Statystyczna Eksploracja Danych

Wykład 7 - analiza skupień

dr inż. Julian Sienkiewicz

16 kwietnia 2019

Cele i ogólny opis

Cele i ogólny opis

- **analiza skupień** ma na celu wykrycie w zbiorze obserwacji struktur zwanych **skupieniami**...

Cele i ogólny opis

- **analiza skupień** ma na celu wykrycie w zbiorze obserwacji struktur zwanych **skupieniami**...
- ... czyli rozłącznych podziorów zbioru obserwacji, wewnątrz których obserwacje są w jakimś określonym sensie bliskie,

Cele i ogólny opis

- **analiza skupień** ma na celu wykrycie w zbiorze obserwacji struktur zwanych **skupieniami**...
- ... czyli rozłącznych podziorów zbioru obserwacji, wewnątrz których obserwacje są w jakimś określonym sensie bliskie,
- podzbiory różne są od siebie odległe (w porównaniu z elementami wewnątrz każdego podzbioru),

Cele i ogólny opis

- **analiza skupień** ma na celu wykrycie w zbiorze obserwacji struktur zwanych **skupieniami**...
- ... czyli rozłącznych podziorów zbioru obserwacji, wewnątrz których obserwacje są w jakimś określonym sensie bliskie,
- podzbiory różne są od siebie odległe (w porównaniu z elementami wewnątrz każdego podzbioru),
- jest to przypadek **klasyfikacji bez nadzoru**, czyli **nie** mamy próby uczącej, ani też wiedzy o tym, jak przypisać klasę do obserwacji,

Cele i ogólny opis

- **analiza skupień** ma na celu wykrycie w zbiorze obserwacji struktur zwanych **skupieniami**...
- ... czyli rozłącznych podziorów zbioru obserwacji, wewnątrz których obserwacje są w jakimś określonym sensie bliskie,
- podzbiory różne są od siebie odległe (w porównaniu z elementami wewnątrz każdego podzbioru),
- jest to przypadek **klasyfikacji bez nadzoru**, czyli **nie** mamy próby uczącej, ani też wiedzy o tym, jak przypisać klasę do obserwacji,
- zakładamy, że liczba skupień jest **z góry ustalona**, co czyni wyznaczenie skupień dobrze zdefiniowanym zadaniem optymalizacyjnym.

Analiza skupień w R^p

Analiza skupień w R^p

Na początek: analiza skupień w przestrzeni euklidesowej R^p

- mamy n-elementowy zbiór obserwacji $\mathbf{x}_i$, $i = 1, ..., n$ o wartościach w R^p,

Analiza skupień w R^p

Na początek: analiza skupień w przestrzeni euklidesowej R^p

- mamy n-elementowy zbiór obserwacji $\mathbf{x}_i$, $i = 1, ..., n$ o wartościach w R^p,
- chcemy podzielic tę próbę na K skupień,

Analiza skupień w R^p

Na początek: analiza skupień w przestrzeni euklidesowej R^p

- mamy n-elementowy zbiór obserwacji $\mathbf{x}_i$, $i = 1, ..., n$ o wartościach w R^p,
- chcemy podzielic tę próbę na K skupień,

$$T = \frac{1}{2} \sum_{i=1}^{n} \sum_{i'=1}^{n} d_{ii'}$$

suma kwadratów odległości pomiędzy parami punktów próby

$$d_{ii'} = d(\mathbf{x}_i, \mathbf{x}_{i'})$$

kwadrat odegłości pomiędzy obserwacjami $\mathbf{x}_i$ i $\mathbf{x}_{i'}$

Analiza skupień w R^p

Na początek: analiza skupień w przestrzeni euklidesowej R^p

- mamy n-elementowy zbiór obserwacji $\mathbf{x}_i$, $i = 1, ..., n$ o wartościach w R^p,
- chcemy podzielic tę próbę na K skupień,

$$T = \frac{1}{2} \sum_{i=1}^{n} \sum_{i'=1}^{n} d_{ii'}$$

suma kwadratów odległości pomiędzy parami punktów próby

$$d_{ii'} = d(\mathbf{x}_i, \mathbf{x}_{i'})$$

kwadrat odegłości pomiędzy obserwacjami $\mathbf{x}_i$ i $\mathbf{x}_{i'}$

- w ten sposób dokonaliśmy arbitralnego podziału obserwacji na K rozłącznych podzbiorów $k = 1, .., K$, gdzie oznaczymy $C(i) = k$ jako przynależność i-tej obserwacji $\mathbf{x}_i$ do k-tego podzbioru.

Analiza skupień w R^p

Analiza skupień w R^p

Sumę kwadratów T można rozłożyć na sumę kwadratów odległości pomiędzy parami

Analiza skupień w R^p

Sumę kwadratów T można rozłożyć na sumę kwadratów odległości pomiędzy parami **należącymi do tego samego skupienia**

Analiza skupień w R^p

Sumę kwadratów T można rozłożyć na sumę kwadratów odległości pomiędzy parami **należącymi do tego samego skupienia** oraz

Analiza skupień w R^p

Sumę kwadratów T można rozłożyć na sumę kwadratów odległości pomiędzy parami **należącymi do tego samego skupienia** oraz parami **należącymi do różnych skupień**

Analiza skupień w R^p

Sumę kwadratów T można rozłożyć na sumę kwadratów odległości pomiędzy parami **należącymi do tego samego skupienia** oraz parami **należącymi do różnych skupień**

$$T = W + B$$

Analiza skupień w R^p

Sumę kwadratów T można rozłożyć na sumę kwadratów odległości pomiędzy parami **należącymi do tego samego skupienia** oraz parami **należącymi do różnych skupień**

$$T = W + B$$

W - within the cluster
B - between clusters

Analiza skupień w R^p

Sumę kwadratów T można rozłożyć na sumę kwadratów odległości pomiędzy parami **należącymi do tego samego skupienia** oraz parami **należącymi do różnych skupień**

$$T = W + B$$

$$W = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} d_{ii'}$$

W - within the cluster
B - between clusters

Analiza skupień w R^p

Sumę kwadratów T można rozłożyć na sumę kwadratów odległości pomiędzy parami **należącymi do tego samego skupienia** oraz parami **należącymi do różnych skupień**

$$T = W + B$$

W - within the cluster
B - between clusters

$$W = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} d_{ii'}$$

$$B = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')\neq k} d_{ii'}$$

Analiza skupień w R^p

Sumę kwadratów T można rozłożyć na sumę kwadratów odległości pomiędzy parami **należącymi do tego samego skupienia** oraz parami **należącymi do różnych skupień**

$$T = W + B$$

$$W = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} d_{ii'}$$

W - within the cluster
B - between clusters

$$B = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')\neq k} d_{ii'}$$

Zmieniając podział punktów na K skupień, zmieniamy W i B (T jest takie samo).

Analiza skupień w R^p

Sumę kwadratów T można rozłożyć na sumę kwadratów odległości pomiędzy parami **należącymi do tego samego skupienia** oraz parami **należącymi do różnych skupień**

$$T = W + B$$

$$W = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} d_{ii'}$$

W - within the cluster
B - between clusters

$$B = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')\neq k} d_{ii'}$$

Zmieniając podział punktów na K skupień, zmieniamy W i B (T jest takie samo).

Czyli analiza skupień to minimalizacja rozrzutu punktów wewnatrz skupień — minimalizacja W (maksymalizacja B).

Oczywiście, ogólnie jest zadanie kombinatoryczne, ale liczba sposobów, na ile można podzielić n obserwacji na K skupień to $\frac{1}{K!}\sum_{k=1}^{K}(-1)^{K-k}\left(\begin{array}{c} K \\ k \end{array}\right)k^n$, co w przypadku $n = 100$ i $K = 5$ daje około 10^{67}.

Oczywiście, ogólnie jest zadanie kombinatoryczne, ale liczba sposobów, na ile można podzielić n obserwacji na K skupień to $\frac{1}{K!}\sum_{k=1}^{K}(-1)^{K-k}\left(\begin{array}{c} K \\ k \end{array}\right)k^n$, co w przypadku $n = 100$ i $K = 5$ daje około 10^{67}.

Sumę W można też zapisać jako

$$W = \sum_{k=1}^{K}\sum_{C(i)=k} d(\mathbf{x}_i, \mathbf{m}_k)n_k,$$

Oczywiście, ogólnie jest zadanie kombinatoryczne, ale liczba sposobów, na ile można podzielić n obserwacji na K skupień to $\frac{1}{K!}\sum_{k=1}^{K}(-1)^{K-k}\left(\begin{array}{c} K \\ k \end{array}\right)k^n$, co w przypadku $n = 100$ i $K = 5$ daje około 10^{67}.

Sumę W można też zapisać jako

$$W = \sum_{k=1}^{K}\sum_{C(i)=k} d(\mathbf{x}_i, \mathbf{m}_k)n_k,$$

gdzie

- n_k - licznośc skupienia k (liczba obserwacji),

Oczywiście, ogólnie jest zadanie kombinatoryczne, ale liczba sposobów, na ile można podzielić n obserwacji na K skupień to $\frac{1}{K!}\sum_{k=1}^{K}(-1)^{K-k}\left(\begin{array}{c} K \\ k \end{array}\right)k^n$, co w przypadku $n = 100$ i $K = 5$ daje około 10^{67}.

Sumę W można też zapisać jako

$$W = \sum_{k=1}^{K}\sum_{C(i)=k} d(\mathbf{x}_i, \mathbf{m}_k)n_k,$$

gdzie

- n_k - licznośc skupienia k (liczba obserwacji),
- $\mathbf{m}_k = \frac{1}{n_k}\sum_{C(i)=k}\mathbf{x}_i$ - średnia wektorowa obserwacji należących do k-tego skupienia (środek skupienia)

Oczywiście, ogólnie jest zadanie kombinatoryczne, ale liczba sposobów, na ile można podzielić n obserwacji na K skupień to $\frac{1}{K!}\sum_{k=1}^{K}(-1)^{K-k}\left(\begin{array}{c} K \\ k \end{array}\right)k^n$, co w przypadku $n = 100$ i $K = 5$ daje około 10^{67}.

Sumę W można też zapisać jako

$$W = \sum_{k=1}^{K}\sum_{C(i)=k} d(\mathbf{x}_i, \mathbf{m}_k)n_k,$$

gdzie

- n_k - licznośc skupienia k (liczba obserwacji),
- $\mathbf{m}_k = \frac{1}{n_k}\sum_{C(i)=k}\mathbf{x}_i$ - średnia wektorowa obserwacji należących do k-tego skupienia (środek skupienia)

Uproszczenie zadania minimalizacji

Zamiast sumy W

$$\widetilde{W} = \sum_{k=1}^{K} \sum_{C(i)=k} d(\mathbf{x}_i, \mathbf{m}_k) = \sum_{i=1}^{N} d(\mathbf{x}_i, \mathbf{m}_{C(i)})$$

Uproszczenie zadania minimalizacji

Zamiast sumy W

$$\widetilde{W} = \sum_{k=1}^{K} \sum_{C(i)=k} d(\mathbf{x}_i, \mathbf{m}_k) = \sum_{i=1}^{N} d(\mathbf{x}_i, \mathbf{m}_{C(i)})$$

Algorytm K-średnich (K-means)

Uproszczenie zadania minimalizacji

Zamiast sumy W

$$\widetilde{W} = \sum_{k=1}^{K} \sum_{C(i)=k} d(\mathbf{x}_i, \mathbf{m}_k) = \sum_{i=1}^{N} d(\mathbf{x}_i, \mathbf{m}_{C(i)})$$

Algorytm K-średnich (K-means)

0. inicjalizacja początkowych K środków $\mathbf{m}_K$,

Uproszczenie zadania minimalizacji

Zamiast sumy W

$$\widetilde{W} = \sum_{k=1}^{K} \sum_{C(i)=k} d(\mathbf{x}_i, \mathbf{m}_k) = \sum_{i=1}^{N} d(\mathbf{x}_i, \mathbf{m}_{C(i)})$$

Algorytm K-średnich (K-means)

0. inicjalizacja początkowych K środków $\mathbf{m}_K$,
1. w pierwszym kroku przypisujemy punkty do najbliższych środków $\mathbf{m}_k$

Uproszczenie zadania minimalizacji

Zamiast sumy W

$$\widetilde{W} = \sum_{k=1}^{K} \sum_{C(i)=k} d(\mathbf{x}_i, \mathbf{m}_k) = \sum_{i=1}^{N} d(\mathbf{x}_i, \mathbf{m}_{C(i)})$$

Algorytm K-średnich (K-means)

0. inicjalizacja początkowych K środków $\mathbf{m}_K$,
1. w pierwszym kroku przypisujemy punkty do najbliższych środków $\mathbf{m}_k$
   - jeżeli mniej niż K skupień $\rightarrow$ powrót do kroku 0

Uproszczenie zadania minimalizacji

Zamiast sumy W

$$\widetilde{W} = \sum_{k=1}^{K} \sum_{C(i)=k} d(\mathbf{x}_i, \mathbf{m}_k) = \sum_{i=1}^{N} d(\mathbf{x}_i, \mathbf{m}_{C(i)})$$

Algorytm K-średnich (K-means)

0. inicjalizacja początkowych K środków $\mathbf{m}_K$,
1. w pierwszym kroku przypisujemy punkty do najbliższych środków $\mathbf{m}_k$
   - jeżeli mniej niż K skupień $\rightarrow$ powrót do kroku 0
2. obliczamy nowe środki skupień i wracamy do kroku 1

Uproszczenie zadania minimalizacji

Zamiast sumy W

$$\widetilde{W} = \sum_{k=1}^{K} \sum_{C(i)=k} d(\mathbf{x}_i, \mathbf{m}_k) = \sum_{i=1}^{N} d(\mathbf{x}_i, \mathbf{m}_{C(i)})$$

Algorytm K-średnich (K-means)

0. inicjalizacja początkowych K środków $\mathbf{m}_K$,
1. w pierwszym kroku przypisujemy punkty do najbliższych środków $\mathbf{m}_k$
   - jeżeli mniej niż K skupień $\rightarrow$ powrót do kroku 0
2. obliczamy nowe środki skupień i wracamy do kroku 1
3. kontynuujemy iteracje, dopóki żaden punkt nie przeniesie się z jednego skupienia do drugiego

Uproszczenie zadania minimalizacji

Zamiast sumy W

$$\widetilde{W} = \sum_{k=1}^{K} \sum_{C(i)=k} d(\mathbf{x}_i, \mathbf{m}_k) = \sum_{i=1}^{N} d(\mathbf{x}_i, \mathbf{m}_{C(i)})$$

Algorytm K-średnich (K-means)

