Wstęp do uczenia maszynowego

Analiza skupień

Ewa Szczurek + BW (modyfikacje)

bartek@mimuw.edu.pl
Instytut Informatyki
Uniwersytet Warszawski

UNIWERSYTET
WARSZAWSKI

PCA a klasteryzacja

- **PCA** pozwala znaleźć nisko wymiarową reprezentację obserwowanych danych, która pozwala wyjaśnić dużą część zmienności (wariancji) w danych.
- **Klasteryzacja** stara się znaleźć homogeniczne grupy wśród obserwowanych danych.
- Dwie metody:
  - Metoda K-średnich
  - Hierarchiczna klastryzacja

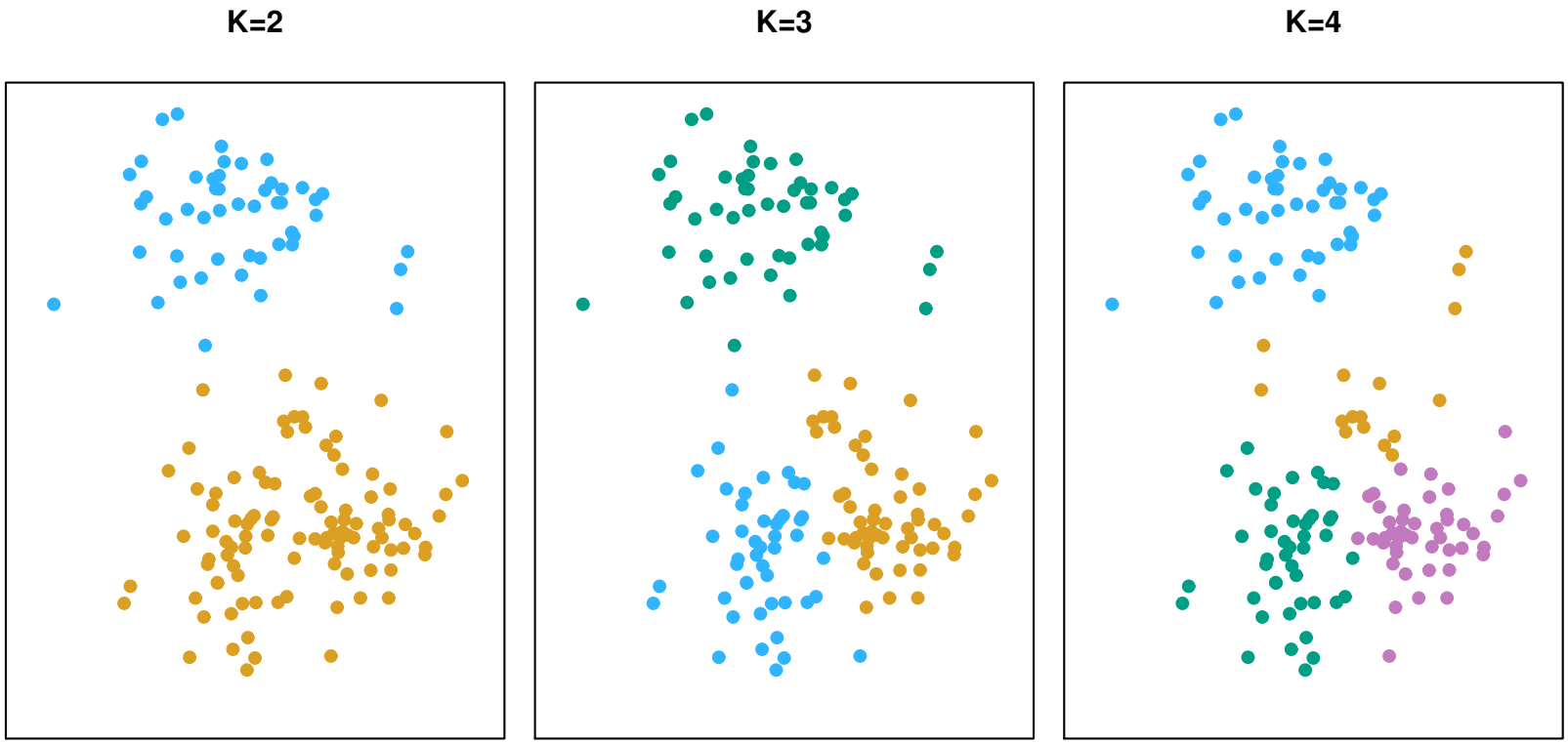
Metoda K-średnich

- Dane: obserwacje (wektory $x_1, \ldots, x_n \in \mathbb{R}^p$).
- Szukane: podział zbioru obserwacji na K niepustych i rozłącznych bloków (K jest ustaloną stałą) $C_1, \ldots, C_K$
  - $C_1 \cup C_2 \cup \ldots \cup C_K = \{x_1, \ldots, x_n\}$
  - $C_k \cap C_{k'} = \emptyset$, dla $k \neq k'$.
  - minimalizującego sumę kwadratów euklidesowych odległości pomiędzy wektorami w ramach każdego bloku podziału podzieloną przez liczbę wektorów w tym bloku.
  - Czyli minimalizujemy po $C_1, \ldots, C_K$ funkcję celu ($|C_k|$ jest liczbą elementów w klastrze C_k):

$$\sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \|x_i - x_{i'}\|^2.$$

- Ten problem jest trudny obliczeniowo (decyzyjna wersja jest **NP-trudnym problemem**).

