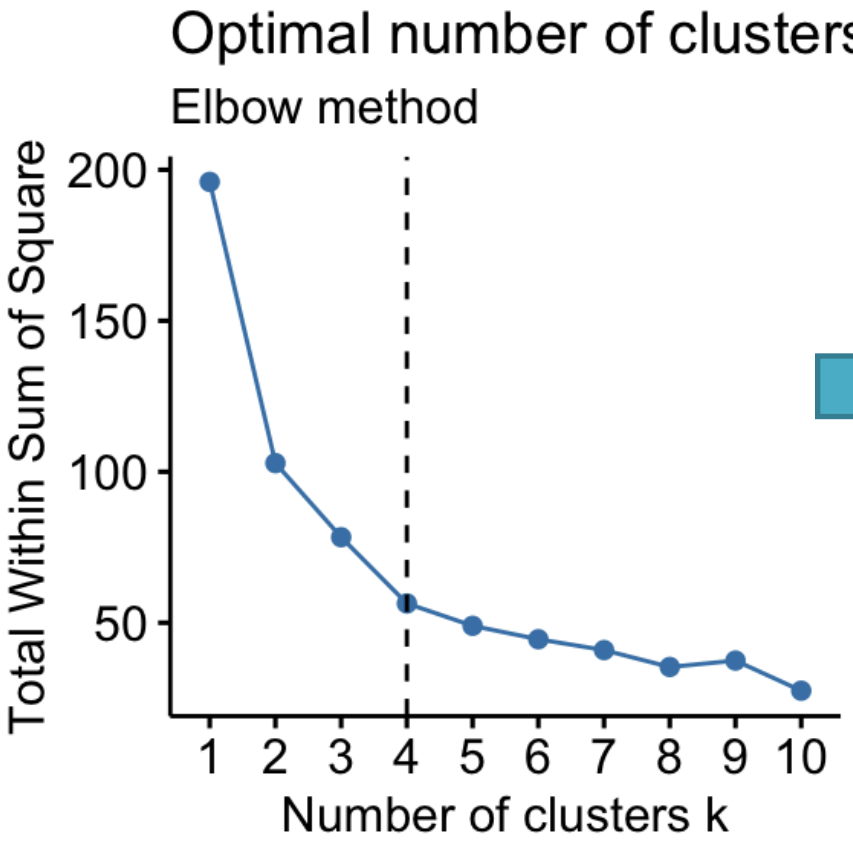
خوشه‌بندی داده‌ها

به روزرسانی مراکز دسته

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

افراز فضا

- تعداد بهینه خوشه
- تکرار الگوریتم به ازای خوشه‌های مختلف



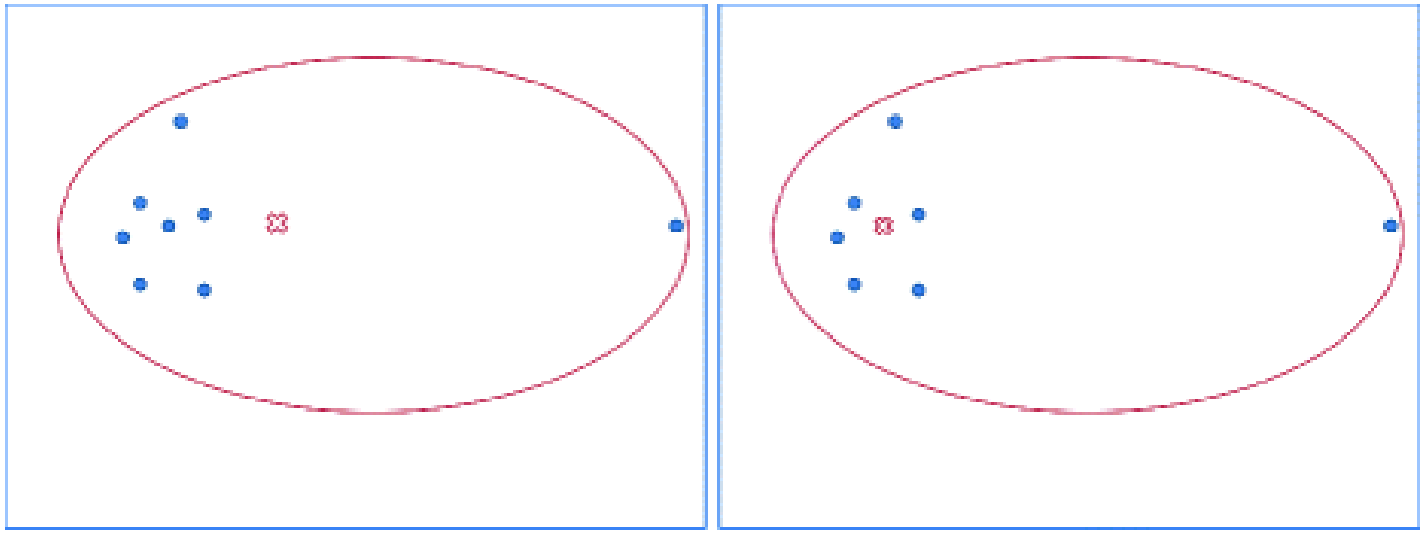
تعداد بهینه خوشه
- پیچیده‌تر شدن مدل تاثیری بر روی تابع هزینه نداشته باشد.