0. inicjalizacja początkowych K środków $\mathbf{m}_K$,
1. w pierwszym kroku przypisujemy punkty do najbliższych środków $\mathbf{m}_k$
   - jeżeli mniej niż K skupień $\rightarrow$ powrót do kroku 0
2. obliczamy nowe środki skupień i wracamy do kroku 1
3. kontynuujemy iteracje, dopóki żaden punkt nie przeniesie się z jednego skupienia do drugiego

Algorytm K-średnich - uwagi

Algorytm *K*-średnich - uwagi

- można też początkowo narzucić skupienia, a potem liczyć srodki

Algorytm *K*-średnich - uwagi

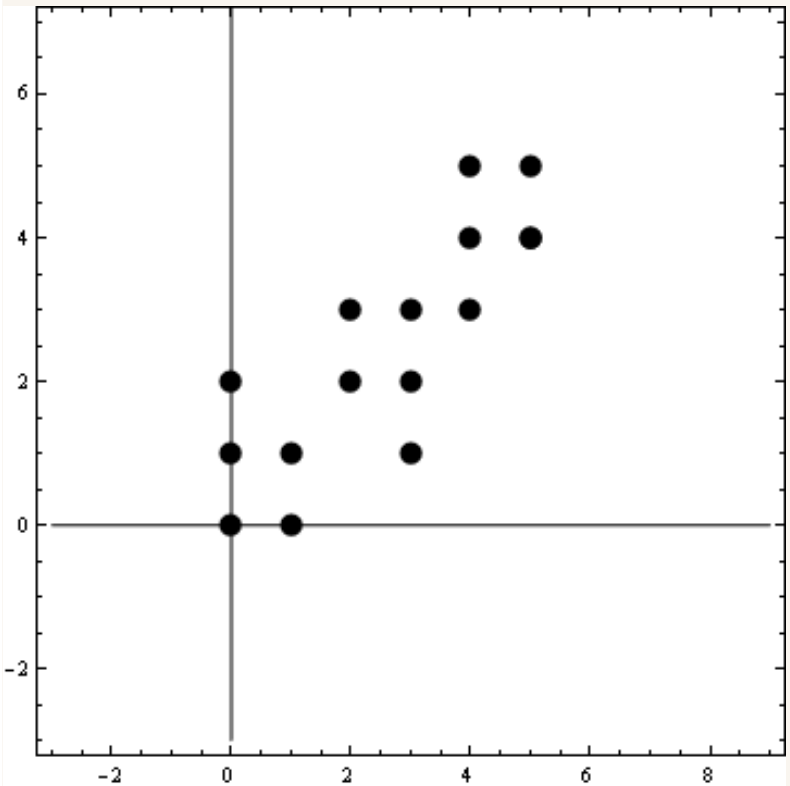
- można też początkowo narzucić skupienia, a potem liczyć srodki
- algorytmy są zbieżne, ale niekoniecznie do rozwiązania **globalnie optymalnego** — mogą to być lokalne minima,

Algorytm *K*-średnich - uwagi

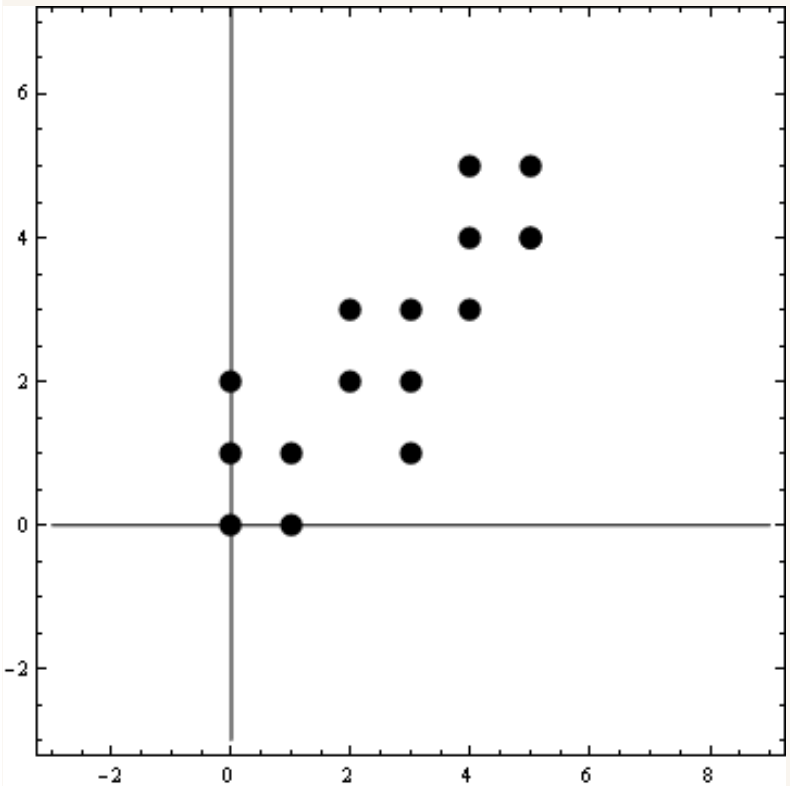
- można też początkowo narzucić skupienia, a potem liczyć srodki
- algorytmy są zbieżne, ale niekoniecznie do rozwiązania **globalnie optymalnego** — mogą to być lokalne minima,
- dlatego warto wielokrotnie stosować dany algoruym dla różnych warunków początkowych

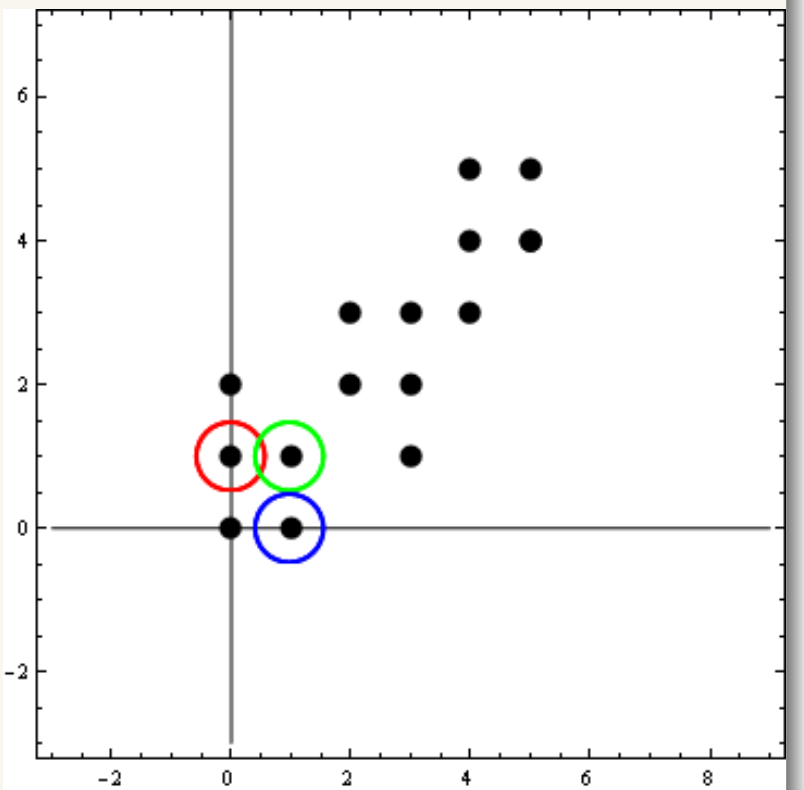
Przykład 1 - 3 klastry, inicjalizacja 1 (kroki 1-2)

Przykład 1 - 3 klastry, inicjalizacja 1 (kroki 1-2)

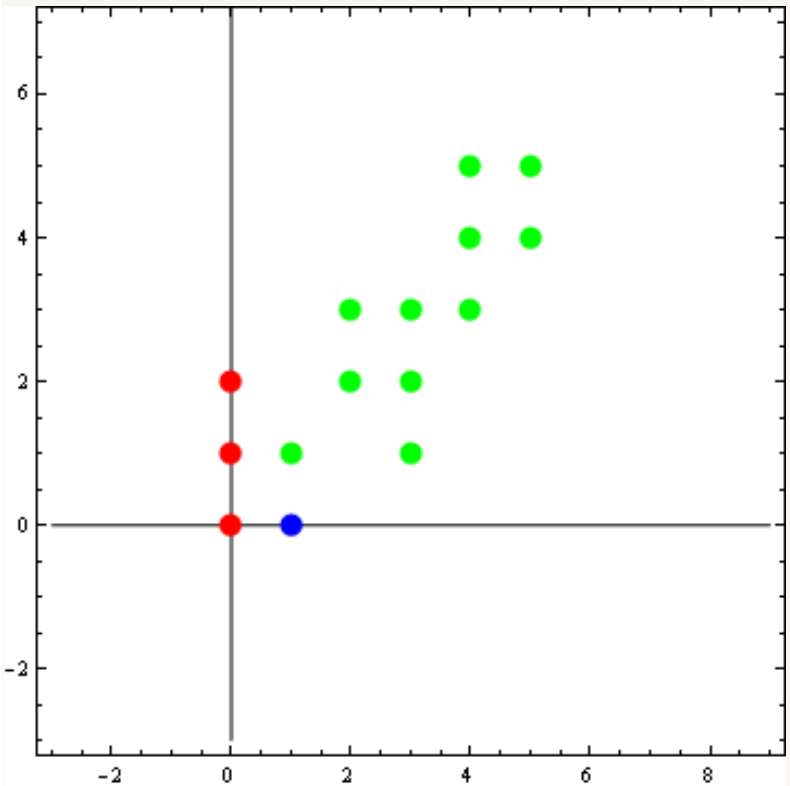


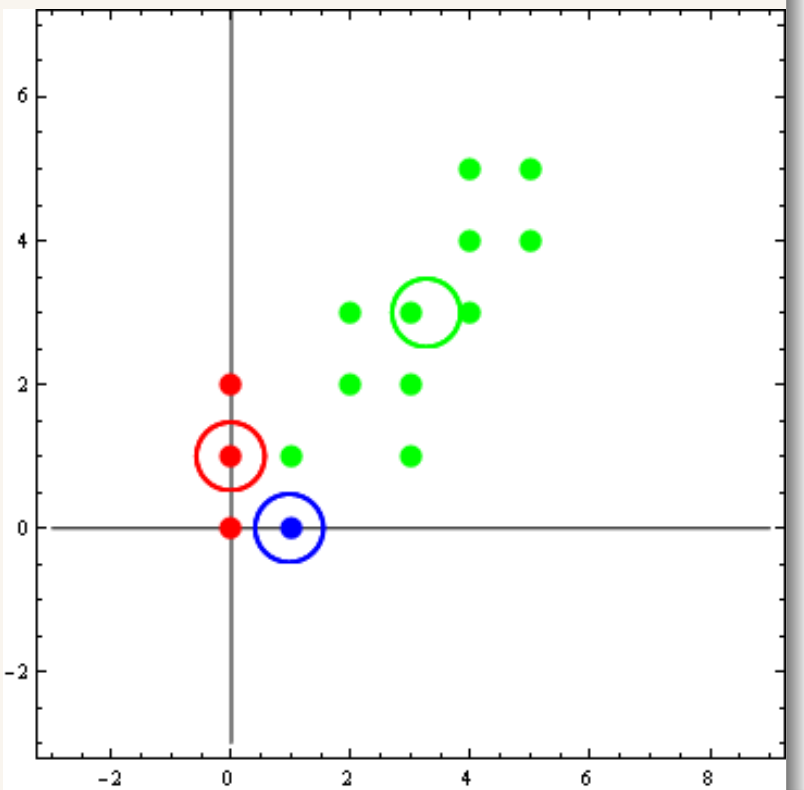
Przykład 1 - 3 klastry, inicjalizacja 1 (kroki 1-2)



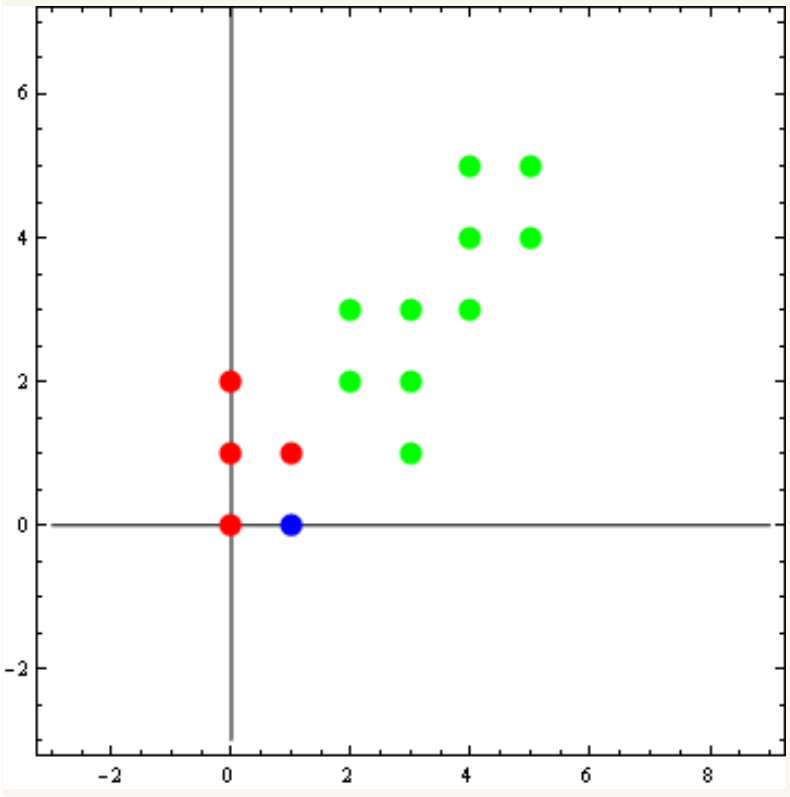


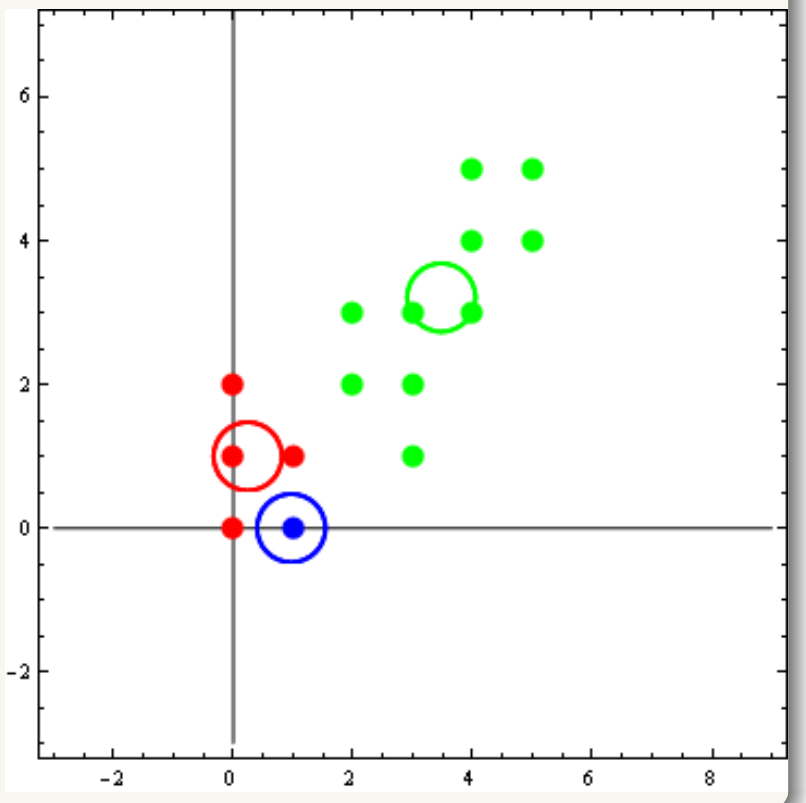
Przykład 1 - 3 klastry, inicjalizacja 1 (kroki 3-4)