Przykład klasteryzacji metodą K-średnich dla różnych K

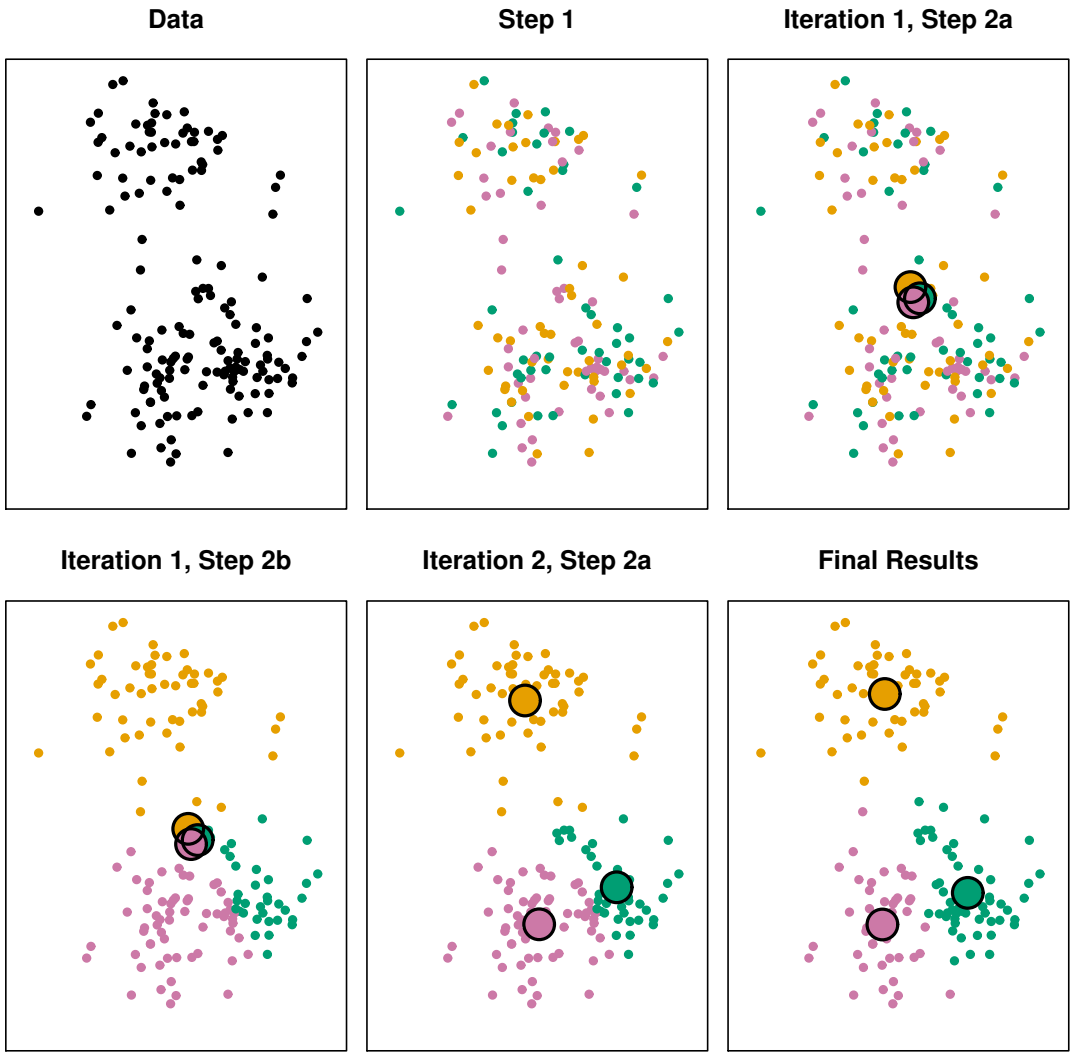


- dane symulowane, dla 150 obserwacji w dwuwymiarowej przestrzeni
- intuicyjnie, K-means szuka klastrowania, dla którego wariancja wewnątrz klastrów jest jak najmniejsza.

Heurystyczny algorytm klasteryzacji K-średnich

1. Losowo przypisz n obserwacji do K grup.
2. Powtarzaj poniższe kroki tak długo jak zmienia się przypisanie obserwacji do klastrów:
   (a) Dla każdego klastra wyznacz **centroid** dla tego klastra (wektor, który jest średnią po współrzędnych dla wszystkich obserwacji z tego klastra).
   (b) Każdą obserwację przypisz do tego klastra, dla którego euklidesowa odległość tej obserwacji od centroidu jest najmniejsza.

Przykładowy wynik algorytmu K-średnich ($K = 3$)



6 wyników algorytmu dla różnych inicjalizacji

Nad wykresami wartości funkcji celu (czerwona oznacza najlepsze wartości)

Nietrywialna równość

Udowodnimy następującą równość

$$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2,$$

gdzie

$$\bar{x}_{kj} = \frac{1}{|C_k|}\sum_{i'\in C_k} x_{i'j},$$

jest j-tą współrzędną centroida dla k-tej klasy.

Z tej równości wynika, że heurystyczny algorytm na każdym kroku poprawia funkcję celu, a zatem zbiega do **minimum lokalnego**. Trzeba wykonywać wiele losowych inicjalizacji i wybrać tę dającą najlepszy wynik.

Dowód równości – zapis wektorowy

Niech $x_i = [x_{i1} \dots x_{ip}]^T \in \mathbb{R}^p$ oraz niech $c = |C_k|$. Wówczas równość, którą chcemy udowodnić to

$$(1/c) \sum_{i,i'} \|x_i - x_{i'}\|^2 = 2 \sum_{i} \|x_i - (1/c) \sum_{i'} x_{i'}\|^2.$$

Lewa strona

$$L = (1/c)\sum_{i,i'} \|x_i - x_{i'}\|^2 = (2/c)\sum_{i<i'} \|x_i - x_{i'}\|^2$$
$$= (2/c)\sum_{i<i'} (x_i - x_{i'})^T (x_i - x_{i'}) = (2/c)\sum_{i<i'} (\|x_i\|^2 + \|x_{i'}\|^2 - 2x_i^T x_{i'})$$
$$= \frac{2(c-1)}{c}\sum_{i} \|x_i\|^2 - \frac{4}{c}\sum_{i'<i''} x_{i'}^T x_{i''}.$$

Ostatnia równość wynika stąd, że każde $\|x_i\|^2$ występuje w $c-1$ parach: $1i, 2i, \ldots, (i-1)i, i(i+1), \ldots, ic$.

Prawa strona (1)

$$P = 2\sum_i \|x_i - (1/c)\sum_{i'} x_{i'}\|^2 = \frac{2}{c^2}\sum_i \|(c-1)x_i - \sum_{i'\neq i} x_{i'}\|^2$$

$$= \frac{2}{c^2}\sum_i ((c-1)x_i - \sum_{i'\neq i} x_{i'})^T((c-1)x_i - \sum_{i'\neq i} x_{i'})$$

$$= \frac{2}{c^2}\sum_i \Big[(c-1)^2\|x_i\|^2 - 2(c-1)\sum_{i'\neq i} x_i^T x_{i'} + 2\sum_{i'<i'',i',i''\neq i} x_{i'}^T x_{i''} + \sum_{i'\neq i}\|x_{i'}\|^2\Big]$$

$$= \frac{2}{c^2}\sum_i \Big[((c-1)^2 - 1)\|x_i\|^2 - 2c\sum_{i'\neq i} x_i^T x_{i'} + 2\sum_{i'<i''} x_{i'}^T x_{i''} + \sum_{i'}\|x_{i'}\|^2\Big]$$