ایراد
- زمان‌بر بودن پروسه بهینه‌سازی

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

افراز فضا

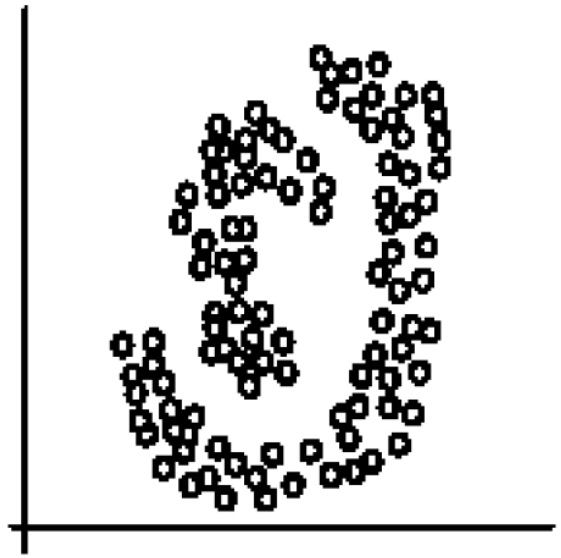
- ایرادات روش K means
- حساس به داده‌ها پرت است – میانگین به داده‌های پرت حساس است
- روش K mediods

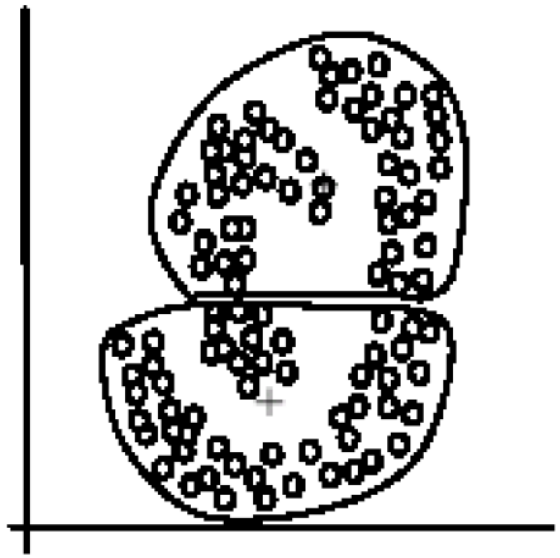


تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

افراز فضا

- ایرادات روش K means
- حساس به شکل پراکندگی داده‌ها در فضا

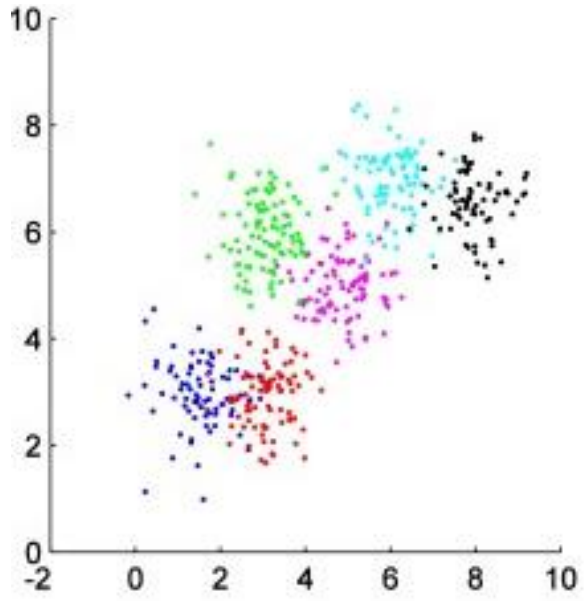




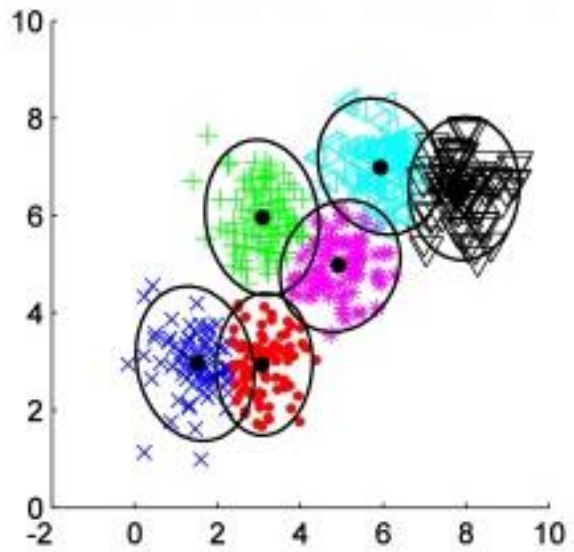
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

افراز فضا

- ایرادات روش K means
  - هر داده مطلقا به یک کلاستر تعلق دارد – خوشه‌بندی سخت
    - خوشه‌بندی نرم – خوشه‌بندی فازی



(a)