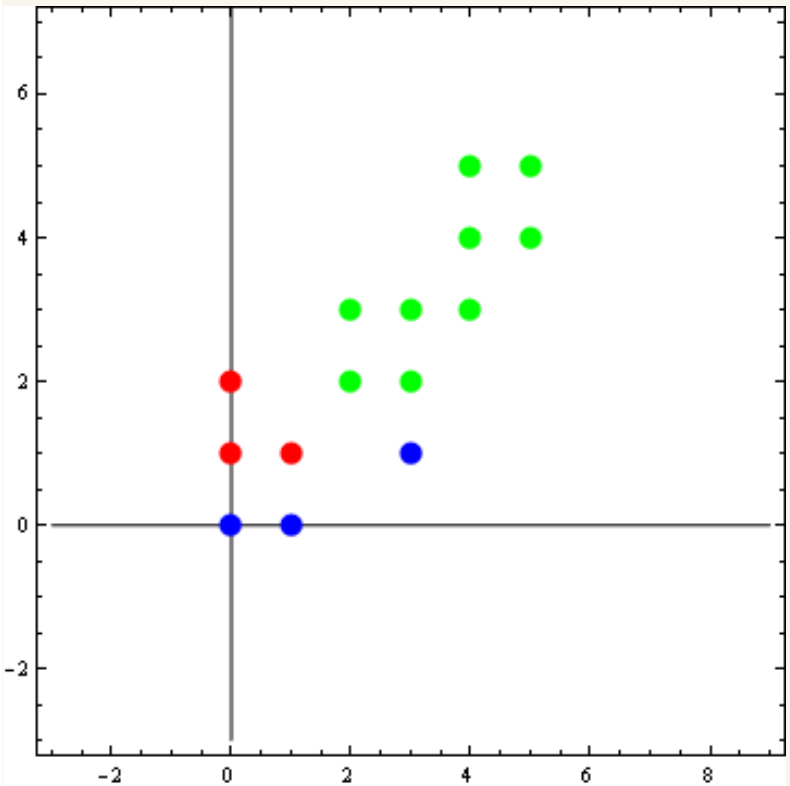


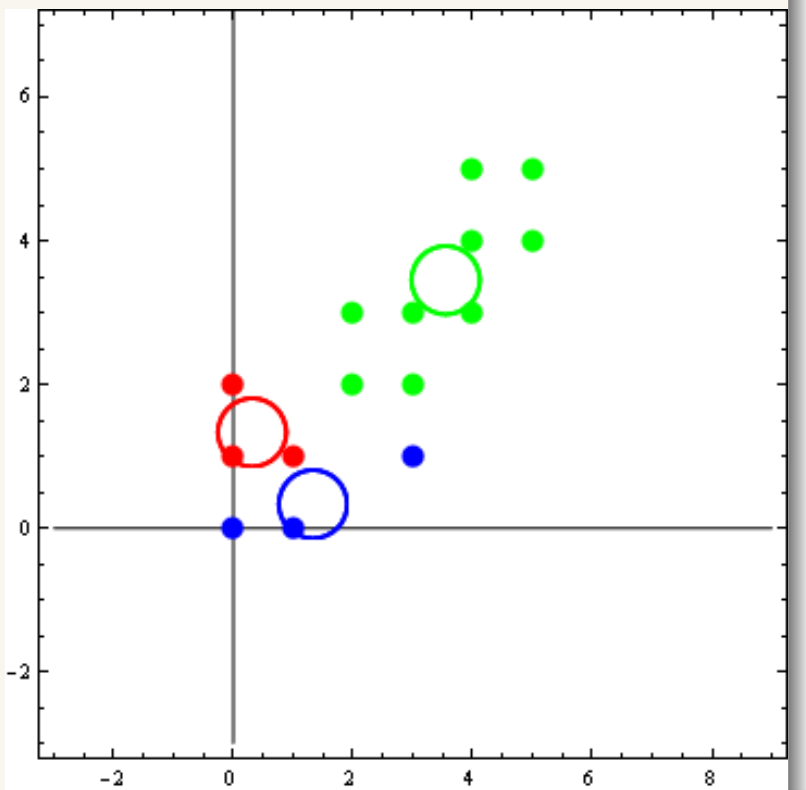
Przykład 1 - 3 klastry, inicjalizacja 1 (kroki 5-6)



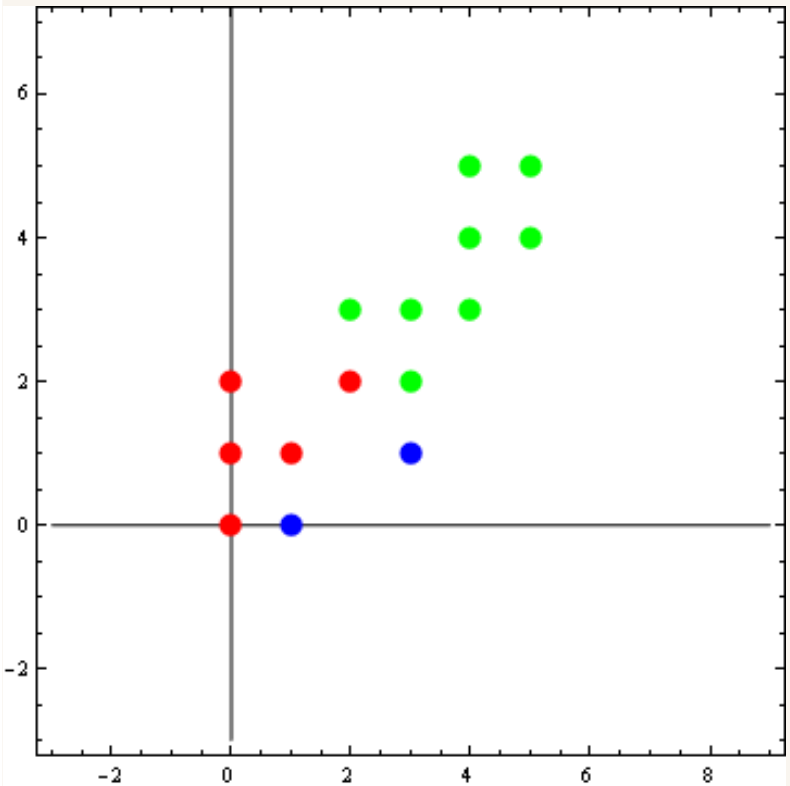


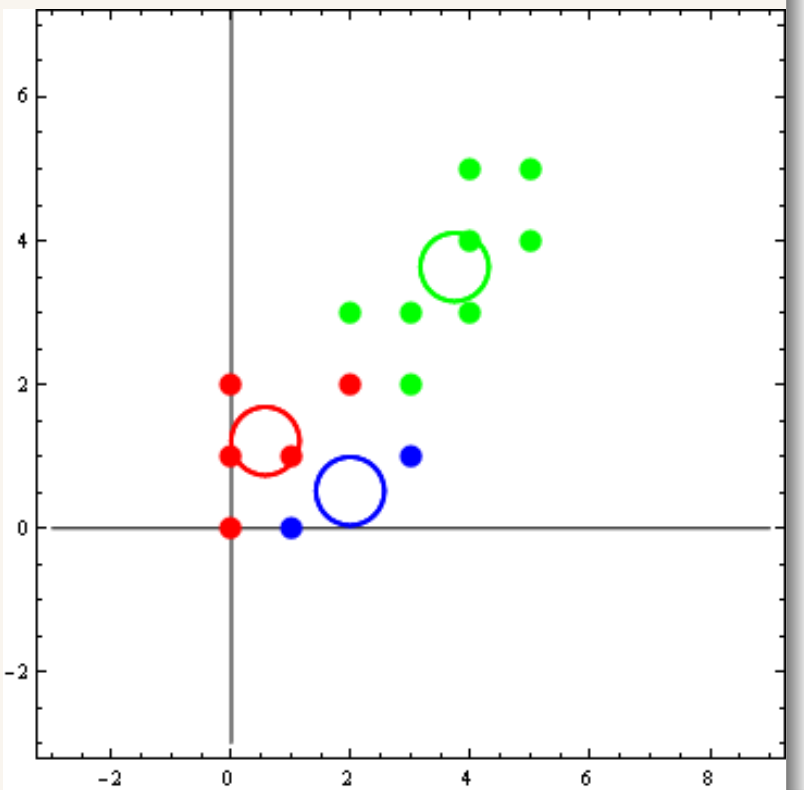
Przykład 1 - 3 klastry, inicjalizacja 1 (kroki 7-8)



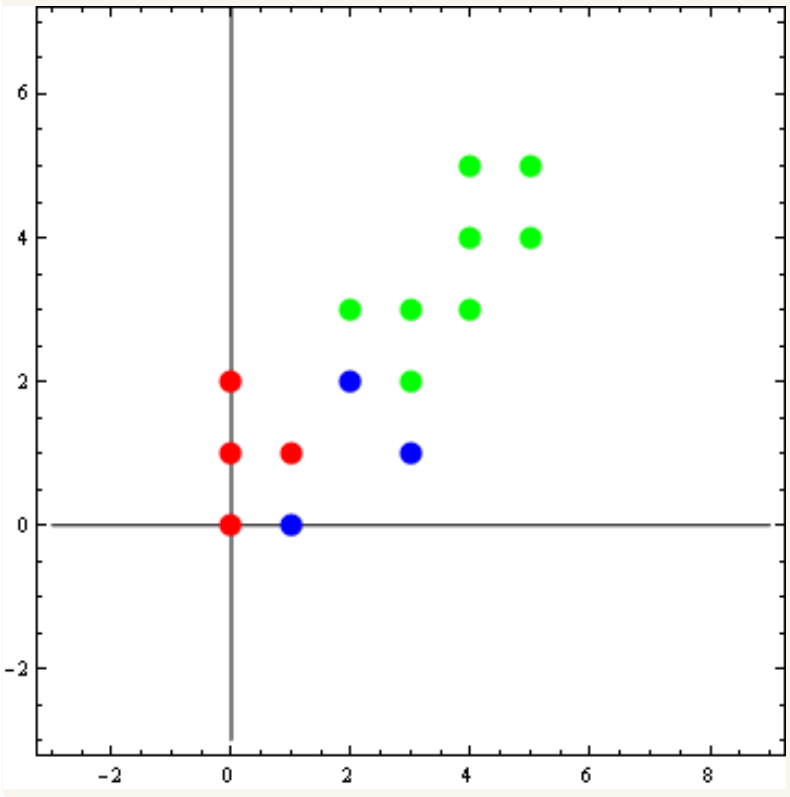


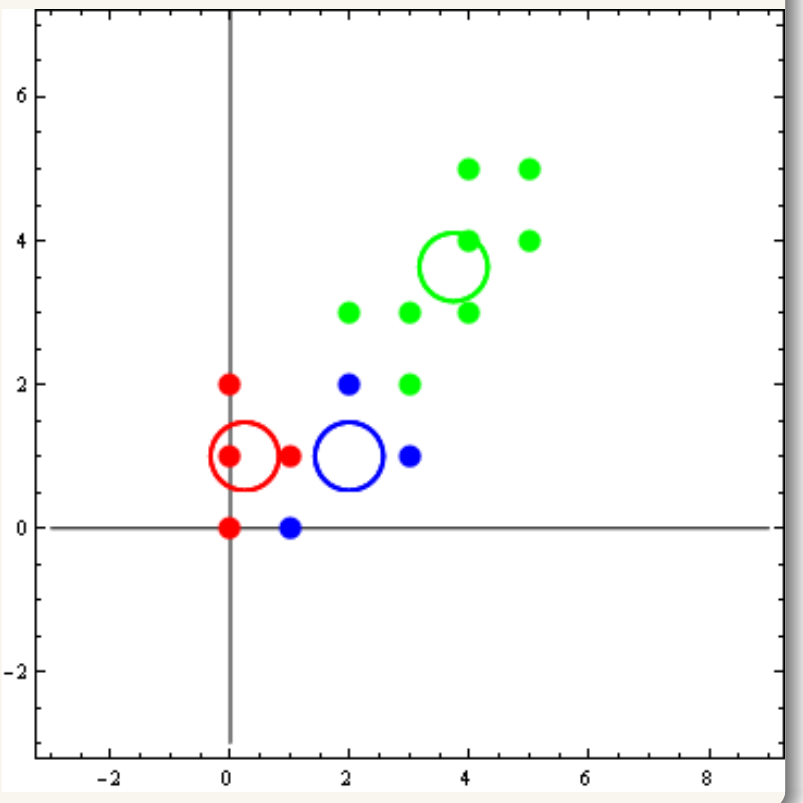
Przykład 1 - 3 klastry, inicjalizacja 1 (kroki 9-10)