Prawa strona (2)

$$P = \frac{2}{c^2}\sum_i \Big[((c-1)^2-1)\|x_i\|^2 - 2c\sum_{i'\neq i} x_i^T x_{i'} + 2\sum_{i'<i''} x_{i'}^T x_{i''} + \sum_{i'}\|x_{i'}\|^2\Big]$$

$$= \frac{2}{c}\Big[(c-2)(\sum_i \|x_i\|^2) - 2(\sum_i\sum_{i'\neq i} x_i^T x_{i'}) + 2(\sum_{i'<i''} x_{i'}^T x_{i''}) + \sum_{i'}\|x_{i'}\|^2\Big]$$

$$= \frac{2(c-1)}{c}\sum_i \|x_i\|^2 + \frac{2}{c}\Big[2\sum_{i'<i''} x_{i'}^T x_{i''} - 2\cdot 2\sum_{i'<i''} x_{i'}^T x_{i''}\Big]$$

$$= \frac{2(c-1)}{c}\sum_i \|x_i\|^2 - \frac{4}{c}\sum_{i'<i''} x_{i'}^T x_{i''}.$$

Lewa strona (3)

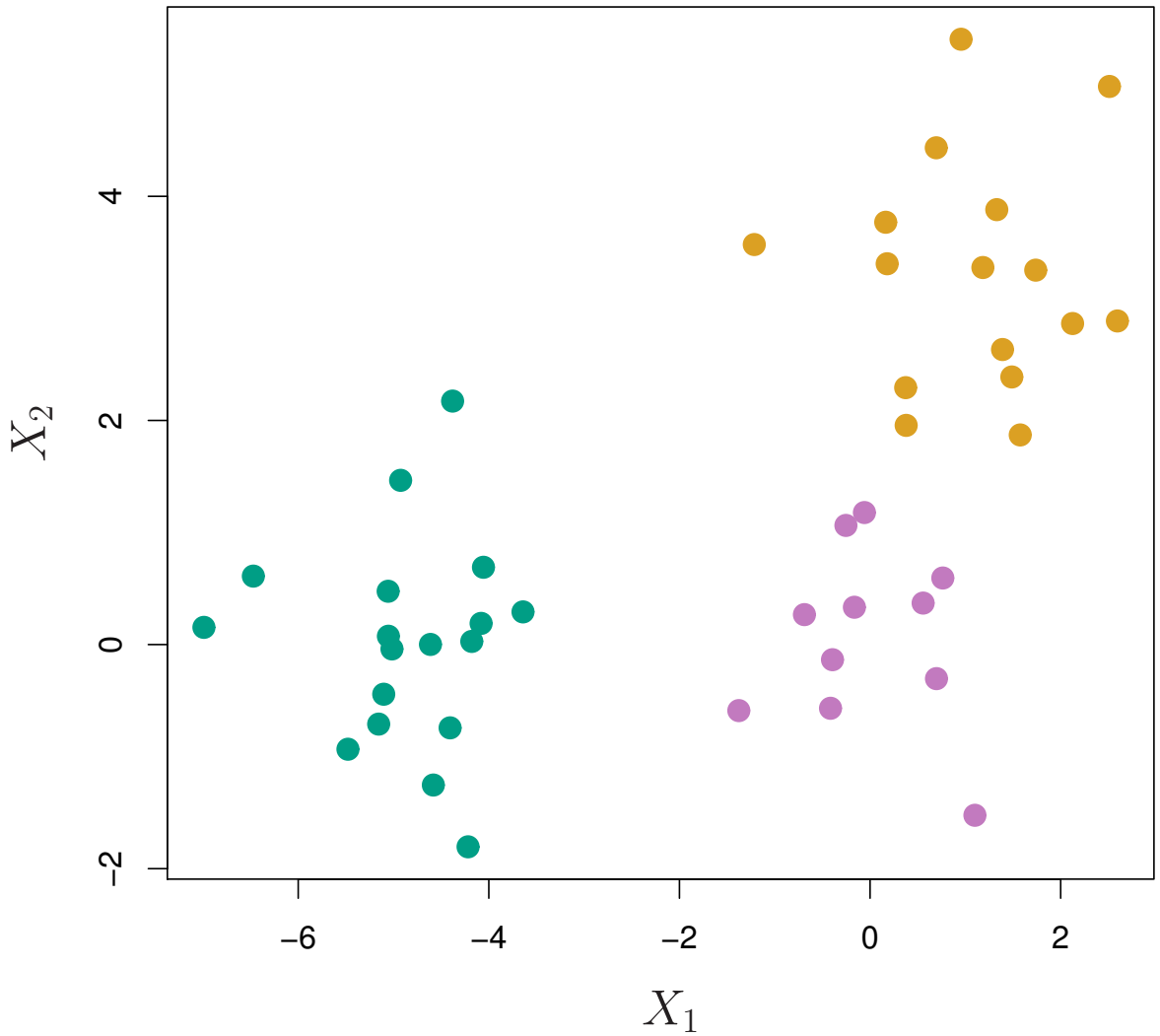
Przedostatnia równość wynika stąd, że

$$\sum_{i} \sum_{i' \neq i} x_i^T x_{i'} = 2 \sum_{i' < i''} x_{i'}^T x_{i''}.$$

Hierarchiczna klasteryzacja

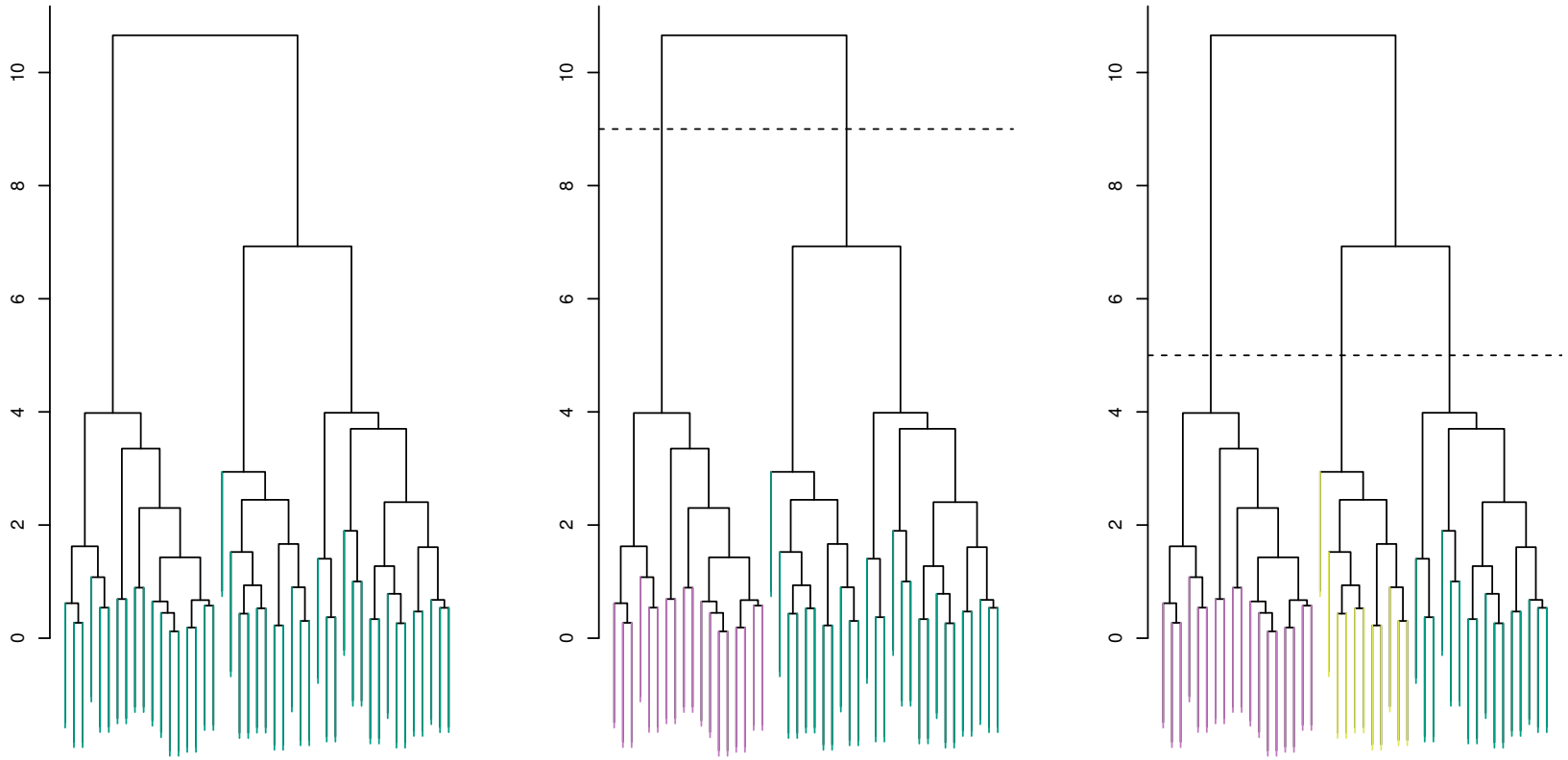
- Algorytmy hierarchicznej klastryzacji produkują **dendrogramy**.
- Dendrogram jest drzewem, które reprezentuje wiele różnych klastrowań (w zależności od poziomu cięcia tego drzewa).
- Liście odpowiadają poszczególnym obserwacjom.
- Struktura dendrogramu opisuje strukturę podobieństwa pomiędzy obserwacjami