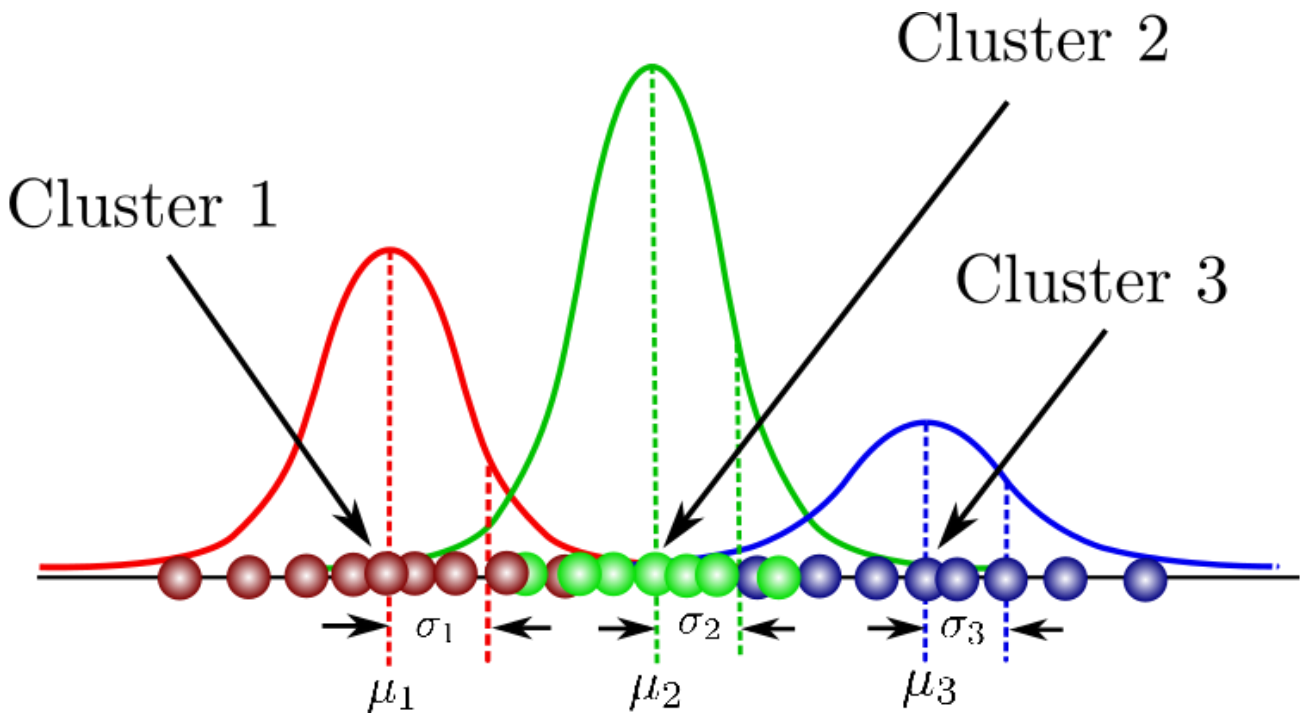


(b)

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

افراز فضا

- ایرادات روش K means
- خوشه‌بندی نرم – خوشه‌بندی مبتنی بر مدل مخلوط گوسی



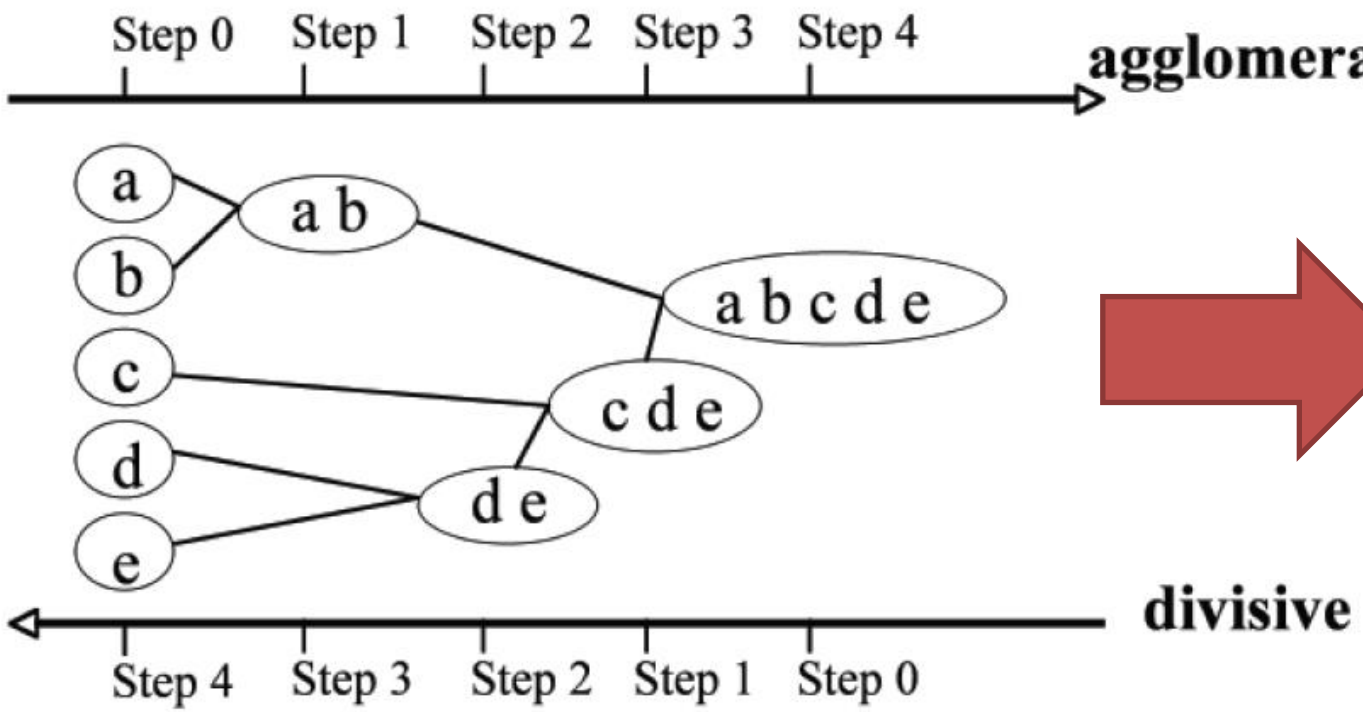
در اینجا محاسبه شباهت بر اساس فاصله در فضای متریک نیست.