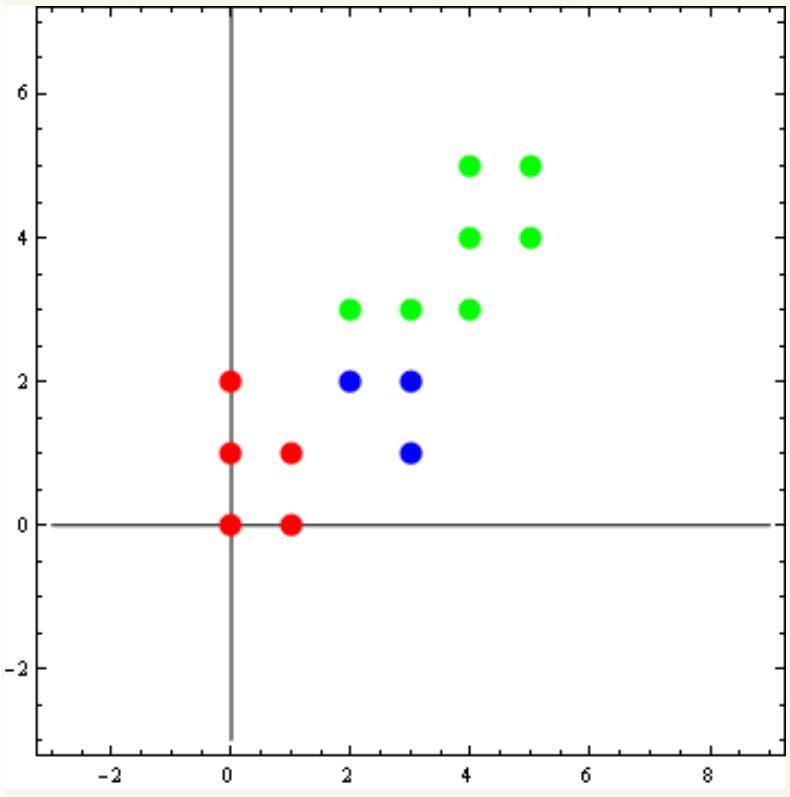


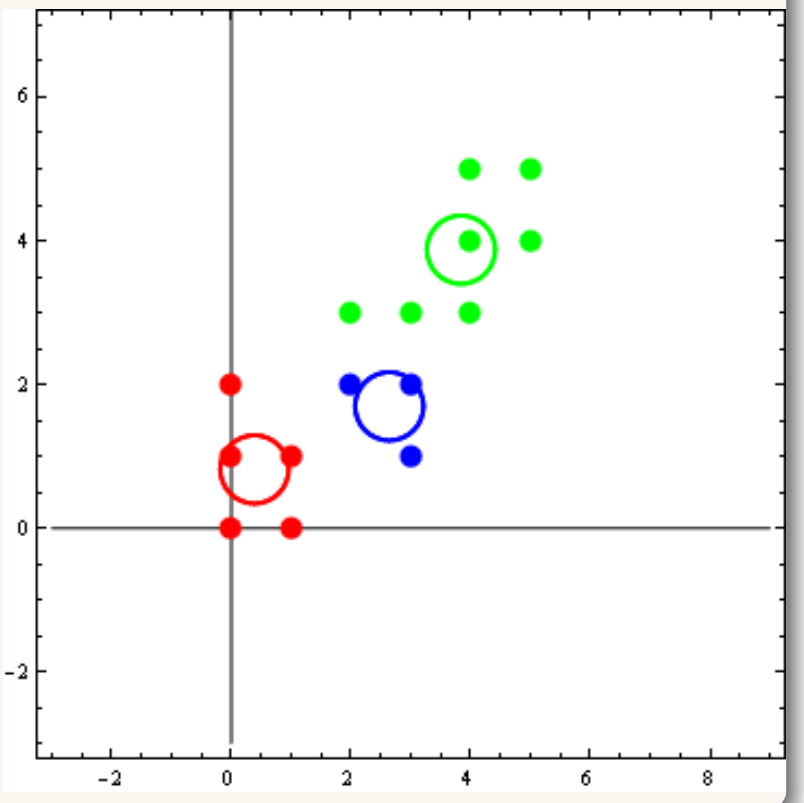
Przykład 1 - 3 klastry, inicjalizacja 1 (kroki 11-12)



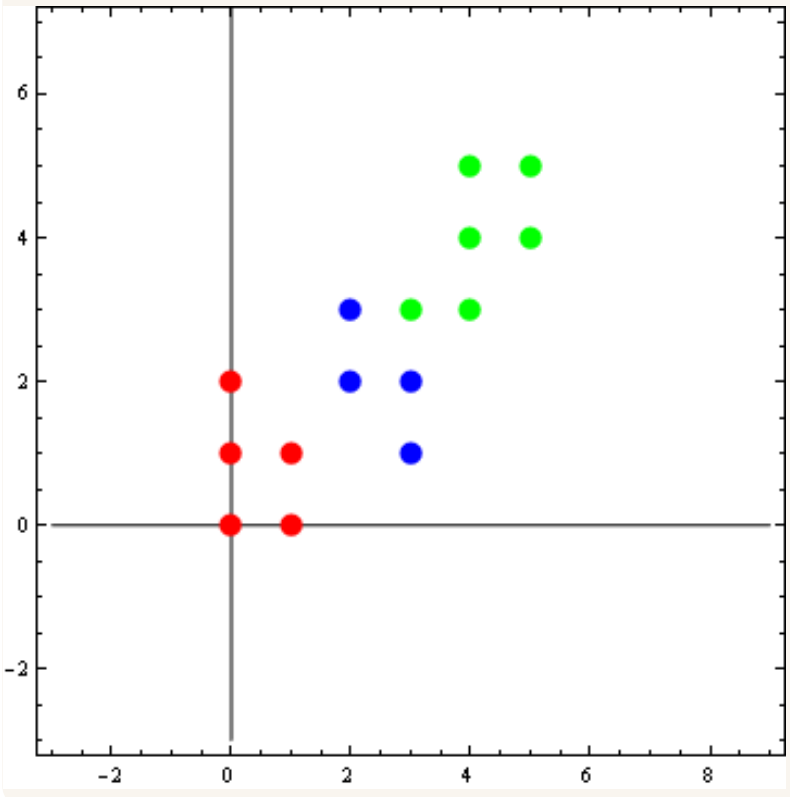


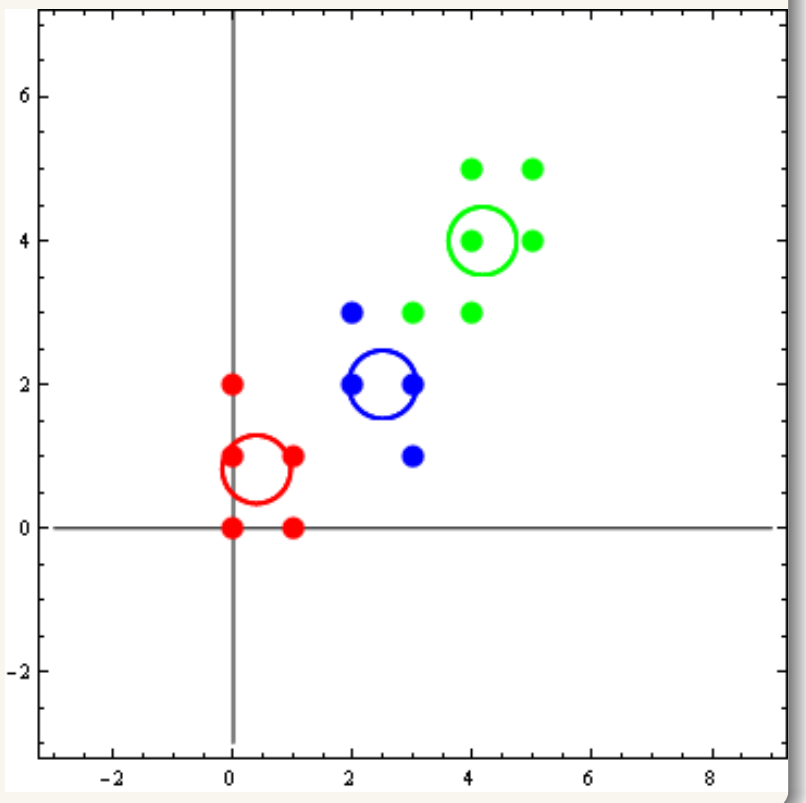
Przykład 1 - 3 klastry, inicjalizacja 1 (kroki 13-14)



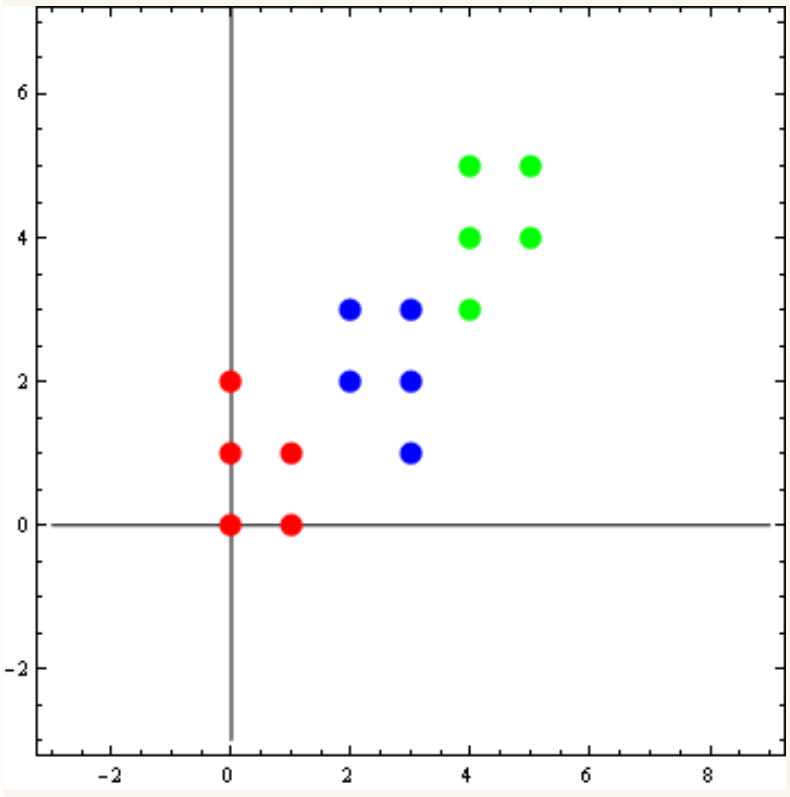


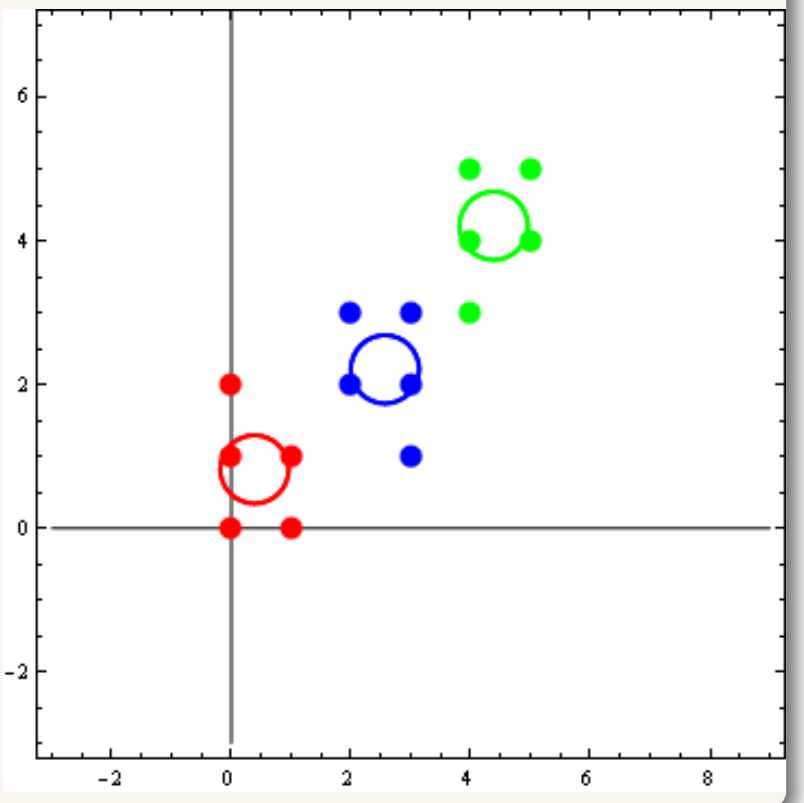
Przykład 1 - 3 klastry, inicjalizacja 1 (kroki 15-16)





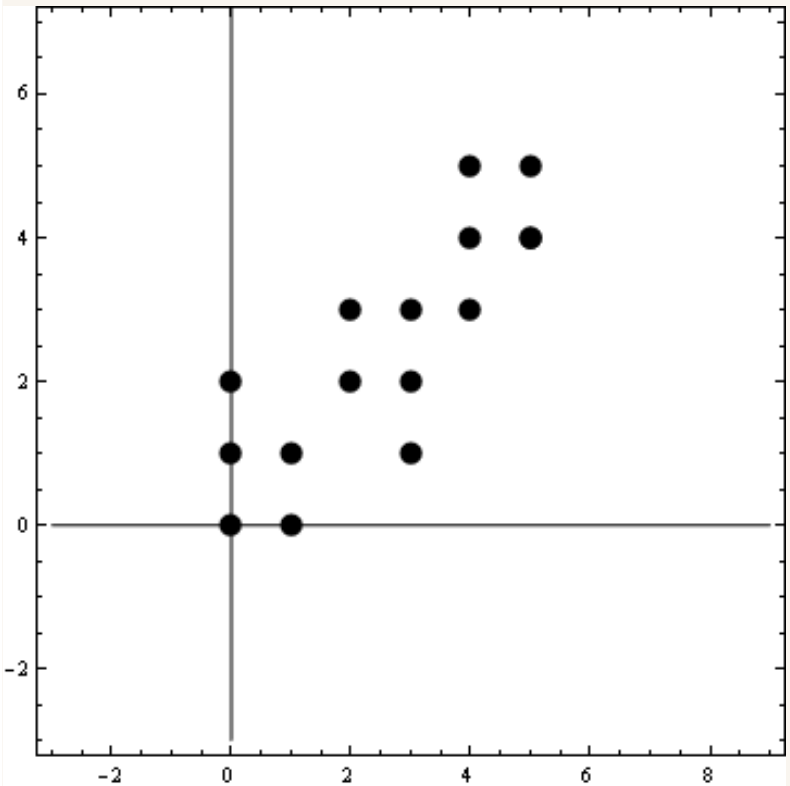
Przykład 1 - 3 klastry, inicjalizacja 1 (kroki 17-18)



