Klasyfikacja i grupowanie

Pojęcia

Podobieństwo a odległość:

- odległość obiektów
- rozproszenie obiektów

Definicje miar podobieństwa

Cechy ilościowe

Cechy jakościowe

Porównanie obiektów do zbioru obiektów

Porównanie zbiorów obiektów

Określenie podobieństwa na podstawie odległości cech obiektów

$$simi(\bullet) = \frac{1}{1 + \alpha\, dist(\bullet)^{\beta}}$$

$$simi(\bullet) = 1 - \frac{dist(\bullet)}{\max(dist(\bullet))}$$

$$simi(\bullet) = e^{-\alpha dist(\bullet)}$$

Problemy określania podobieństwa

- Skalowanie i normalizacja
- Cechy ilościowe
- Cechy jakościowe
- Cechy ilościowo-jakościowe

Cechy ilościowe

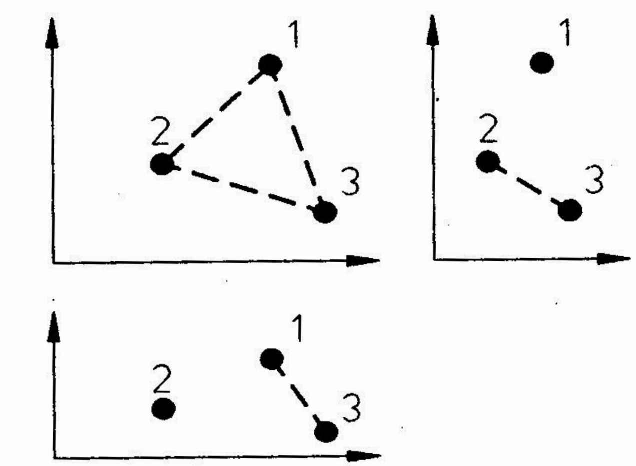
- Miara odległości Minkowskiego:

$$\mathrm{dist}_{\mathbf{M}}(\mathbf{v}_i, \mathbf{v}_j) = \left(\sum_{k=1}^{\dim(\underline{\mathbf{v}})} \left| v_{i,k} - v_{j,k} \right|^s \right)^{1/s}$$

- Miara odległości euklidesowej:

$$\mathrm{dist}_{\mathbf{E}}(\mathbf{v}_i, \mathbf{v}_j) = \sqrt{(\mathbf{v}_i - \mathbf{v}_j)^{\mathrm{T}} (\mathbf{v}_i - \mathbf{v}_j)}$$

- Wpływ skal współrzędnych na wynik grupowania



Cechy ilościowe

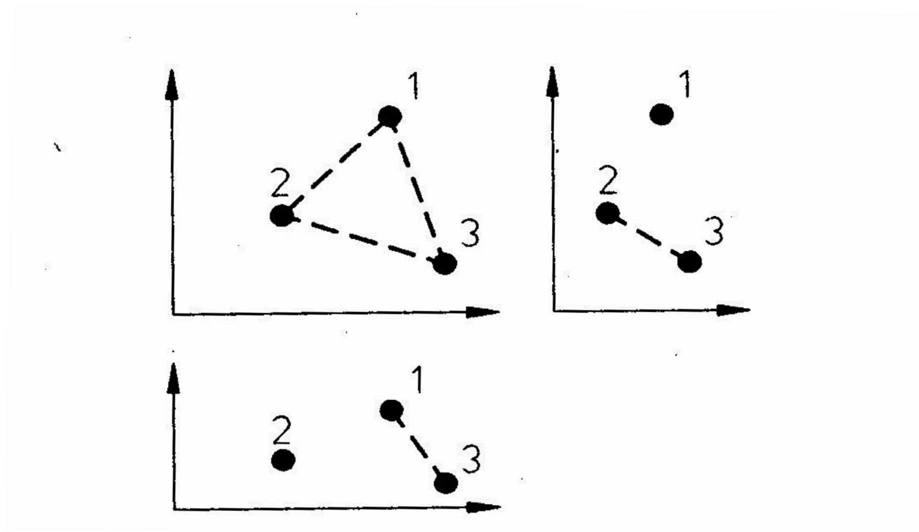
- Miara odległości Minkowskiego:

$$\mathrm{dist}_{\mathbf{M}}(\mathbf{v}_i, \mathbf{v}_j) = \left(\sum_{k=1}^{\dim(\underline{\mathbf{v}})} \left| v_{i,k} - v_{j,k} \right|^s \right)^{1/s}$$