- شباهت بر اساس احتمال رخداد

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

خوشه‌بندی سلسله مراتبی

- روش افزار فضا
  - تعداد خوشه‌ها باید از قبل مشخص باشد – اگر مشخص نبود، بر اساس روش Elbow محاسبه می‌شود.
- خوشه‌بندی سلسله مراتبی



- عدم نیاز به مشخص بودن تعداد کلاسترها
- مطابق با درک انسان

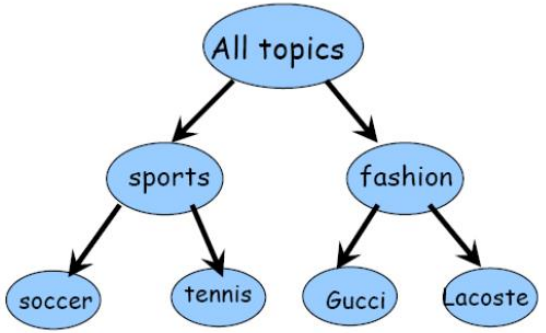
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

خوشه‌بندی سلسله مراتبی

- خوشه‌بندی سلسله مراتبی

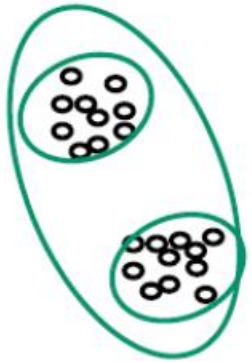
Agglomerative (Bottom-up)

قرار دادن هر داده در یک کلاستر → ترکیب کردن کلاسترها → کاهش کلاسترها



Divisive (Up-down)

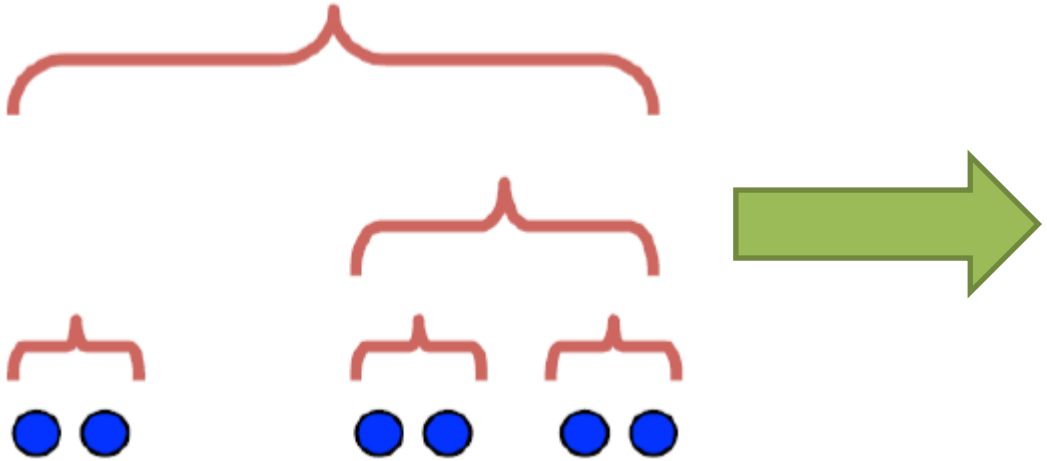
قرار دادن کل داده‌ها در یک کلاستر → تقسیم کردن → افزایش کلاسترها



تولید محتوا: وحید محمدزاده ایوقی

خوشه‌بندی سلسله مراتبی

- خوشه‌بندی سلسله مراتبی
  - Agglomerative



الگوریتم خوشه‌بندی سلسله مراتبی
- هر داده در یک خوشه
- ترکیب خوشه‌های نزدیک بهم
- ادامه ترکیب تا جایی که یک خوشه کلی حاصل شود

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

خوشه‌بندی سلسله مراتبی

- خوشه‌بندی سلسله مراتبی – Agglomerative

الگوریتم

- مقدار دهی اولیه – هر داده در یک خوشه $R = \{C_i\}, i = 1, \dots, N$
- یافتن نزدیک‌ترین کلاسترها $(i^*, j^*) = \arg\min d(C_i, C_j)$
- ادغام کلاسترها $C_q = C_{i^*} \cup C_{j^*}$
- به روزرسانی خوشه‌های اولیه $R^+ == \left\{R^- - \left\{C_{i^*}, C_{j^*}\right\}\right\} \cup C_q$
- برگشت به گام دوم

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

خوشه‌بندی سلسله مراتبی

- خوشه‌بندی سلسله مراتبی – Agglomerative
  - ایراد اساسی – اگر یک عضو به اشتباه دسته‌بندی شود دیگر امکان خارج از دسته خود را ندارد.
  - هزینه محاسباتی – محاسبه فاصله

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

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

- دیگر نیازی به محاسبه مجدد فاصله‌ها نیست! همه در ماتریس نزدیکی موجود است.