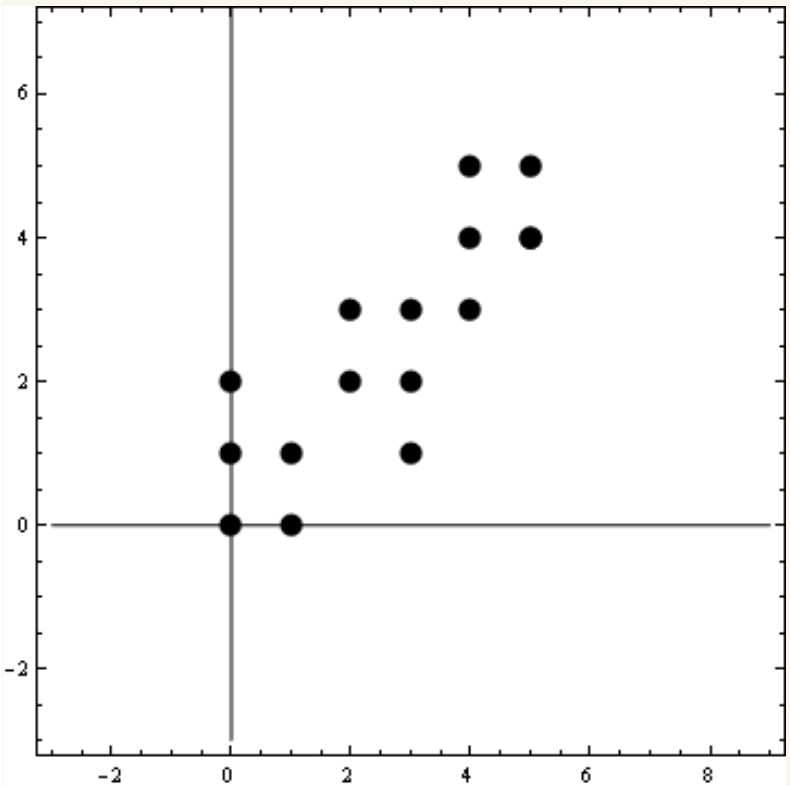


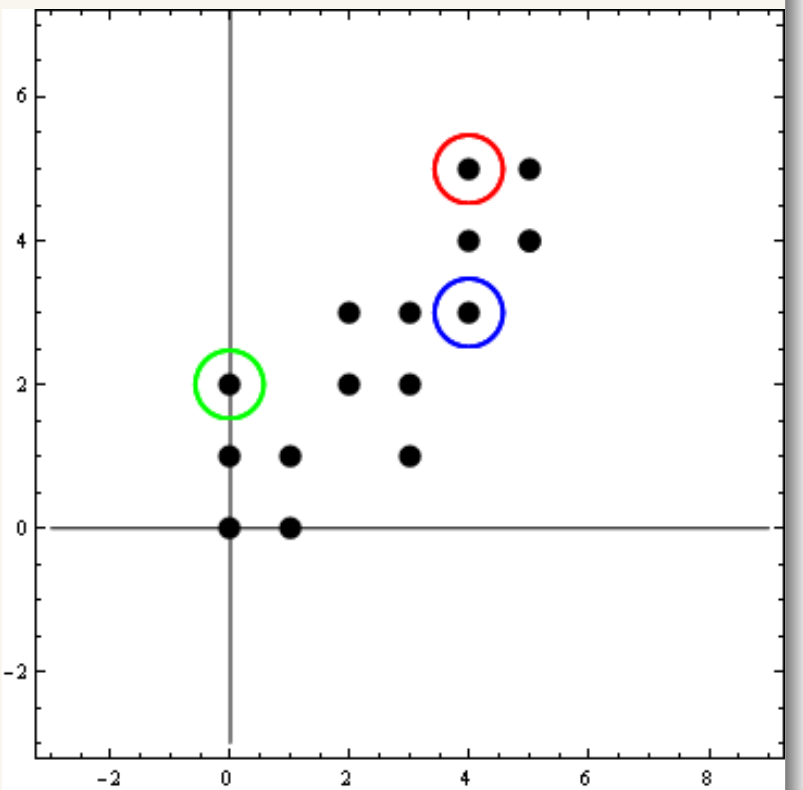
Przykład 2 - 3 klastry, inicjalizacja 2 (kroki 1-2)

Przykład 2 - 3 klastry, inicjalizacja 2 (kroki 1-2)

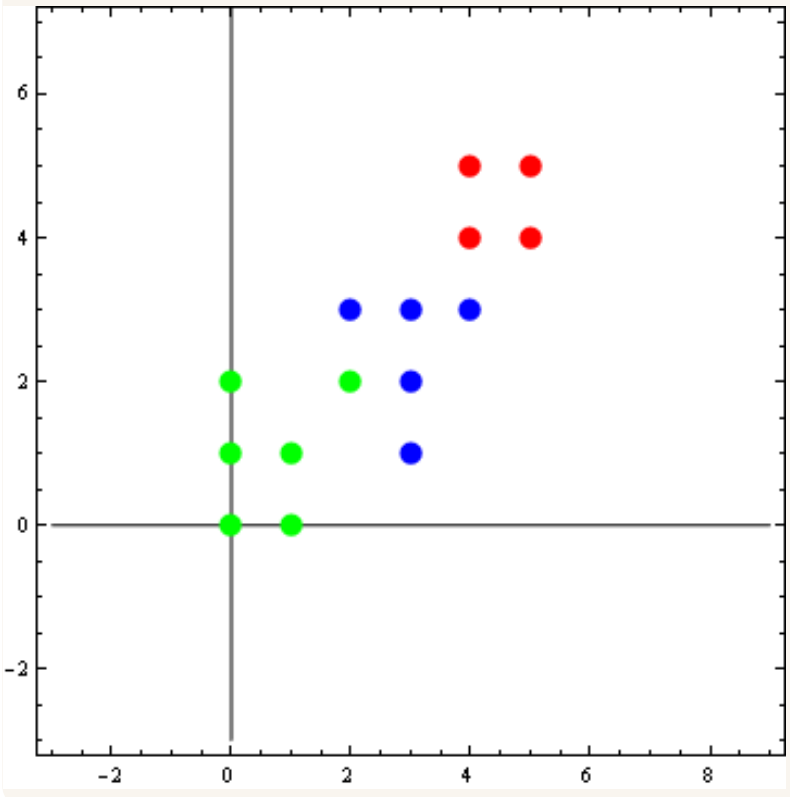


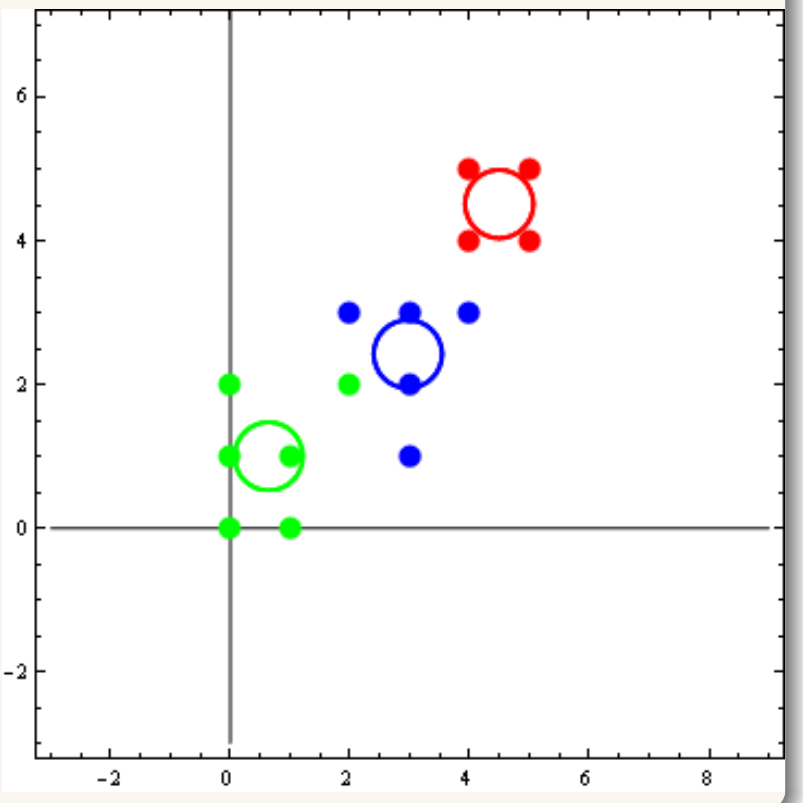
Przykład 2 - 3 klastry, inicjalizacja 2 (kroki 1-2)



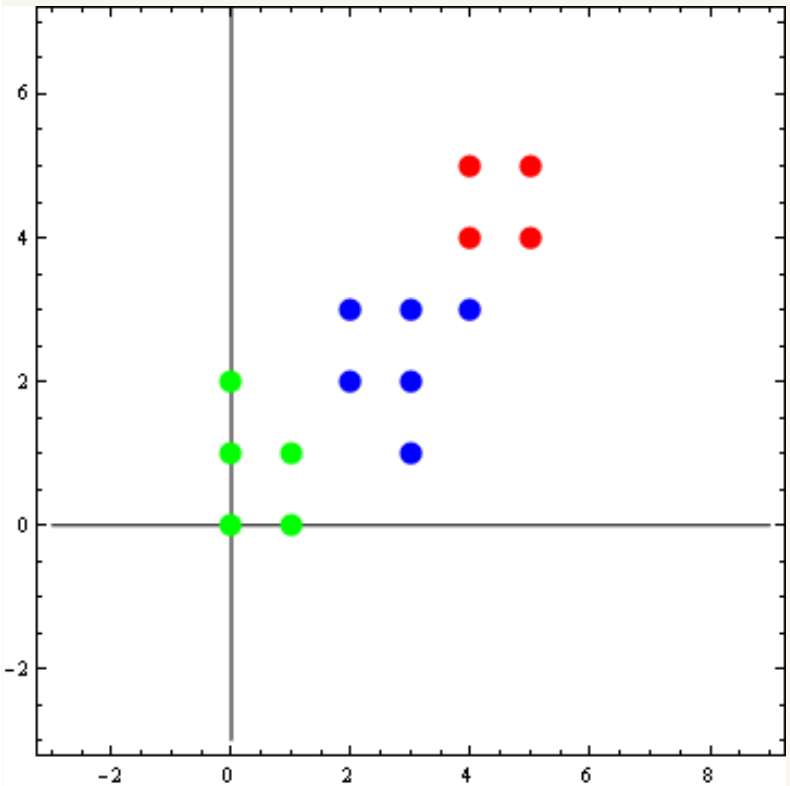


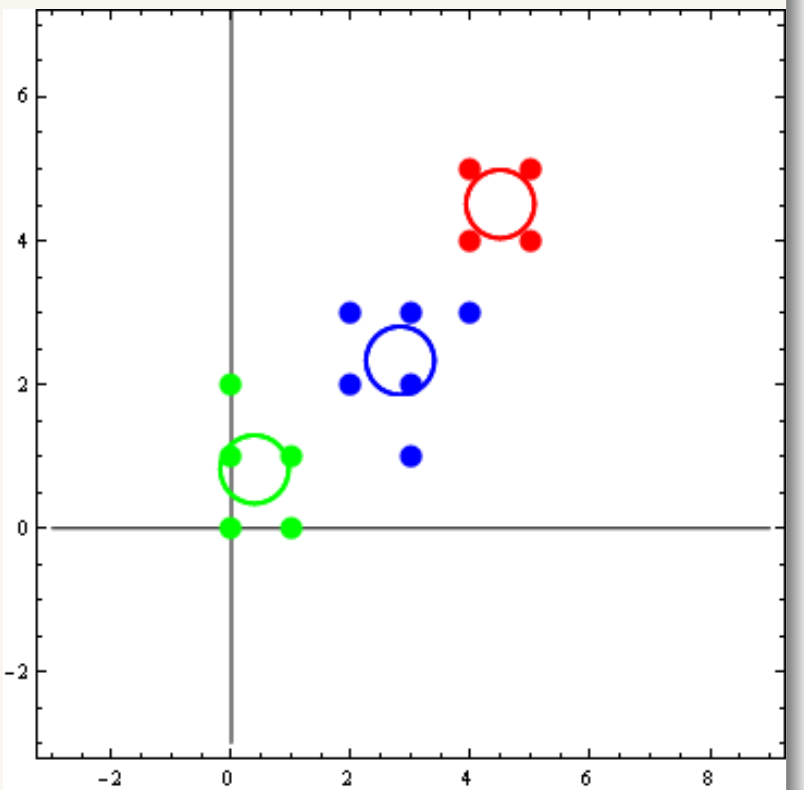
Przykład 2 - 3 klastry, inicjalizacja 2 (kroki 3-4)



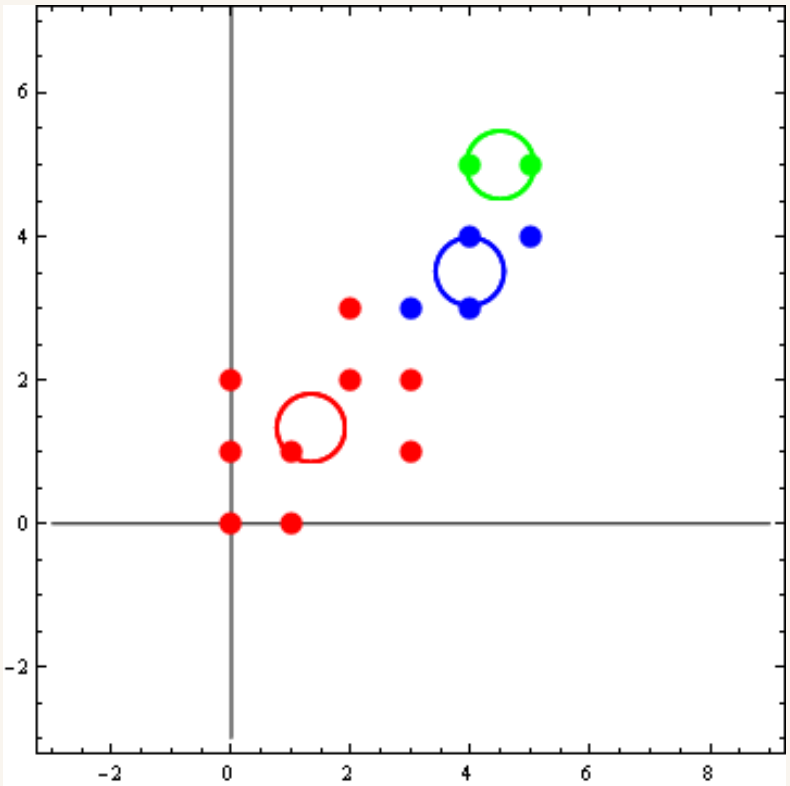


Przykład 2 - 3 klastry, inicjalizacja 2 (kroki 5-6)