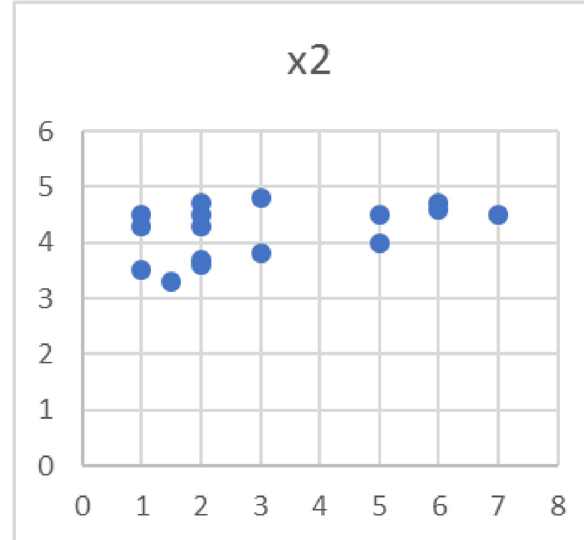
- Miara odległości euklidesowej:

$$\mathrm{dist}_{\mathbf{E}}(\mathbf{v}_i, \mathbf{v}_j) = \sqrt{\left(\mathbf{v}_i - \mathbf{v}_j\right)^{\mathrm{T}\left(\mathbf{v}_i - \mathbf{v}_j\right)}}$$

- Wpływ skal współrzędnych na wynik grupowania

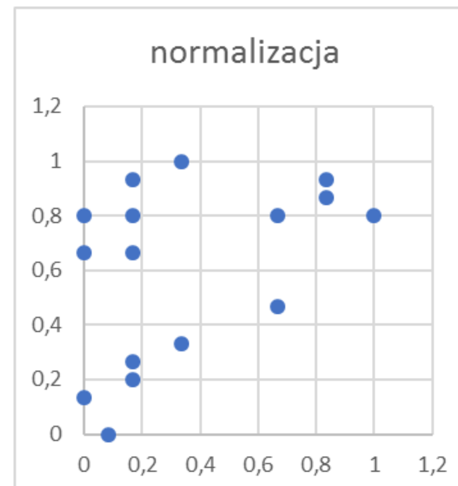


Cechy ilościowe



- Normalizacja

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

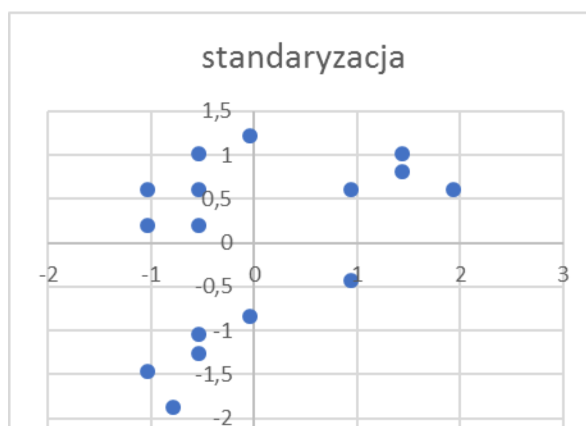


- Standaryzacja

$$x_i' = \frac{x_i - \bar{x}}{\sigma}$$

gdzie:

$$\sigma = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$



Cechy ilościowe – wagi cech

- Ważona miara odległości

$$\mathrm{dist}_{\mathbf{W}}(\underline{\mathbf{v}}_i, \underline{\mathbf{v}}_j) = \sqrt{\left(\underline{\mathbf{v}}_i - \underline{\mathbf{v}}_j\right)^{\mathrm{T}} \underline{\mathbf{W}} \left(\underline{\mathbf{v}}_i - \underline{\mathbf{v}}_j\right)}$$

$$\underline{\mathbf{W}} = \begin{bmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & w_n \end{bmatrix}$$