  - poziom podobieństwa pomiędzy dwoma obserwacjami reprezentowany jest wysokością **najniższego wspólnego przodka** tych obserwacji (*least common ancestor* (LCA)).
  - Im ta wysokość jest mniejsza tym obserwacje są do siebie bardziej podobne.

Przykład: Dane losowe



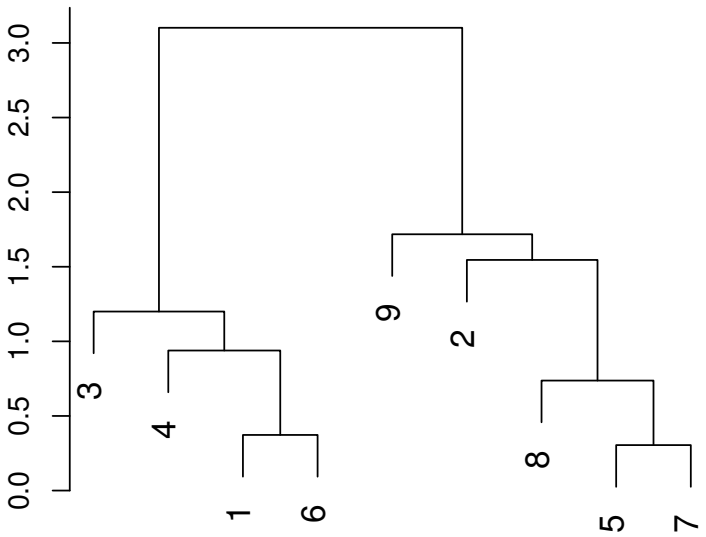
45 obserwacji, 3 klastry. Spróbujmy je odtworzyć!

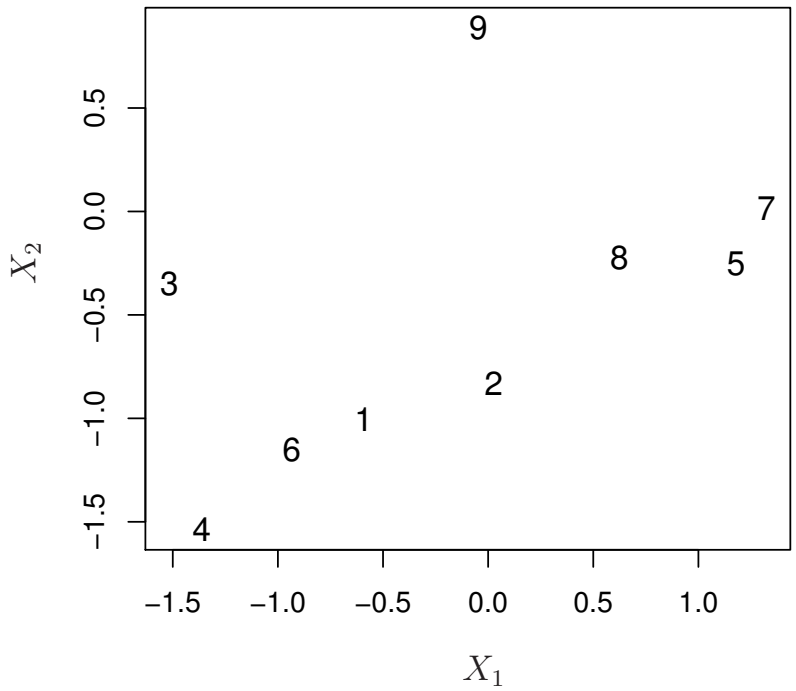
Przykład: Dendrogram i dwa różne klastrowania



- Klastrowanie hierarchiczne, pełne wiązanie.
- Wysokość cięcia determinuje liczbę klastrów

Dendrogram dla 9 obserwacji w przestrzeni 2-wymiarowej





- Obserwacje 5 i 7 są blisko siebie i klastrują się jako pierwsze
- Podobieństwo wg odległości euklidesowej
- Błędem jest stwierdzenie, że 9 i 2 są podobne, bo usytuowane blisko siebie na dendogramie!
- 9 nie musi być bardziej podobna do 2 niż do 8, 5, czy 7.
- O podobieństwie (odległości) wnioskujemy na podstawie osi pionowej

Algorytm hierarchicznego klastrowania

Dendrogramy są konstruowane w stylu *bottom-up* – zaczynamy od liści i tworzymy coraz większe klastry idąc w kierunku korzenia.
Wejście: n obserwacji oraz miara odległości pomiędzy klastrami.