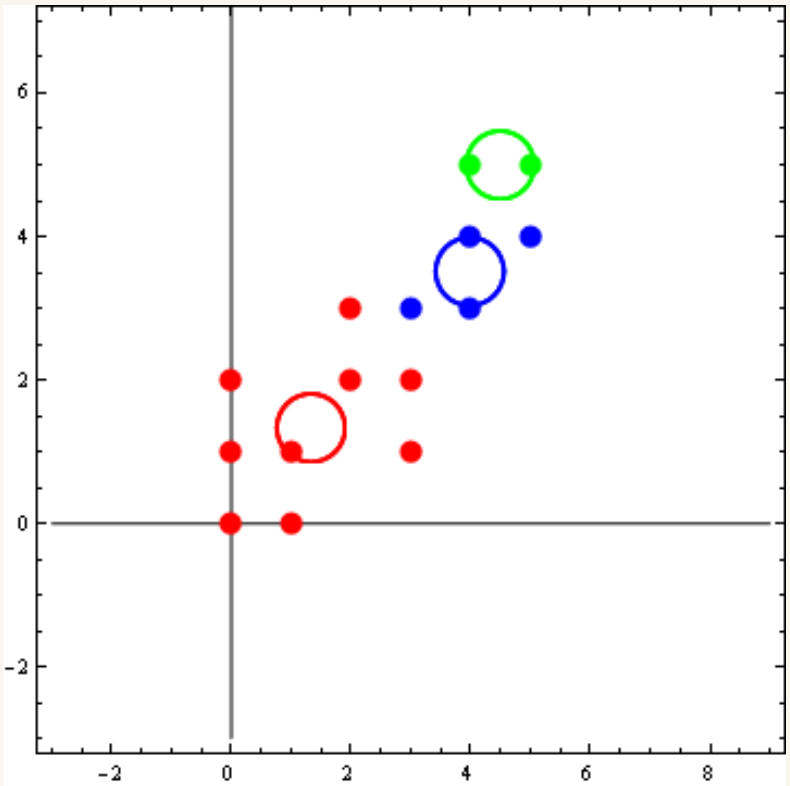




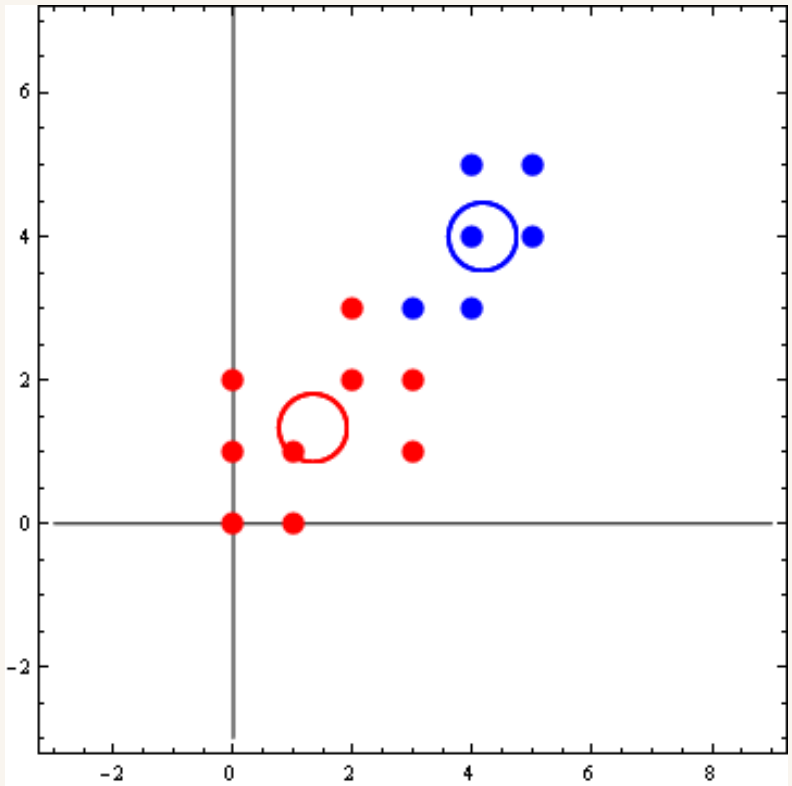
Przykład 3 - 3 klastry



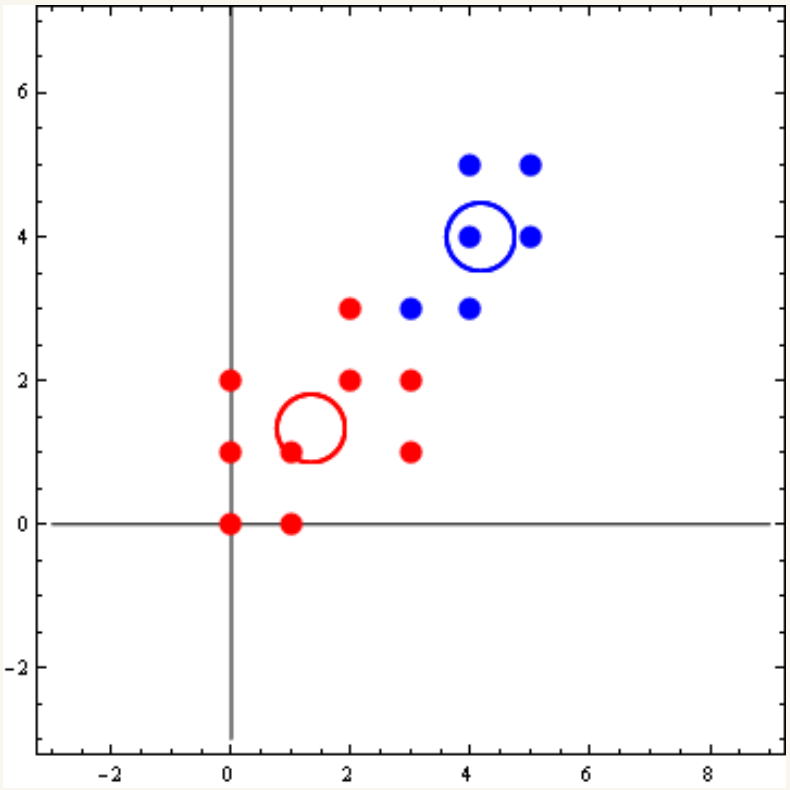
Przykład 3 - 3 klastry



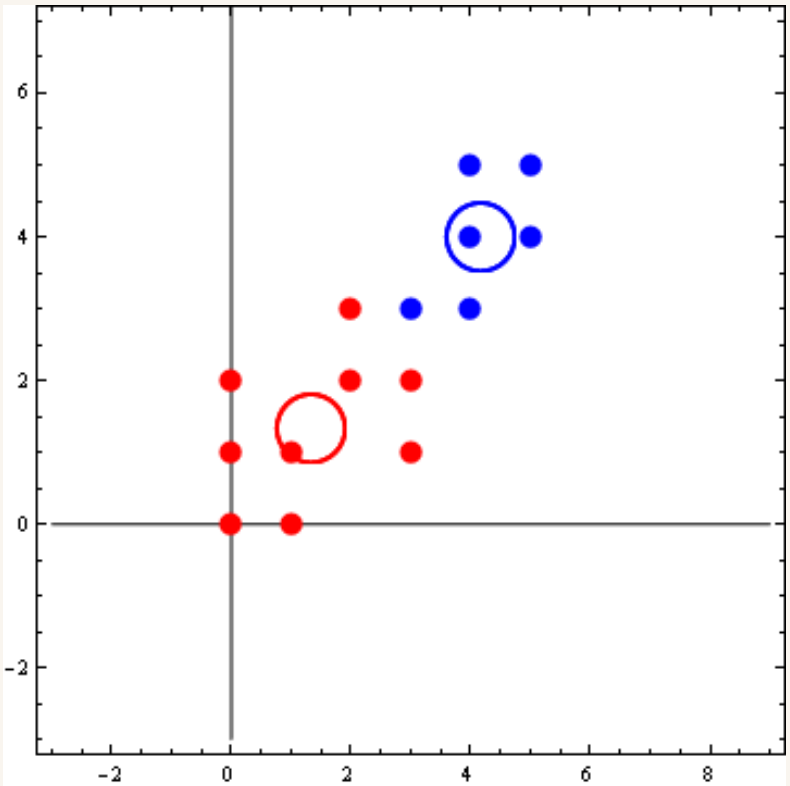
Przykład 4 - 2 klastry



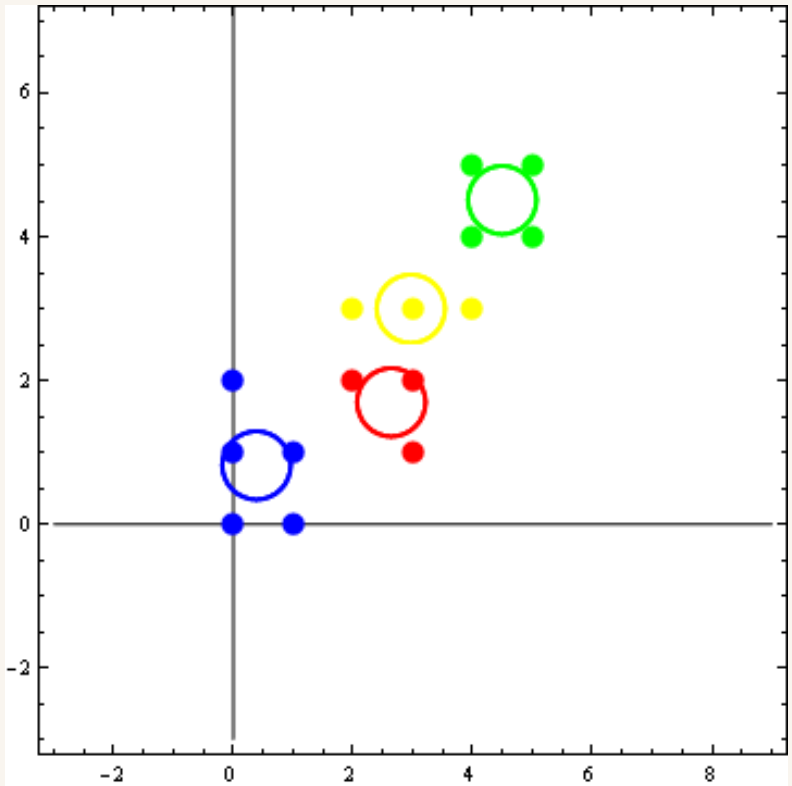
Przykład 5 - 4 klastry



Przykład 5 - 4 klastry



Przykład 6 - 5 klastrów



Optymalna liczba skupień

- w przypadku algorytmów K-means konieczne jest podanie z góry liczby skupień K,

Optymalna liczba skupień

- w przypadku algorytmów K-means konieczne jest podanie z góry liczby skupień K,
- często faktycznie K jest narzucone poprzez typ problemu np. firma może zatrudnić K sprzedawców i należy rozdzielić bazę danych klientów tak, aby klienci byli jak najbardziej podobni

Optymalna liczba skupień

- w przypadku algorytmów K-means konieczne jest podanie z góry liczby skupień K,
- często faktycznie K jest narzucone poprzez typ problemu np. firma może zatrudnić K sprzedawców i należy rozdzielić bazę danych klientów tak, aby klienci byli jak najbardziej podobni
- często jednak konieczne jest optymalne rozdzielenie obserwacji na skupienia i tym samym wyestymowanie z danych optymalnej wartości $K^{\star}$,

Optymalna liczba skupień

- w przypadku algorytmów K-means konieczne jest podanie z góry liczby skupień K,
- często faktycznie K jest narzucone poprzez typ problemu np. firma może zatrudnić K sprzedawców i należy rozdzielić bazę danych klientów tak, aby klienci byli jak najbardziej podobni
- często jednak konieczne jest optymalne rozdzielenie obserwacji na skupienia i tym samym wyestymowanie z danych optymalnej wartości $K^{\star}$,
- może się wydawać, że odpowiednim wskaźnikiem jest $\widetilde{W}$...

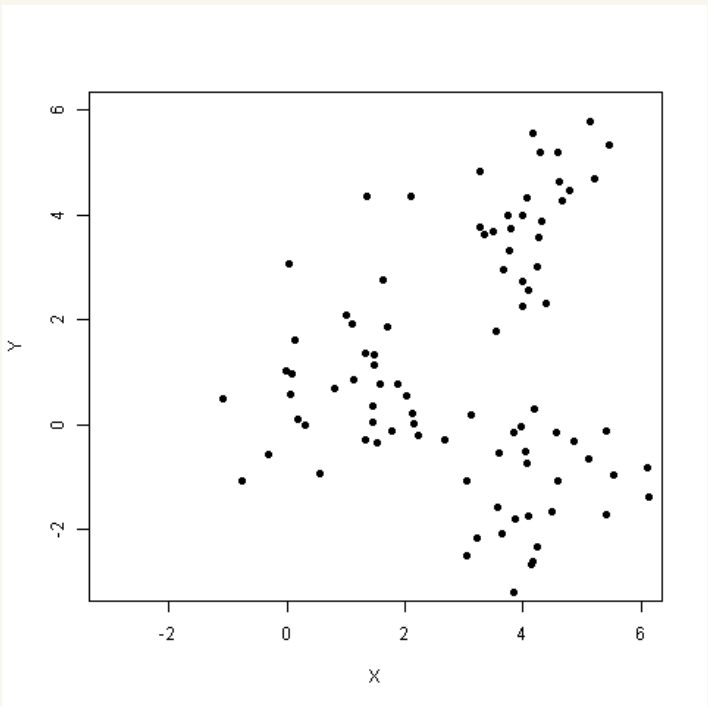
Optymalna liczba skupień

- ... ale $\widetilde{W}$ spada wraz z liczbą skupień,

Optymalna liczba skupień

- ... ale $\widetilde{W}$ spada wraz z liczbą skupień,
- dlaczego? istnieje coraz więcej środków, więc średnia odległość będzie się zmniejszała

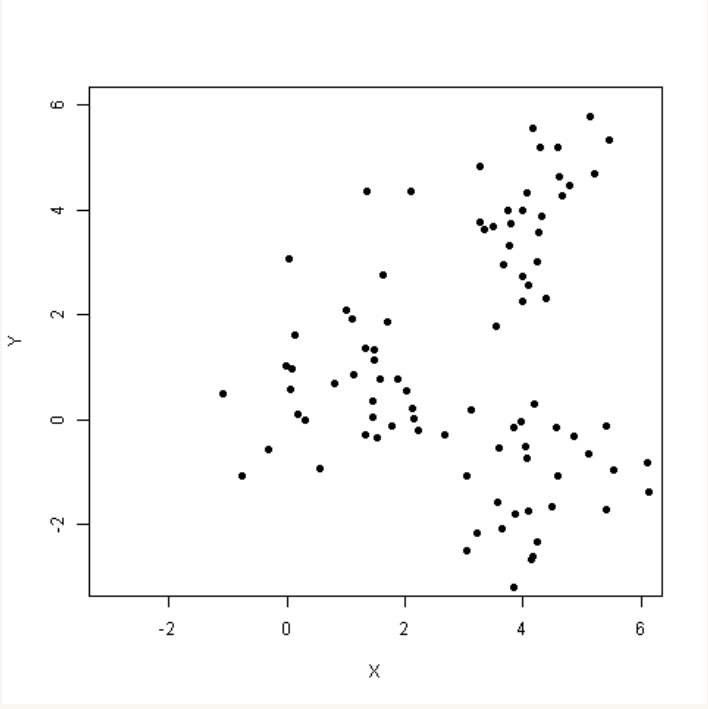
Dane (3 rozkłady Gaussa)