خوشه‌بندی سلسله مراتبی

- خوشه‌بندی سلسله مراتبی – Agglomerative
- اگر دو داده، و یا دو کلاس، ادغام شدند، وضعیت ماتریس نزدیکی چگونه خواهد بود؟

$$d(C_q, C_s) = a_i d(C_i, C_s) + a_j d(C_j, C_s) + b d(C_i, C_j) + c\left|d(C_i, C_s) - d(C_j, C_s)\right|$$

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

Single link: $d(C_q,C_s) = \min\left(d(C_i,C_s),d(C_j,C_s)\right)$

Complete link: $d(C_q,C_s) = \max\left(d(C_i,C_s),d(C_j,C_s)\right)$

Average link: $d(C_q,C_s) = \text{ave}\left(d(C_i,C_s),d(C_j,C_s)\right)$

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

خوشه‌بندی سلسله مراتبی – Agglomerative

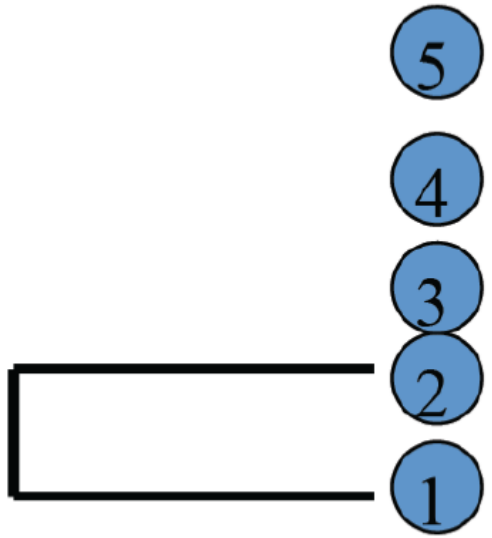
مثال

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 3 & 0 & & \\ 10 & 9 & 7 & 0 & \\ 9 & 8 & 5 & 4 & 0 \end{bmatrix} \end{array} \Rightarrow \begin{array}{c} \\ (1,2) \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccccc} (1,2) & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 9 & 7 & 0 & \\ 8 & 5 & 4 & 0 \end{bmatrix} \end{array}$$

$$d_{(1,2),3} = \min\{d_{1,3}, d_{2,3}\} = \min\{6,3\} = 3$$

$$d_{(1,2),4} = \min\{d_{1,4}, d_{2,4}\} = \min\{10,9\} = 9$$

$$d_{(1,2),5} = \min\{d_{1,5}, d_{2,5}\} = \min\{9,8\} = 8$$



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

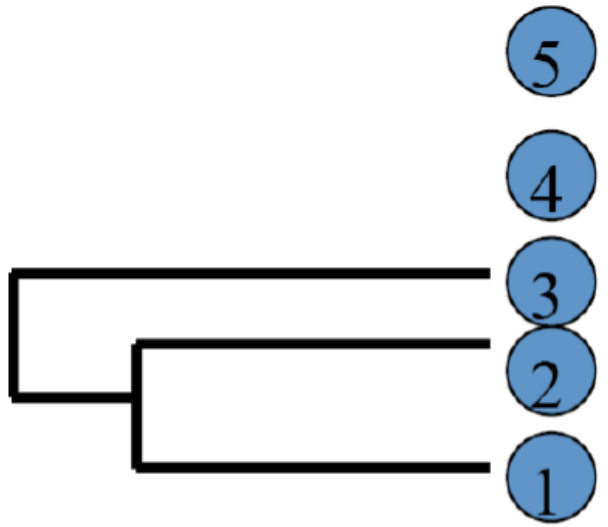
خوشه‌بندی سلسله مراتبی – Agglomerative

مثال

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 3 & 0 & & \\ 10 & 9 & 7 & 0 & \\ 9 & 8 & 5 & 4 & 0 \end{bmatrix} \end{array} \Rightarrow \begin{array}{c} \\ (1,2) \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{cccc} (1,2) & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 9 & 7 & 0 & \\ 8 & 5 & 4 & 0 \end{bmatrix} \end{array} \Rightarrow \begin{array}{c} \\ (1,2,3) \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccc} (1,2,3) & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 5 & 4 & 0 \end{bmatrix} \end{array}$$