1. Inicjujemy n klastrów, wszystkie jednoelementowe.
2. Dla $i = n, n-1, \ldots, 2$ wykonuj
   (a) Mamy i klastrów oraz obliczone odległości dla wszystkich $\binom{i}{2} = i(i-1)/2$ par klastrów. Wybierz dwa klastry, które są do siebie najbardziej podobne (najmniej odległe). Połącz je w nowy klaster, tworząc w ten sposób $i-1$ klastrów. Długość gałęzi w dendrogramie prowadzących do tych dwóch klastrów odpowiada odległości pomiędzy nimi (im mniej podobne tym krawędzie są dłuższe).
   (b) Oblicz odległości pomiędzy nowym klastrem i wszystkimi pozostałymi $i-2$ klastrami.

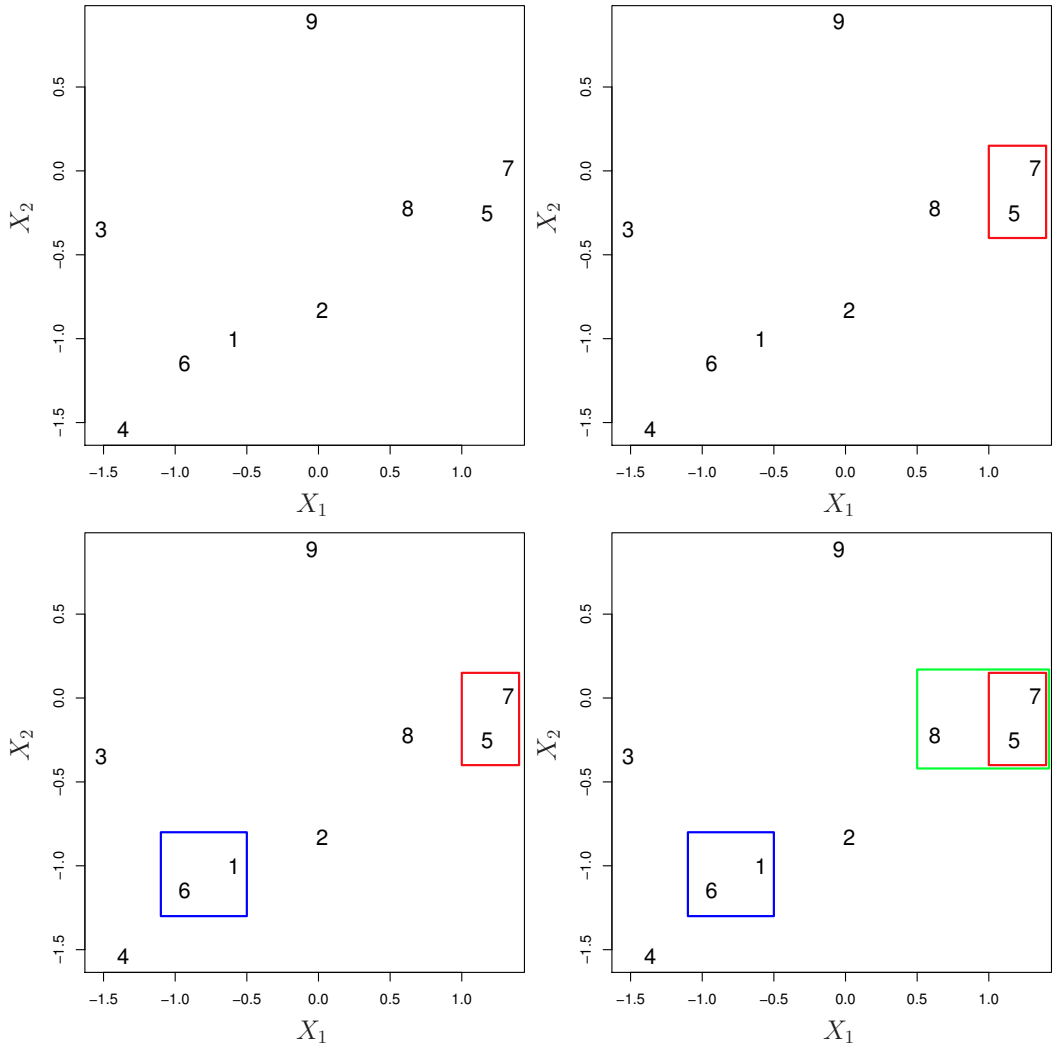
Cztery sposoby obliczania odległości $d(-,-)$ pomiędzy klastrami C, C'