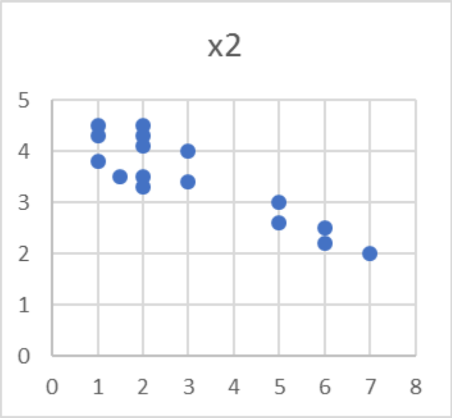
- Kowariancja

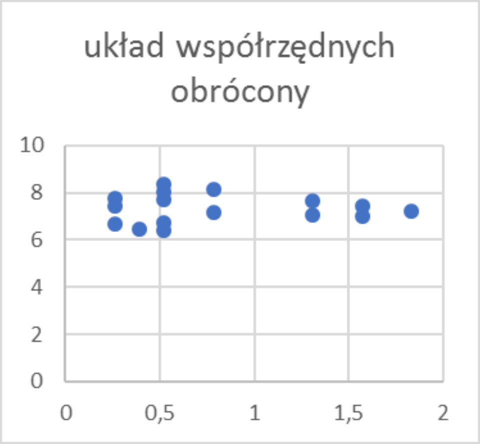
$$\boldsymbol{W} = \boldsymbol{C}^{-1}$$

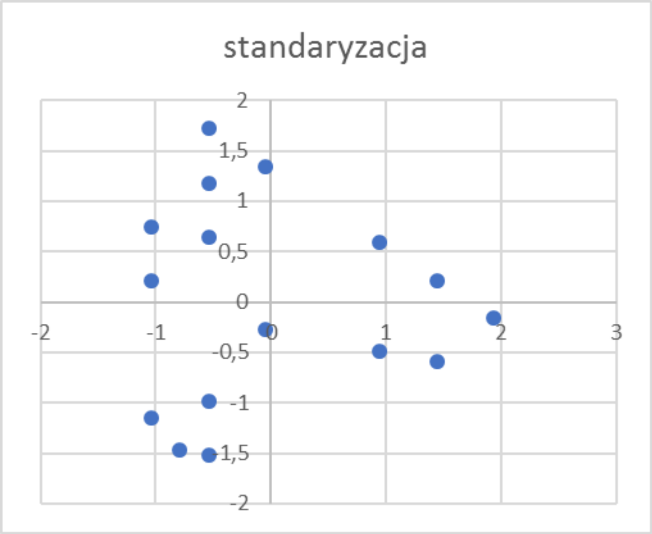
$$\boldsymbol{C} = \frac{1}{n} \sum_i (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T$$

Cechy ilościowe – wagi cech

- Ważona miara odległości – transformacje przestrzeni wartości cech; osie główne

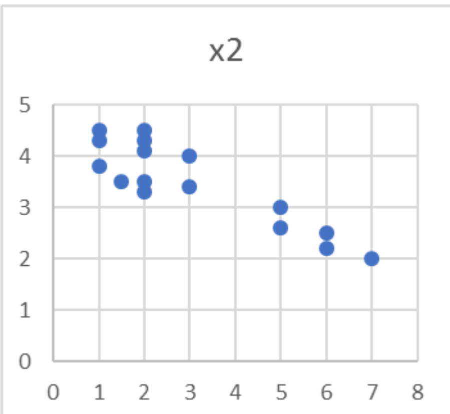


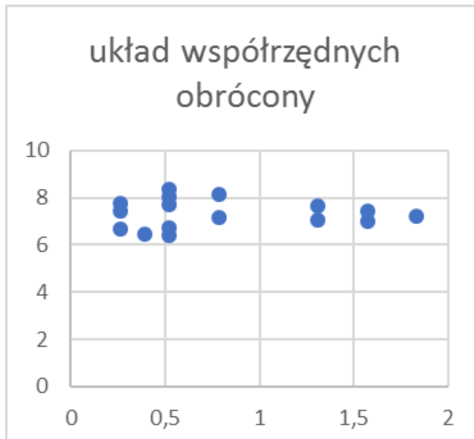


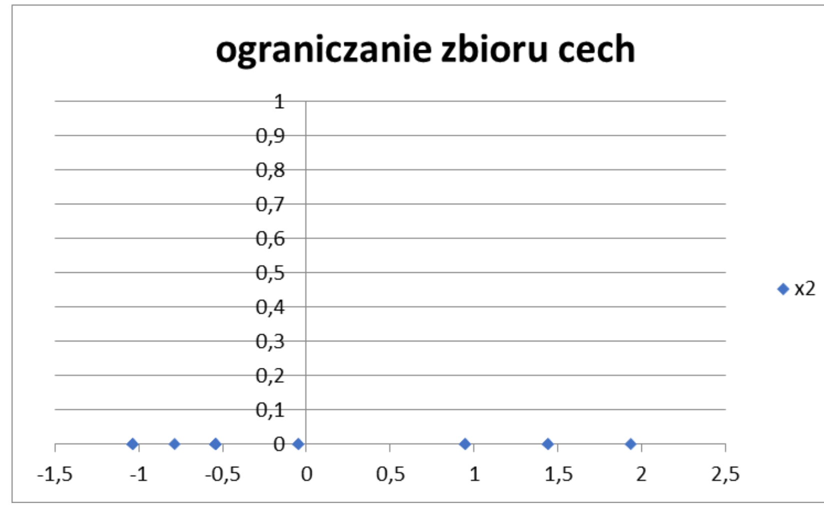


Cechy ilościowe – wagi cech

- Ważona miara odległości – ograniczanie zbioru uwzględnianych cech







Cechy ilościowe – wagi cech

- miara odległości – modyfikacja Hamminga:

$$\mathrm{dist}_{\mathbf{H}}(\underline{\mathbf{v}}_i, \underline{\mathbf{v}}_j) = \frac{1}{\dim(\underline{\mathbf{v}})} \sum_{k=1}^{\dim(\underline{\mathbf{v}})} \left| v_{ik} - v_{jk} \right|$$

gdzie: $v_{\mathrm{i},k}$ – jest wartością odpowiedniej *i*-tej cechy (obiektu) i *k*-tej współrzędnej.

Normalizacja tej modyfikacji:

$$\mathrm{dist}_{\mathbf{H}\mathrm{n}}(\underline{\mathbf{v}}_i, \underline{\mathbf{v}}_j) = \frac{1}{\dim(\underline{\mathbf{v}})} \sum_{k=1}^{\dim(\underline{\mathbf{v}})} \frac{\left| v_{ik} - v_{jk} \right|}{\max\limits_{m}(v_{mk}) - \min\limits_{m}(v_{mk})}$$

Cechy jakościowe – przejście na cechy ilościowe

1. Metryka nominalna (*overlap metric*):

$$\mathrm{dist}(\underline{\mathbf{v}}_i, \underline{\mathbf{v}}_j) = \begin{cases} 0 & \Leftrightarrow \underline{v}_i = \underline{v}_j \\ \infty & \Leftrightarrow \underline{v}_i \neq \underline{v}_j \end{cases}$$

2. Rangi dla wartości jakościowych:

$$\mathrm{dist}_{\underline{\mathbf{v}}_i \underline{\mathbf{v}}_j} = \frac{1}{\dim(\underline{\mathbf{v}})} \sum_{k=1}^{\dim(\underline{\mathbf{v}})} \left| \mathrm{rng}(v_{ik}) - \mathrm{rng}(v_{jk}) \right|$$

Cechy mieszane

- Zamiana wartości cech jakościowych na ilościowe i dopiero wtedy wyznaczenie podobieństwa

- Wyznaczenie osobnych podobieństw dla poszczególnych cech:

$$simi(\underline{\mathbf{v}}_i, \underline{\mathbf{v}}_j) = \alpha simi_{quan}(\underline{\mathbf{v}}_i, \underline{\mathbf{v}}_j) + (1 - \alpha) simi_{qual}(\underline{\mathbf{v}}_i, \underline{\mathbf{v}}_j)$$

$$simi(\underline{\mathbf{v}}_i, \underline{\mathbf{v}}_j) = simi_{quan}(\underline{\mathbf{v}}_i, \underline{\mathbf{v}}_j)^{\alpha} \cdot simi_{qual}(\underline{\mathbf{v}}_i, \underline{\mathbf{v}}_j)^{(1-\alpha)}$$

Porównanie obiektów ze zbiorami obiektów

- Podobieństwo średnie:

$$simi\left(\underline{\mathbf{v}}_i, \left\{\underline{\mathbf{v}}_j\right\}\right) = \frac{1}{\left|\left\{\underline{\mathbf{v}}_j\right\}\right|} \sum_{\underline{v} \in \left\{\underline{v}_j\right\}} simi\left(\underline{\mathbf{v}}_i, \underline{\mathbf{v}}\right)$$

- Podobieństwo między danym elementem a reprezentantem zbioru:

$$simi\left(\underline{\mathbf{v}}_i, \left\{\underline{\mathbf{v}}_j\right\}\right) = simi\left(\underline{\mathbf{v}}_i, \frac{1}{\left|\left\{\underline{\mathbf{v}}_j\right\}\right|} \sum_{\underline{v} \in \left\{\underline{v}_j\right\}} \underline{\mathbf{v}}\right)$$

$$simi\left(\underline{\mathbf{v}}_i, \left\{\underline{\mathbf{v}}_j\right\}\right) = \max_{\underline{\mathbf{v}} \in \left\{\underline{\mathbf{v}}_j\right\}} \left(simi\left(\underline{\mathbf{v}}_i, \underline{\mathbf{v}}\right)\right)$$