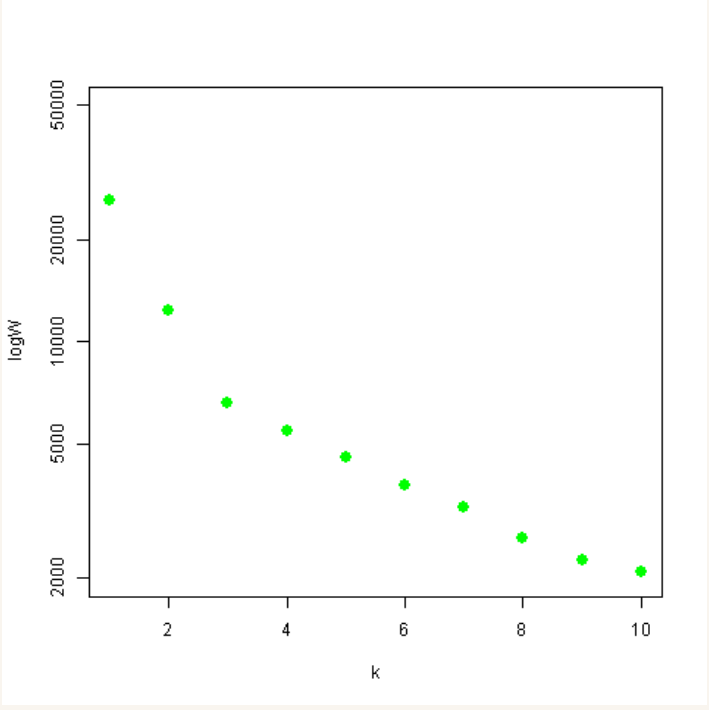
Optymalna liczba skupień

- ... ale $\widetilde{W}$ spada wraz z liczbą skupień,
- dlaczego? istnieje coraz więcej środków, więc średnia odległość będzie się zmniejszała

Dane (3 rozkłady Gaussa)



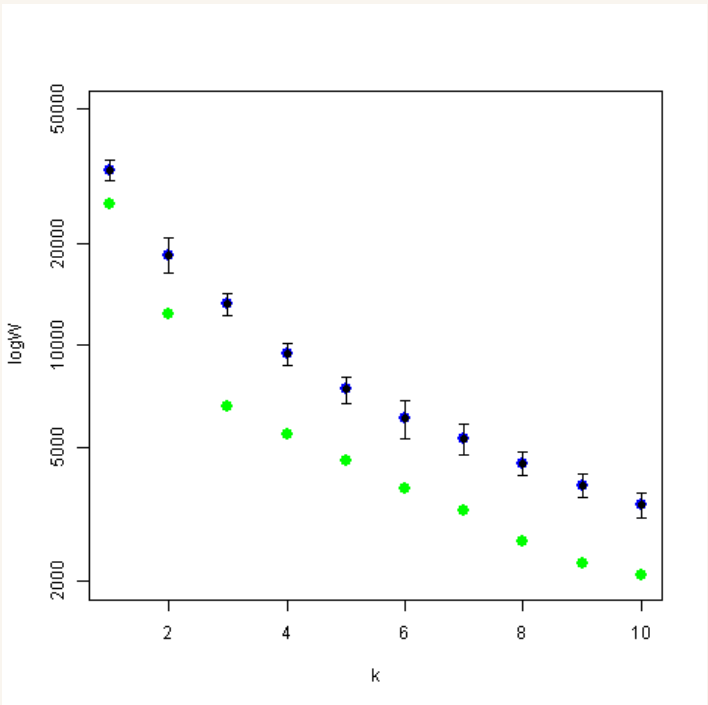
Wartośc log $\widetilde{W}(K)$



Optymalna liczba skupień

Konstruujemy kolejno rozwiązania z $K = 1, 2, ...$ skupieniami i zaprzestajemy z chwila, gdy różnica pomiędzy dwiema kolejnymi $\widetilde{W}$ przestaje być duża

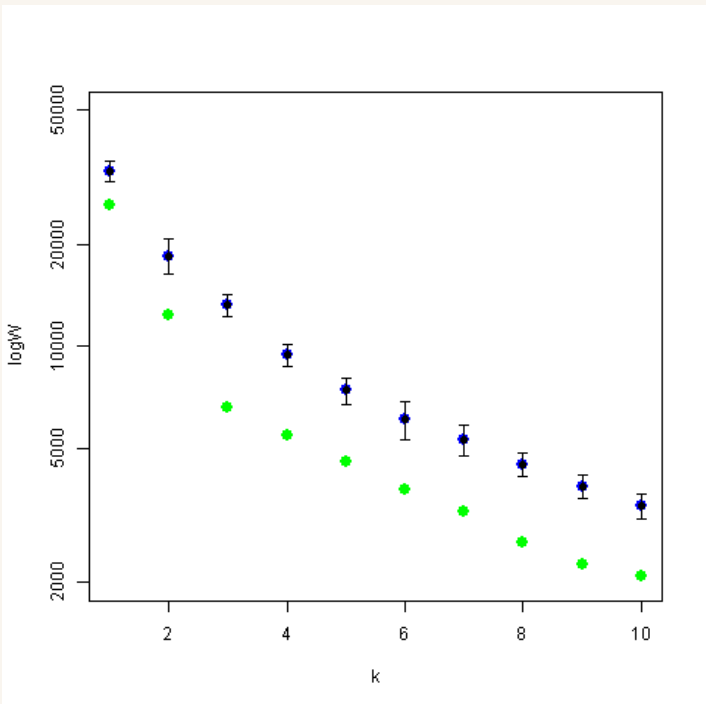
Wartośc log $\widetilde{W}(K)$ (obs. i teoret.)



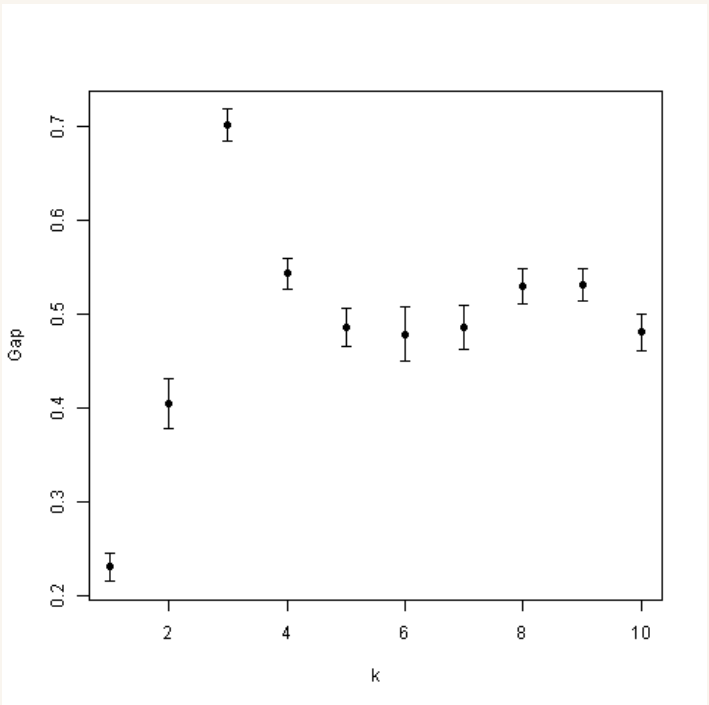
Optymalna liczba skupień

Konstruujemy kolejno rozwiązania z $K = 1, 2, \ldots$ skupieniami i zaprzestajemy z chwila, gdy różnica pomiędzy dwiema kolejnymi $\widetilde{W}$ przestaje być duża

Wartośc log $\widetilde{W}(K)$ (obs. i teoret.)



Statystyka odstępu



Co zrobić, gdy obserwacje **x** przyjmują atrybuty nieliczbowe?
Odległość można zastąpić **odmiennością**.

Dla danych binarnych

Co zrobić, gdy obserwacje **x** przyjmują atrybuty nieliczbowe?
Odległość można zastąpić **odmiennością**.

Dla danych binarnych

Każda z p składowych obserwacji i przyjmuje wartośc $x_i^{(k)} \in \{0, 1\}$.
Oznaczmy:

- a współrzednych ma tę samą wartość 1,

Co zrobić, gdy obserwacje **x** przyjmują atrybuty nieliczbowe?
Odległość można zastąpić **odmiennością**.

Dla danych binarnych

Każda z p składowych obserwacji i przyjmuje wartośc $x_i^{(k)} \in \{0, 1\}$.
Oznaczmy:

- a współrzednych ma tę samą wartość 1,
- b współrzednych ma cechę $x_i^{(k)} = 1$ i $x_j^{(k)} = 0$,

Co zrobić, gdy obserwacje **x** przyjmują atrybuty nieliczbowe?
Odległość można zastąpić **odmiennością**.

Dla danych binarnych

Każda z p składowych obserwacji i przyjmuje wartośc $x_i^{(k)} \in \{0, 1\}$.
Oznaczmy:

- a współrzednych ma tę samą wartość 1,
- b współrzednych ma cechę $x_i^{(k)} = 1$ i $x_j^{(k)} = 0$,
- c współrzednych ma cechę $x_i^{(k)} = 0$ i $x_j^{(k)} = 1$,

Co zrobić, gdy obserwacje **x** przyjmują atrybuty nieliczbowe?
Odległość można zastąpić **odmiennością**.

Dla danych binarnych

Każda z p składowych obserwacji i przyjmuje wartośc $x_i^{(k)} \in \{0, 1\}$.
Oznaczmy:

- a współrzednych ma tę samą wartość 1,
- b współrzednych ma cechę $x_i^{(k)} = 1$ i $x_j^{(k)} = 0$,
- c współrzednych ma cechę $x_i^{(k)} = 0$ i $x_j^{(k)} = 1$,
- d współrzednych ma tę samą wartość 0,

Co zrobić, gdy obserwacje **x** przyjmują atrybuty nieliczbowe?
Odległość można zastąpić **odmiennością**.

Dla danych binarnych

Każda z p składowych obserwacji i przyjmuje wartośc $x_i^{(k)} \in \{0, 1\}$.
Oznaczmy:

- a współrzednych ma tę samą wartość 1,
- b współrzednych ma cechę $x_i^{(k)} = 1$ i $x_j^{(k)} = 0$,
- c współrzednych ma cechę $x_i^{(k)} = 0$ i $x_j^{(k)} = 1$,
- d współrzednych ma tę samą wartość 0,

Odległość Hamminga

Co zrobić, gdy obserwacje **x** przyjmują atrybuty nieliczbowe?
Odległość można zastąpić **odmiennością**.

Dla danych binarnych

Każda z p składowych obserwacji i przyjmuje wartośc $x_i^{(k)} \in \{0, 1\}$.
Oznaczmy:

- a współrzednych ma tę samą wartość 1,
- b współrzednych ma cechę $x_i^{(k)} = 1$ i $x_j^{(k)} = 0$,
- c współrzednych ma cechę $x_i^{(k)} = 0$ i $x_j^{(k)} = 1$,
- d współrzednych ma tę samą wartość 0,

Odległość Hamminga

Odległość Hamminga jest równa liczbie współprzędnych, których wartości są dla obydwu obserwacji różne ($d_{ij} = b + c$).

Co zrobić, gdy obserwacje **x** przyjmują atrybuty nieliczbowe?
Odległość można zastąpić **odmiennością**.

Dla danych binarnych

Każda z p składowych obserwacji i przyjmuje wartośc $x_i^{(k)} \in \{0, 1\}$.
Oznaczmy:

- a współrzednych ma tę samą wartość 1,
- b współrzednych ma cechę $x_i^{(k)} = 1$ i $x_j^{(k)} = 0$,
- c współrzednych ma cechę $x_i^{(k)} = 0$ i $x_j^{(k)} = 1$,
- d współrzednych ma tę samą wartość 0,

Odległość Hamminga

Odległość Hamminga jest równa liczbie współprzędnych, których wartości są dla obydwu obserwacji różne ($d_{ij} = b + c$).

$\mathbf{x}_i$	1	1	0	1	0	1	0	

Co zrobić, gdy obserwacje **x** przyjmują atrybuty nieliczbowe?
Odległość można zastąpić **odmiennością**.

Dla danych binarnych

Każda z p składowych obserwacji i przyjmuje wartośc $x_i^{(k)} \in \{0, 1\}$.
Oznaczmy:

- a współrzednych ma tę samą wartość 1,
- b współrzednych ma cechę $x_i^{(k)} = 1$ i $x_j^{(k)} = 0$,
- c współrzednych ma cechę $x_i^{(k)} = 0$ i $x_j^{(k)} = 1$,
- d współrzednych ma tę samą wartość 0,

Odległość Hamminga

Odległość Hamminga jest równa liczbie współprzędnych, których wartości są dla obydwu obserwacji różne ($d_{ij} = b + c$).

$\mathbf{x}_i$	1	1	0	1	0	1	0
$\mathbf{x}_j$	1	0	1	0	0	1	1

Co zrobić, gdy obserwacje **x** przyjmują atrybuty nieliczbowe?
Odległość można zastąpić **odmiennością**.

Dla danych binarnych

Każda z p składowych obserwacji i przyjmuje wartośc $x_i^{(k)} \in \{0, 1\}$.
Oznaczmy:

- a współrzednych ma tę samą wartość 1,
- b współrzednych ma cechę $x_i^{(k)} = 1$ i $x_j^{(k)} = 0$,
- c współrzednych ma cechę $x_i^{(k)} = 0$ i $x_j^{(k)} = 1$,
- d współrzednych ma tę samą wartość 0,

Odległość Hamminga

Odległość Hamminga jest równa liczbie współprzędnych, których wartości są dla obydwu obserwacji różne ($d_{ij} = b + c$).

$\mathbf{x}_i$	1	1	0	1	0	1	0	
$\mathbf{x}_j$	1	0	1	0	0	1	1	
		x	x	x			x	

Co zrobić, gdy obserwacje **x** przyjmują atrybuty nieliczbowe?
Odległość można zastąpić **odmiennością**.

Dla danych binarnych

Każda z p składowych obserwacji i przyjmuje wartośc $x_i^{(k)} \in \{0, 1\}$.
Oznaczmy:

- a współrzednych ma tę samą wartość 1,
- b współrzednych ma cechę $x_i^{(k)} = 1$ i $x_j^{(k)} = 0$,
- c współrzednych ma cechę $x_i^{(k)} = 0$ i $x_j^{(k)} = 1$,
- d współrzednych ma tę samą wartość 0,

Odległość Hamminga

Odległość Hamminga jest równa liczbie współprzędnych, których wartości są dla obydwu obserwacji różne ($d_{ij} = b + c$).