Im większa wartość $d(x, x')$ tym mniej x oraz x' są do siebie podobne,

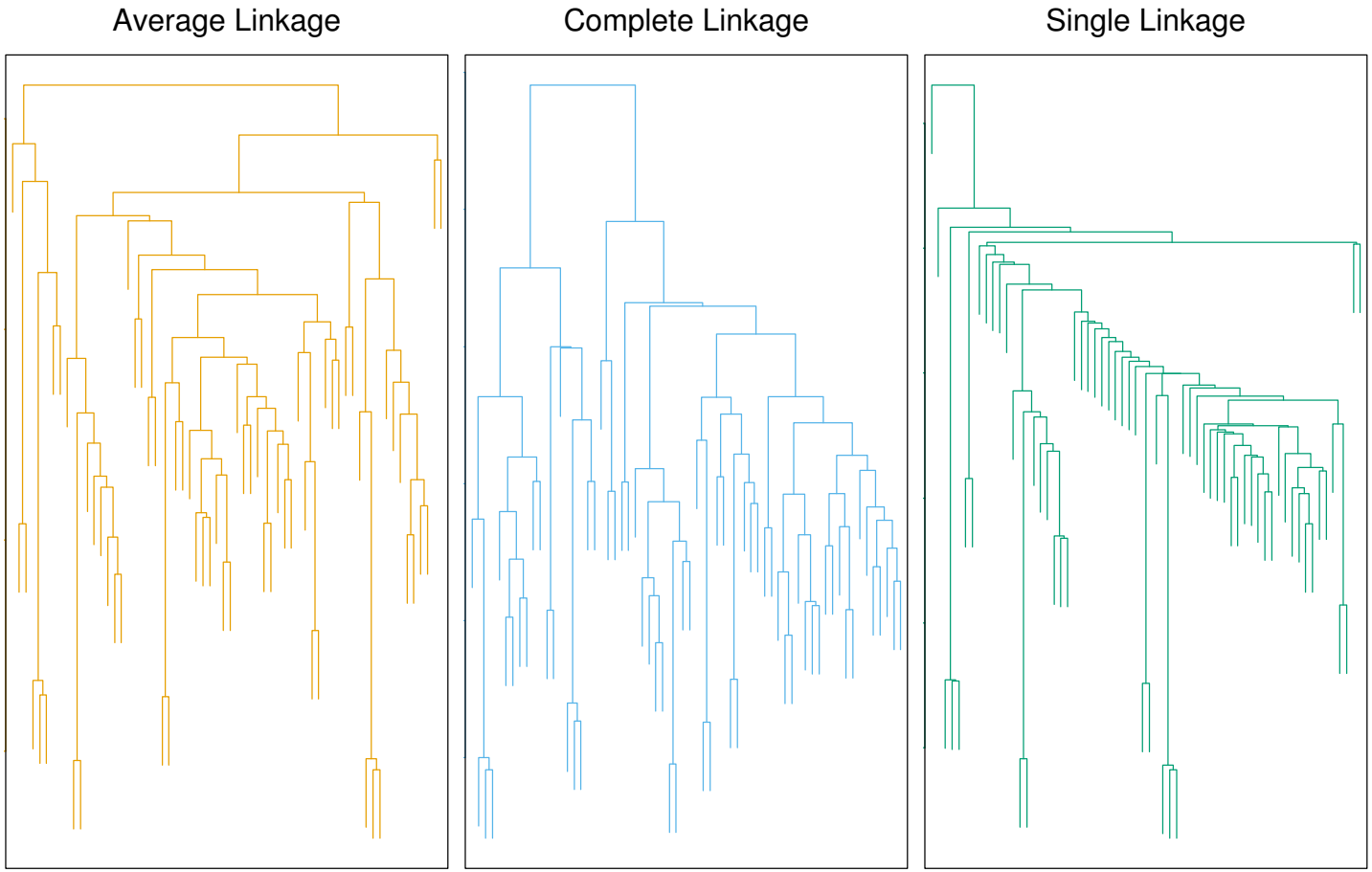
- (**Pełne wiązanie**, (*complete linkage*)): maksimum odległości pomiędzy elementami klastrów $\max\{d(x, x') \mid x \in C, x' \in C'\}$.
- (**Pojedyncze wiązanie** (*single linkage*)): minimum odległości pomiędzy elementami klastrów $\min\{d(x, x') \mid x \in C, x' \in C'\}$.
- (**Wiązanie średnich** (*average linkage*)): średnia z odległości pomiędzy elementami klastrów $\frac{1}{|C||C'|} \sum_{x \in C} \sum_{x' \in C'} d(x, x')$.
- (**Wiązanie centroidów** (*centroid linkage*)): odległość pomiędzy centroidami $d(\frac{1}{|C|} \sum_{x \in C} x, \frac{1}{|C'|} \sum_{x' \in C'} x')$.

Musi być określone dodawanie elementów i mnożenie ich przez skalar.

Kilka kroków algorytmu hierarchicznego klastrowania dla metody pełnego wiązania. Dane 2-wymiarowe



Dendrogramy otrzymane dla tych samych danych trzema metodami wiązań. Wiązania pełne i średnich dają bardziej zbalansowane dendrogramy