$$d_{(1,2,3),4} = \min\{d_{(1,2),4}, d_{3,4}\} = \min\{9,7\} = 7$$

$$d_{(1,2,3),5} = \min\{d_{(1,2),5}, d_{3,5}\} = \min\{8,5\} = 5$$



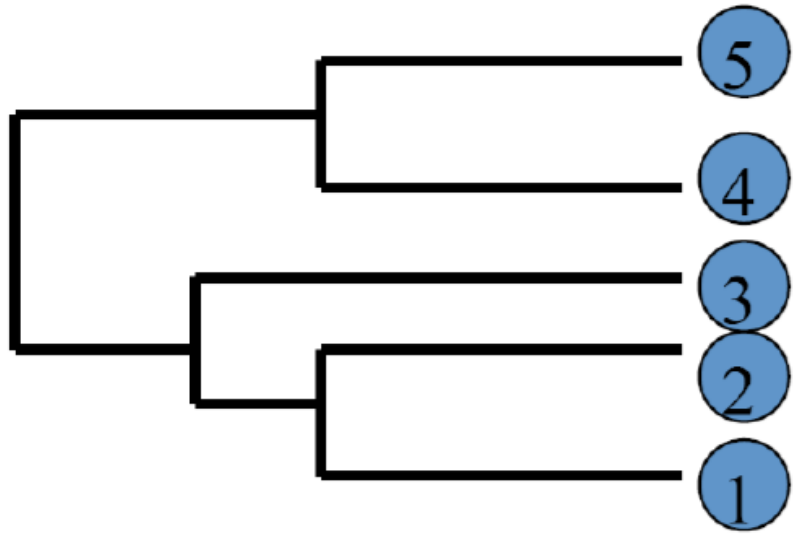
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

خوشه‌بندی سلسله مراتبی – Agglomerative

مثال

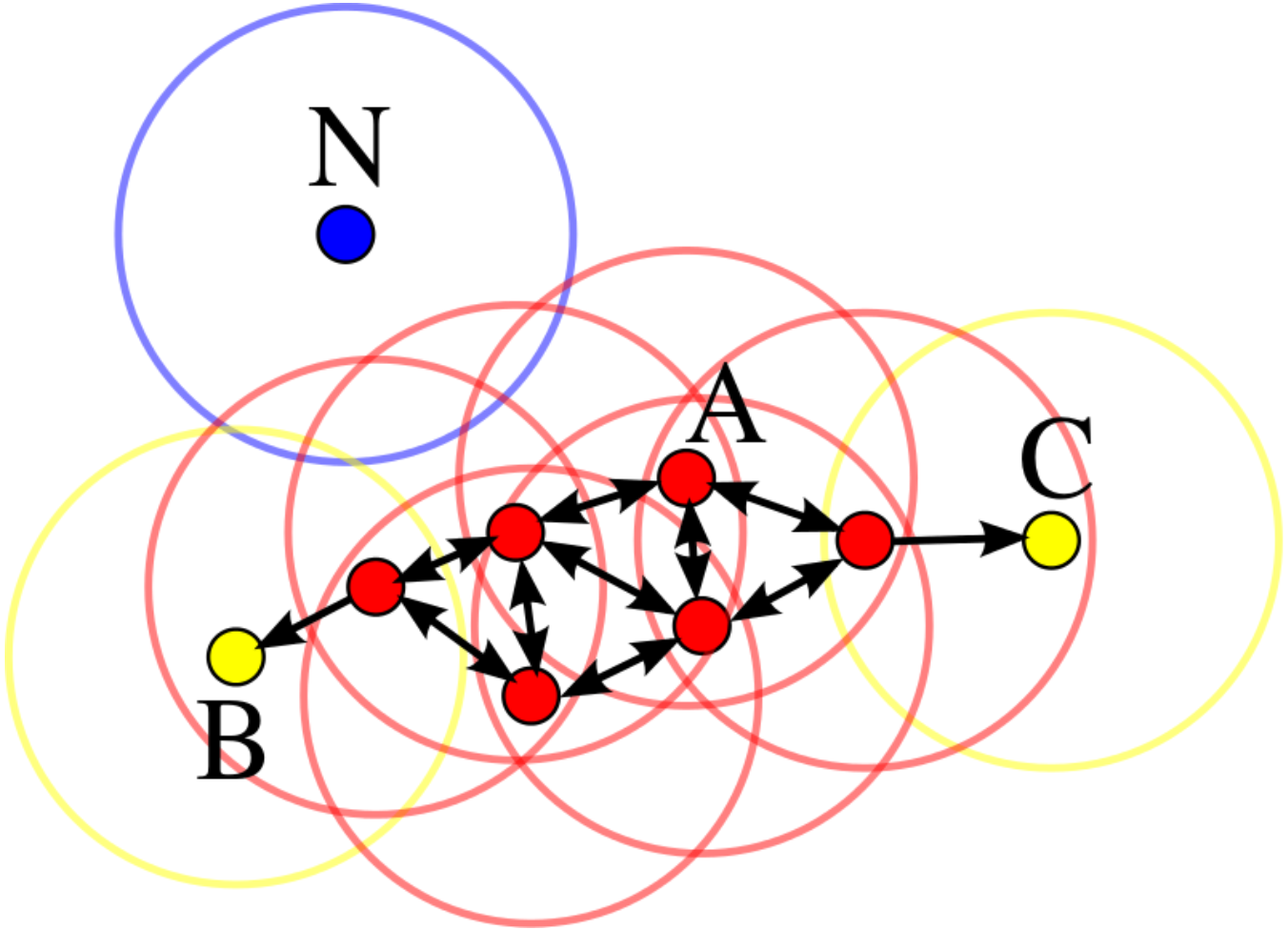
$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 3 & 0 & & \\ 10 & 9 & 7 & 0 & \\ 9 & 8 & 5 & 4 & 0 \end{bmatrix} \end{array} \Rightarrow \begin{array}{c} \\ (1,2) \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{cccc} (1,2) & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 9 & 7 & 0 & \\ 8 & 5 & 4 & 0 \end{bmatrix} \end{array} \Rightarrow \begin{array}{c} \\ (1,2,3) \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccc} (1,2,3) & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 5 & 4 & 0 \end{bmatrix} \end{array}$$