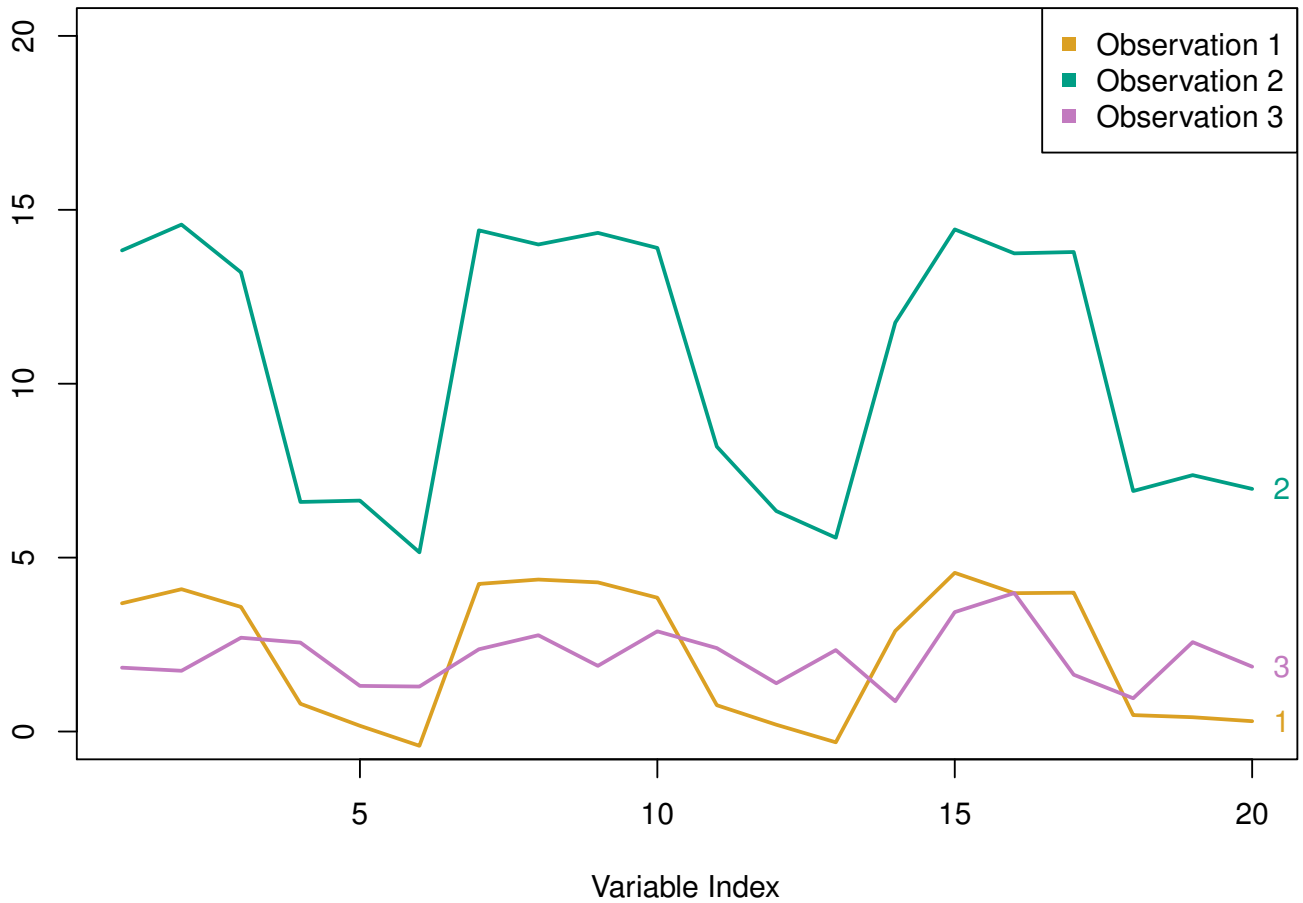


Różne miary podobieństwa

- Na przykład, można stosować:
  - Odległość euklidesową
  - Odległość opartą na korelacji
- Dendrogramy otrzymane dla różnych miar podobieństwa mogą być bardzo różne.
- Wybór miary podobieństwa zależy od rodzaju problemu.

Porównanie trzech obserwacji przy pomocy odległości euklidesowej i odległości korelacyjnej

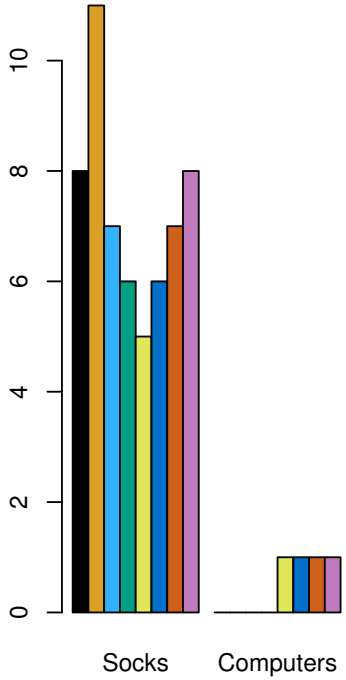
Obserwacje 1 i 3 są bliskie euklidesowo, a obserwacje 1 i 2 są bliskie korelacyjnie.

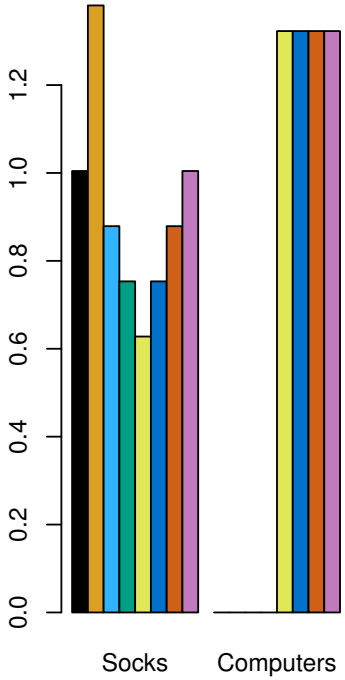


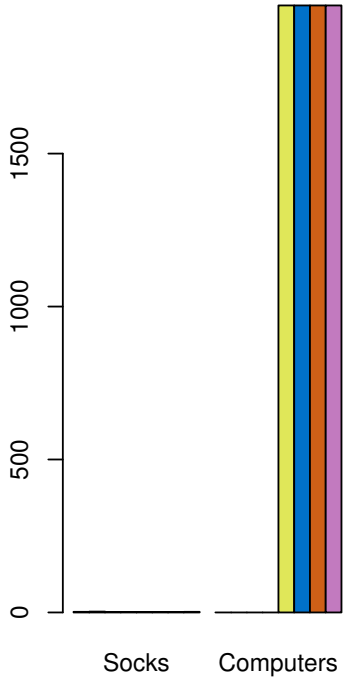
Podejmowane decyzje przed wykonaniem klasteryzacji

- Czy obserwacje powinny być standaryzowane przed klasteryzacją (np. czy zmienne mają zostać scentrowane (średnia=0, a standardowe odchylenie=1)?
- W przypadku hierarchicznego klastrowania:
  - Jakiej miary podobieństwa użyć?
  - Jaki wybrać typ wiązań?
  - Na jakiej wysokości wyznaczyć cięcie dendrogramu?
- W przypadku klasteryzacji metodą K-średnich, jak duże ma być K?

Czemu warto skalować zmienne







- # par skarpet i komputerów. Kolory: kupujący (obserwacje)
- Skarpety dominują odległość euklidesowa
- Te same dane po skalowaniu
- Komputery mają teraz większy wpływ
- Dolary wydane na skarpety i na komputery
- Komputery (dużo droższe) dominują

Zalety K-means

- Gdy struktura klastrowania nie jest zagnieżdżona, lepsze od hiearchicznego
  - Przykład: dane o kobietach i mężczyznach, trzech narodowości. Podział na narodowości (3 klastry) nie jest zagnieżdżony w podziale na płcie.

Zalety klastrowania hierarchicznego

- Brak konieczności zadania K przed klastrowaniem
- Dobór K często na podstawie oglądu dendogramu

Ocena jakości klastrowania

Intuicja:

1. dobre klastrowanie to takie, gdzie elementy wewnątrz klastrów są bardziej podobne do siebie niż pomiędzy klastrami. Przykład: *silhouette score*
2. dobre klastrowanie to takie, które jest podobne do jakiegoś innego klastrowania. Przykład: *rand index*

Silhouette score

Niech $i \in C_I$ punkt danych w klastrze C_I, $d(i,j)$ odległość między punktami i oraz j