$$d_{(1,2,3),(4,5)} = \min\{d_{(1,2,3),4}, d_{(1,2,3),5}\} = 5$$



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

روش DBSCAN



- خوشه‌ها محلی چگال از فضا هستند که توسط نواحی کم چگال از هم تفکیک شده‌اند
- زمانی که نویز و داده پرت وجود دارد، این الگوریتم‌ها مناسب هستند. (پراکندگی داده‌ها در فضا به صورت فشرده نیست)

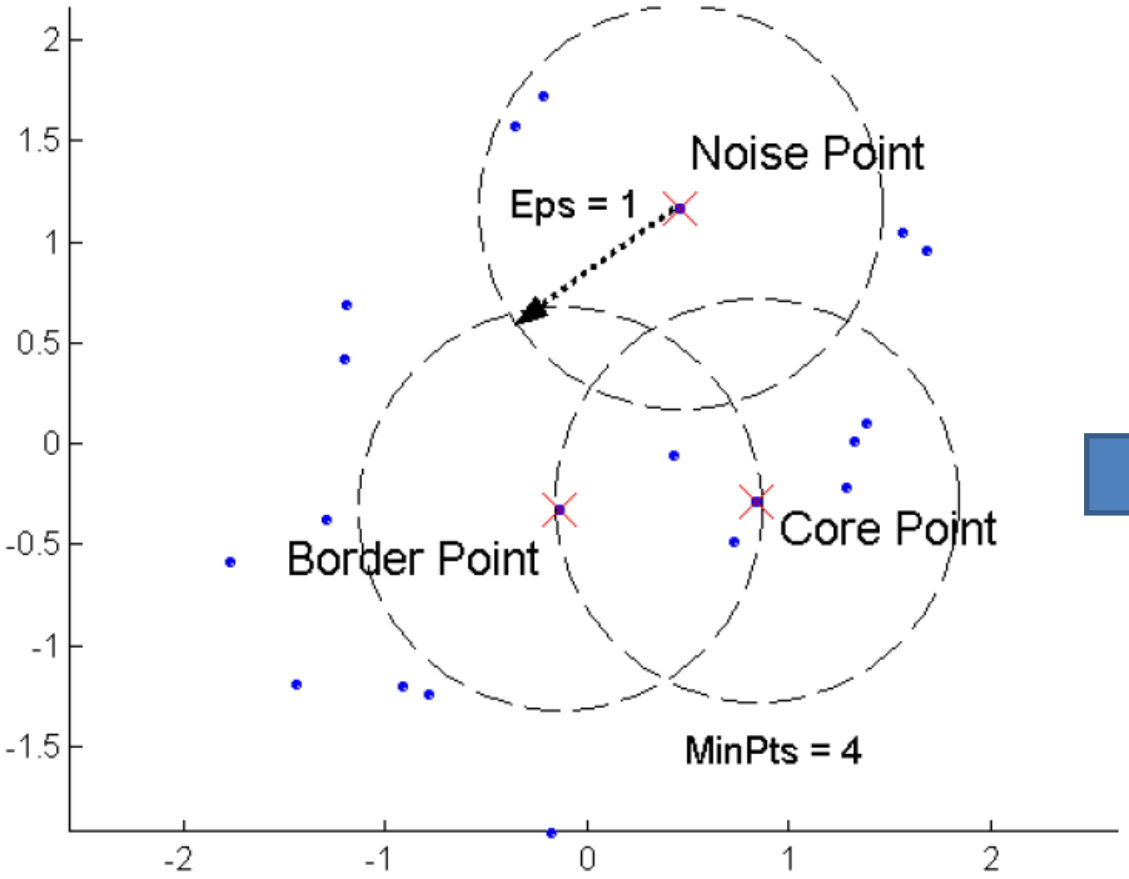
چگالی چگونه تعریف می‌شود؟

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

روش DBSCAN

دارای دو پارامتر است:

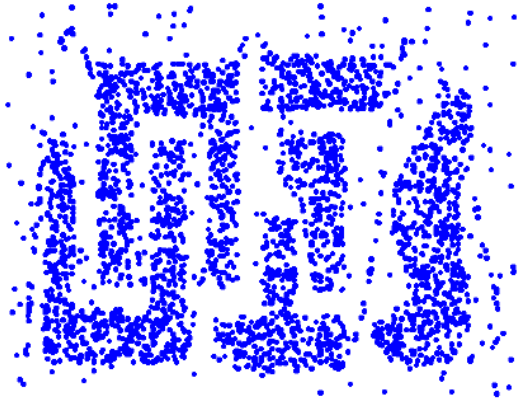
- شعاع همسایگی
- تعداد نقاط موجود در همسایگی

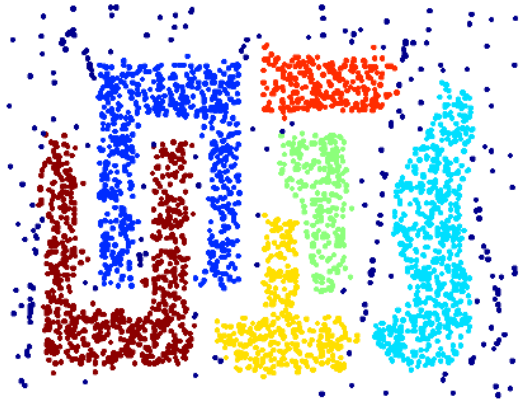


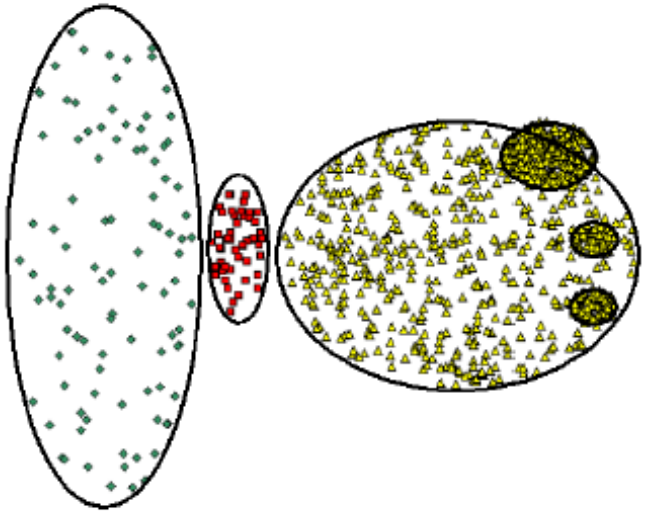
با سعی و خطا این دو ابرپارامتر محاسبه می‌شوند

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

روش DBSCAN







مزایا:

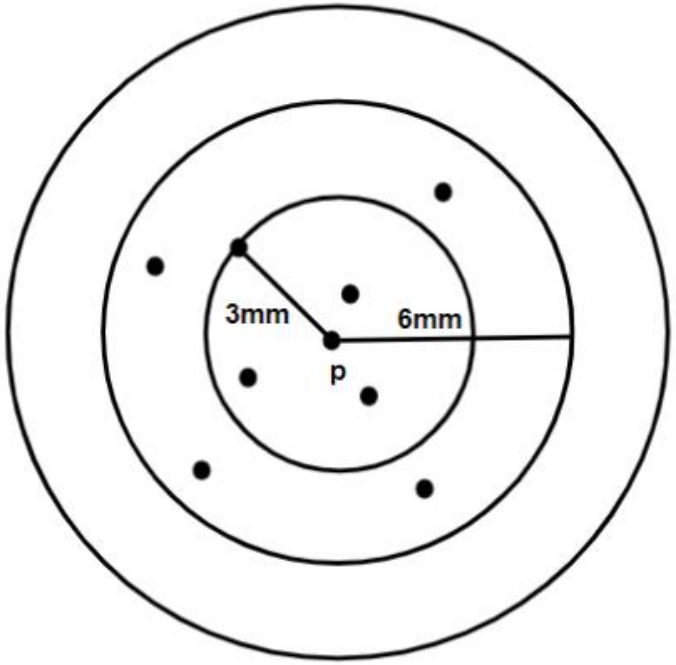
- مقاوم نسبت به نویز
- قادر به خوشه‌بندی داده‌ها با هر شکلی

معایب

- نامناسب برای چگالی‌های متغیر

تولید محتوا: وحید محمدزاده ایوقی

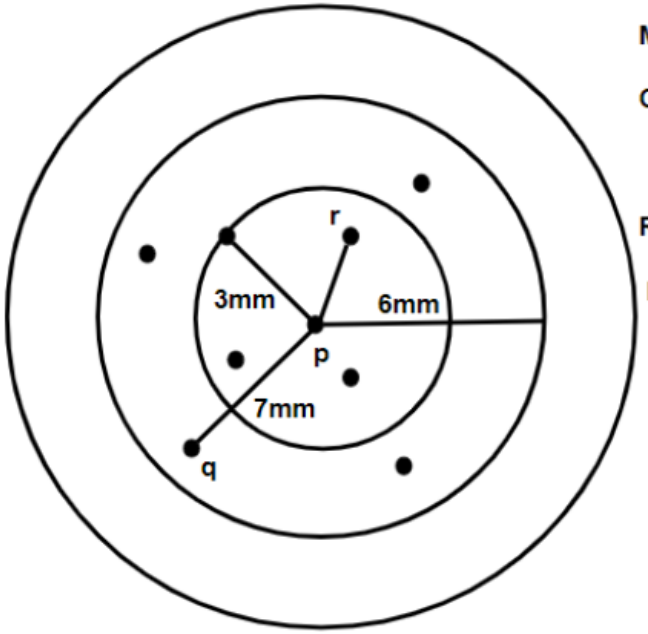
روش OPTICS



Eps = 6mm

MinPts = 5

Core_Distance(p) = 3mm



Eps = 6mm

MinPts = 5

Core_Distance(p) = 3mm

Reachability_Distance(q,p) = 7mm

Reachability_Distance(r,p) = 3mm

تولید محتوا: وحید محمدزاده ایوقی

daychegroup

daychegroup

dayche.com | گروه دایکه