- Średnia odległość punktu i od innych punktów *wewnątrz* klastra

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i,j)$$

(dzielimy przez $|C_I| - 1$ bo nie uwzględniamy odległości i od siebie

- Najmniejsza średnia odległość do punktów z innych klastrów

$$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j)$$

- Silhouette score

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ dla} |C_I| > 1$$

oraz

$$s(i) = 0 \ dla \ |C_I| = 1$$

Silhouette score

Ten sam wzór

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ dla} |C_I| > 1$$

oraz

$$s(i) = 0 \ dla \ |C_I| = 1$$

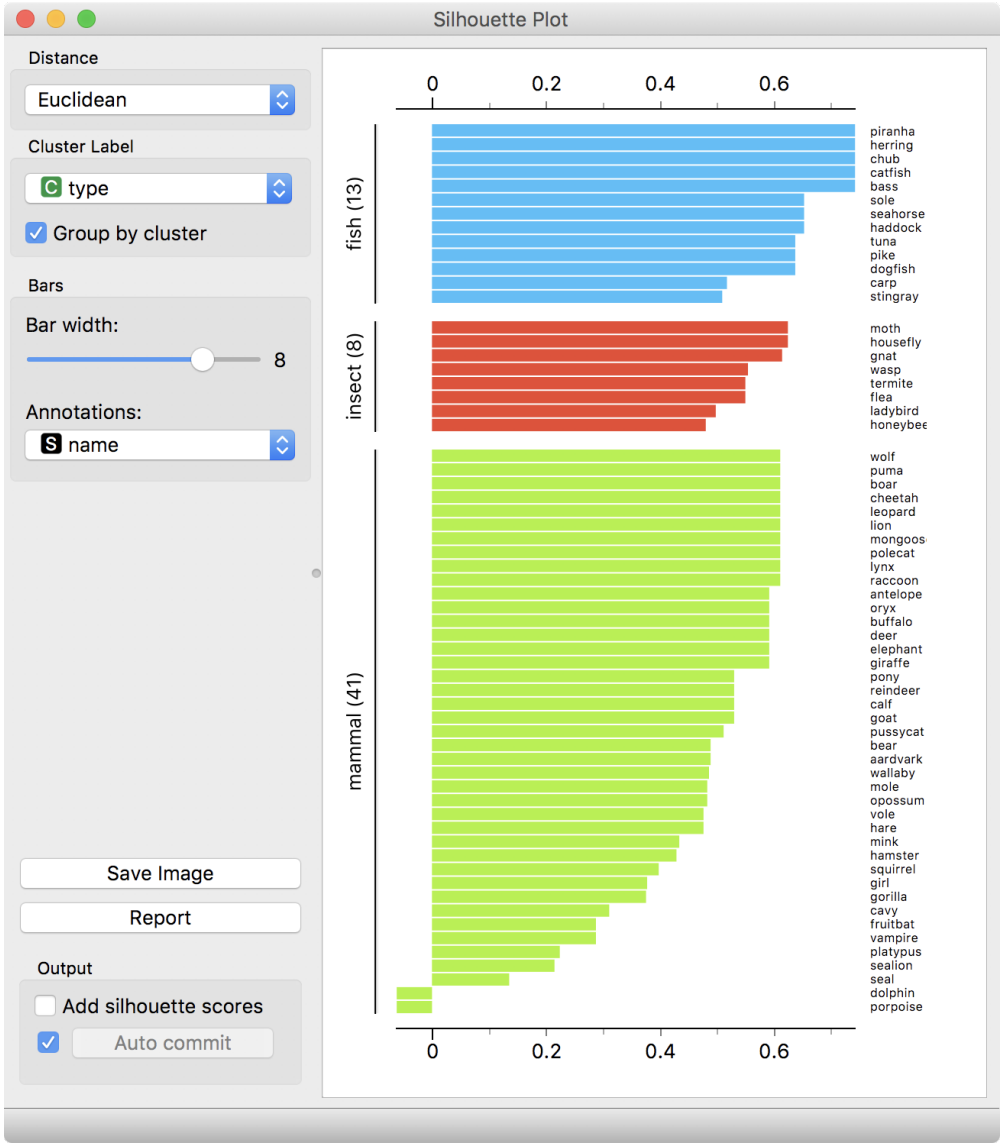
można zapisać inaczej

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{dla } a(i) < b(i) \\ 0, & \text{dla } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{dla } a(i) > b(i) \end{cases}$$

A zatem mamy $-1 \leq s(i) \leq 1$.

Uwaga: dla klastrów wielkości 1 ustawiamy $s(i) = 0$.

Silhouette score



Silhouette Coefficient

- Współczynnik $s(i)$ jest zdefiniowany dla pojedyńczej obserwacji i w klastrze C_I.
- Możemy też zdefiniować współczynnik średni $\bar{s}$, będący średnią wartością $s(i)$ po wszystkich obserwacjach we wszystkich klastrach.
- Jeżeli chcemy wybrać najlepsze k, w sensie *Silhouette score*, dla klastrowania k-średnich, możemy to zrobić znadjując

$$argmax_k \bar{s}(k)$$

gdzie $\bar{s}(k)$ jest współczynnikiem średniego silhouette dla wyniku k-średnich przy parametrze k

Rand index

Zadane:

- zbiór n punktów $S = \{o_1, \ldots, o_n\}$
- dwa klastrowania zbioru S, które porównujemy: $X = \{X_1, \ldots, X_r\}$ (podział na r podzbiorów) i $Y = \{Y_1, \ldots, Y_s\}$ (podział na s podzbiorów

Oznaczenia:

- a: liczba par punktów z S, które są w **tych samych** podzbiorach w X oraz w **tych samych** podzbiorach w Y
- b: liczba par punktów z S, które są w **innych** podzbiorach w X oraz w **innych** podzbiorach w Y
- c: liczba par punktów z S, które są w **tych samych** podzbiorach w X oraz w **innych** podzbiorach w Y
- c: liczba par punktów z S, które są w **innych** podzbiorach w X oraz w **tych samych** podzbiorach w Y

Rand index

Oznaczenia:

- a: liczba par punktów z S, które są w **tych samych** podzbiorach w X oraz w **tych samych** podzbiorach w Y
- b: liczba par punktów z S, które są w **innych** podzbiorach w X oraz w **innych** podzbiorach w Y
- c: liczba par punktów z S, które są w **tych samych** podzbiorach w X oraz w **innych** podzbiorach w Y
- c: liczba par punktów z S, które są w **innych** podzbiorach w X oraz w **tych samych** podzbiorach w Y

Rand index:

$$RI = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

A zatem $RI \in < 0, 1 >$

Adjusted rand index: intuicja

Idea: poprawić, unormować Rand index, tak, aby uwzględnić zgodność podziałów występującą 'przez przypadek'. Intuicyjnie,

$$ARI = \frac{RI - ExpectedRI}{MaxRI - ExpectedRI}$$

Adjusted rand index

Zgodność podziałów X i Y może być podsumowana tablicą kontyngencji $[n_{ij}]$, w której każde wejście n_{ij} oznacza kardynalność przecięcia zbiorów X_i i Y_j, czyli $n_{ij} = |X_i \cap Y_j|$:

$X \backslash Y$	Y_1	Y_2	$\cdots$	Y_s	sumy
X_1	n_{11}	n_{12}	$\cdots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\cdots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	a_r
sumy	b_1	b_2	$\cdots$	b_s	

Wówczas

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}$$

Podsumowanie

- k-means
- klastrowanie hierarchiczne
- Silhouette score
- Rand index