$\mathbf{x}_i$	1	1	0	1	0	1	0	
$\mathbf{x}_j$	1	0	1	0	0	1	1	
		x	x	x			x	$\mathbf{d_{ij} = 4}$

Unormowana odległość Hamminga

Unormowana odległość Hamminga

Unormowana odległość Hamminga (współczynnik dopasowania):

Unormowana odległość Hamminga

Unormowana odległość Hamminga (współczynnik dopasowania):

$$\hat{d}_{ij} = \frac{b+c}{p} = 1 - \frac{a+d}{p}$$

W poprzednim przykładzie $\hat{d}_{ij} = \frac{4}{7}$.

Unormowana odległość Hamminga

Unormowana odległość Hamminga (współczynnik dopasowania):

$$\hat{d}_{ij} = \frac{b+c}{p} = 1 - \frac{a+d}{p}$$

W poprzednim przykładzie $\hat{d}_{ij} = \frac{4}{7}$.

Współczynnik Jaccarda

Unormowana odległość Hamminga

Unormowana odległość Hamminga (współczynnik dopasowania):

$$\hat{d}_{ij} = \frac{b+c}{p} = 1 - \frac{a+d}{p}$$

W poprzednim przykładzie $\hat{d}_{ij} = \frac{4}{7}$.

Współczynnik Jaccarda

Współczynnik Jaccarda określa stopień odmienności jako

Unormowana odległość Hamminga

Unormowana odległość Hamminga (współczynnik dopasowania):

$$\hat{d}_{ij} = \frac{b+c}{p} = 1 - \frac{a+d}{p}$$

W poprzednim przykładzie $\hat{d}_{ij} = \frac{4}{7}$.

Współczynnik Jaccarda

Współczynnik Jaccarda określa stopień odmienności jako

$$d_{ij} = \frac{b+c}{a+b+c} = \frac{b+c}{p-d}$$

Unormowana odległość Hamminga

Unormowana odległość Hamminga (współczynnik dopasowania):

$$\hat{d}_{ij} = \frac{b+c}{p} = 1 - \frac{a+d}{p}$$

W poprzednim przykładzie $\hat{d}_{ij} = \frac{4}{7}$.

Współczynnik Jaccarda

Współczynnik Jaccarda określa stopień odmienności jako

$$d_{ij} = \frac{b+c}{a+b+c} = \frac{b+c}{p-d}$$

Innymi słowy ten współczynnik traktuje wartość 0 jako brak atrybutu — brak ten nie przyczynia się do lepszego rozróżnienia obiektów.

A co w przypadku danych mieszanych (np. liczbowych i jakościowych)?

A co w przypadku danych mieszanych (np. liczbowych i jakościowych)?

Współczynniki Gowera

A co w przypadku danych mieszanych (np. liczbowych i jakościowych)?

Współczynniki Gowera

Aby je policzyć:

A co w przypadku danych mieszanych (np. liczbowych i jakościowych)?

Współczynniki Gowera

Aby je policzyć:

- należy najpierw określić współczynnik podobieństwa oddzielnie dla każdej składowej wektora,

A co w przypadku danych mieszanych (np. liczbowych i jakościowych)?

Współczynniki Gowera

Aby je policzyć:

- należy najpierw określić współczynnik podobieństwa oddzielnie dla każdej składowej wektora,
- przyjmuje się, że wartości mogą być nieporównywalne gdy

A co w przypadku danych mieszanych (np. liczbowych i jakościowych)?

Współczynniki Gowera

Aby je policzyć:

- należy najpierw określić współczynnik podobieństwa oddzielnie dla każdej składowej wektora,
- przyjmuje się, że wartości mogą być nieporównywalne gdy
  1. brakuje wartości w jednym z obiektów,

A co w przypadku danych mieszanych (np. liczbowych i jakościowych)?

Współczynniki Gowera

Aby je policzyć:

- należy najpierw określić współczynnik podobieństwa oddzielnie dla każdej składowej wektora,
- przyjmuje się, że wartości mogą być nieporównywalne gdy
  1. brakuje wartości w jednym z obiektów,
  2. zmienna jest binarna i nie występuje przynajmniej w jednym z obiektów,

A co w przypadku danych mieszanych (np. liczbowych i jakościowych)?

Współczynniki Gowera

Aby je policzyć:

- należy najpierw określić współczynnik podobieństwa oddzielnie dla każdej składowej wektora,
- przyjmuje się, że wartości mogą być nieporównywalne gdy
  1. brakuje wartości w jednym z obiektów,
  2. zmienna jest binarna i nie występuje przynajmniej w jednym z obiektów,
- zarówno sam współczynnik s_{ij} jak i jego wartości cząstkowe s_{ijk} są unormowane,

A co w przypadku danych mieszanych (np. liczbowych i jakościowych)?

Współczynniki Gowera

Aby je policzyć:

- należy najpierw określić współczynnik podobieństwa oddzielnie dla każdej składowej wektora,
- przyjmuje się, że wartości mogą być nieporównywalne gdy
  1. brakuje wartości w jednym z obiektów,
  2. zmienna jest binarna i nie występuje przynajmniej w jednym z obiektów,
- zarówno sam współczynnik s_{ij} jak i jego wartości cząstkowe s_{ijk} są unormowane,
- możliwość porównania k-tej składowej opisuje współczynnik δ_{ijk} przyjmujący wartość 1 (możliwe porównanie) lub 0 (przeciwna sytuacja).

Współczynniki Gowera - definicja

Współczynniki Gowera - definicja

$$s_{ij} = \frac{\sum_{k=1}^{p} s_{ijk}}{\sum_{k=1}^{p} \delta_{ijk}}$$

Współczynniki Gowera - definicja

$$s_{ij} = \frac{\sum_{k=1}^{p} s_{ijk}}{\sum_{k=1}^{p} \delta_{ijk}}$$

Dla zmiennych liczbowych

Współczynniki Gowera - definicja

$$s_{ij} = \frac{\sum_{k=1}^{p} s_{ijk}}{\sum_{k=1}^{p} \delta_{ijk}}$$

Dla zmiennych liczbowych

$$s_{ijk} = 1 - \frac{|x_i^{(k)} - x_j^{(k)}|}{\text{zakres } k\text{-tej zmiennej}}$$

Współczynniki Gowera - definicja

$$s_{ij} = \frac{\sum_{k=1}^{p} s_{ijk}}{\sum_{k=1}^{p} \delta_{ijk}}$$

Dla zmiennych liczbowych

$$s_{ijk} = 1 - \frac{|x_i^{(k)} - x_j^{(k)}|}{\text{zakres } k\text{-tej zmiennej}}$$

Dla zmiennych jakościowych

Współczynniki Gowera - definicja

$$s_{ij} = \frac{\sum_{k=1}^{p} s_{ijk}}{\sum_{k=1}^{p} \delta_{ijk}}$$

Dla zmiennych liczbowych

$$s_{ijk} = 1 - \frac{|x_i^{(k)} - x_j^{(k)}|}{\text{zakres } k\text{-tej zmiennej}}$$

Dla zmiennych jakościowych

$$s_{ijk} = \begin{cases} 1, \text{ gdy } x_i^{(k)} = x_j^{(k)} \\ 0, \text{ w przeciwnym przypadku} \end{cases}$$

Współczynniki Gowera - definicja

$$s_{ij} = \frac{\sum_{k=1}^{p} s_{ijk}}{\sum_{k=1}^{p} \delta_{ijk}}$$

Dla zmiennych jakościowych

$$s_{ijk} = \begin{cases} 1, \text{ gdy } x_i^{(k)} = x_j^{(k)} \\ 0, \text{ w przeciwnym przypadku} \end{cases}$$

Dla zmiennych liczbowych

$$s_{ijk} = 1 - \frac{|x_i^{(k)} - x_j^{(k)}|}{\text{zakres } k\text{-tej zmiennej}}$$

Dla zmiennych binarnych

Współczynniki Gowera - definicja

$$s_{ij} = \frac{\sum_{k=1}^{p} s_{ijk}}{\sum_{k=1}^{p} \delta_{ijk}}$$

Dla zmiennych jakościowych

$$s_{ijk} = \begin{cases} 1, \text{ gdy } x_i^{(k)} = x_j^{(k)} \\ 0, \text{ w przeciwnym przypadku} \end{cases}$$

Dla zmiennych liczbowych

$$s_{ijk} = 1 - \frac{|x_i^{(k)} - x_j^{(k)}|}{\text{zakres } k\text{-tej zmiennej}}$$

Dla zmiennych binarnych

	wartość zmiennej k			
obiekt i-ty	1	1	0	0
obiekt j-ty	1	1	0	0
s_{ijk}	1	0	0	0
δ_{ijk}	1	1	1	0

Jak sobie radzić z wartością średnią?

Średnia zbioru

Średnia zbioru Z_k wyznacza punkt w przestrzeni $\mathcal{R}^p$ minimalizujący sumę kwadratów odegłości od tego punktu do wszystkich punktów zbioru Z_k:

$$\bar{\mathbf{x}}_{Z_k} = \arg\min \sum_{\mathbf{x}_i \in Z_k} d(\mathbf{x}_i, \mathbf{y})$$

Czyli zadanie minimalizacji $\widetilde{W}$ jest równoważne następującemu zadaniu minimalizacji względem rodziny C wszystkich możliwych podziałów próby na K rozłącznych skupień i jednocześnie względem środków tych skupień.

$$\min_{C, \{\mathbf{y}_k\}_{k=1}^K} \sum_{i=1}^{K} d(\mathbf{x}_i, \mathbf{y}_{C(i)}) = \min_{C, \{\mathbf{y}_k\}_{k=1}^K} \sum_{i=1}^{K} \sum_{C(i)=k} d(\mathbf{x}_i, \mathbf{y}_k)$$

W efekcie środkiem takiego skupienia może być tylko **jeden z elementów**.

Metody hierarchiczne

- opierają sie na pomiarze uogólnionej odmienności między dwoma dowolnymi zbiorami obserwacji,

Metody hierarchiczne

- opierają sie na pomiarze uogólnionej odmienności między dwoma dowolnymi zbiorami obserwacji,
- nie wymagają z góry określenia liczby skupień,

Metody hierarchiczne

- opierają sie na pomiarze uogólnionej odmienności między dwoma dowolnymi zbiorami obserwacji,
- nie wymagają z góry określenia liczby skupień,
- w pierwszym kroku **metody algomeracyjnej** tworzymy tyle skupień, ile jest obserwacji,

Metody hierarchiczne

- opierają sie na pomiarze uogólnionej odmienności między dwoma dowolnymi zbiorami obserwacji,
- nie wymagają z góry określenia liczby skupień,
- w pierwszym kroku **metody algomeracyjnej** tworzymy tyle skupień, ile jest obserwacji,
- w nastepnym kroku w jedno skupienie łączona jest para najmniej odległych obserwacji,

Metody hierarchiczne

- opierają sie na pomiarze uogólnionej odmienności między dwoma dowolnymi zbiorami obserwacji,
- nie wymagają z góry określenia liczby skupień,
- w pierwszym kroku **metody algomeracyjnej** tworzymy tyle skupień, ile jest obserwacji,
- w nastepnym kroku w jedno skupienie łączona jest para najmniej odległych obserwacji,
- z koroku na krok skupień jest coraz mniej, aż w ostatnim powstaje cała próba w jednym skupieniu,

Metody hierarchiczne

- opierają sie na pomiarze uogólnionej odmienności między dwoma dowolnymi zbiorami obserwacji,
- nie wymagają z góry określenia liczby skupień,
- w pierwszym kroku **metody algomeracyjnej** tworzymy tyle skupień, ile jest obserwacji,
- w nastepnym kroku w jedno skupienie łączona jest para najmniej odległych obserwacji,
- z koroku na krok skupień jest coraz mniej, aż w ostatnim powstaje cała próba w jednym skupieniu,
- w efekcie otrzymuje się nieskierowane drzewo (dendrogram),

Metody hierarchiczne

- opierają sie na pomiarze uogólnionej odmienności między dwoma dowolnymi zbiorami obserwacji,
- nie wymagają z góry określenia liczby skupień,
- w pierwszym kroku **metody algomeracyjnej** tworzymy tyle skupień, ile jest obserwacji,
- w nastepnym kroku w jedno skupienie łączona jest para najmniej odległych obserwacji,
- z koroku na krok skupień jest coraz mniej, aż w ostatnim powstaje cała próba w jednym skupieniu,
- w efekcie otrzymuje się nieskierowane drzewo (dendrogram),
- to samo można otrzymać "idąc od góry"(**metoda oparta na dzieleniu**), ale jest to dużo bardziej złożony obliczeniowo.

Rodzaje odmienności

1. odmienność **najbliższego sąsiada** (single linkage) miedzy skupieniami *i* oraz *j* jest równa **najmniejszej** spośród $n_i n_j$ odmienności między parami obserwacji, z których jedna jest z jednego a druga z drugiego skupienia,

Rodzaje odmienności

1. odmienność **najbliższego sąsiada** (single linkage) miedzy skupieniami *i* oraz *j* jest równa **najmniejszej** spośród $n_i n_j$ odmienności między parami obserwacji, z których jedna jest z jednego a druga z drugiego skupienia,
2. odmienność **najdalszego sąsiada** (complete linkage) miedzy skupieniami *i* oraz *j* jest równa **największej** spośród $n_i n_j$ odmienności między parami obserwacji, z których jedna jest z jednego a druga z drugiego skupienia,

Rodzaje odmienności

1. odmienność **najbliższego sąsiada** (single linkage) miedzy skupieniami *i* oraz *j* jest równa **najmniejszej** spośród $n_i n_j$ odmienności między parami obserwacji, z których jedna jest z jednego a druga z drugiego skupienia,
2. odmienność **najdalszego sąsiada** (complete linkage) miedzy skupieniami *i* oraz *j* jest równa **największej** spośród $n_i n_j$ odmienności między parami obserwacji, z których jedna jest z jednego a druga z drugiego skupienia,
3. odmienność **średnia** (average linkage) — uśredniona wartość odmienności między parami obserwacji.

Związki rekurencyjne

1. $D_{k,ij} = \min\{D_{ki}, D_{kj}\} = \frac{1}{2}\left(D_{ik} + D_{kj} - |D_{ik} - D_{kj}|\right)$

Związki rekurencyjne

1. $D_{k,ij} = \min\{D_{ki}, D_{kj}\} = \frac{1}{2}\left(D_{ik} + D_{kj} - |D_{ik} - D_{kj}|\right)$
2. $D_{k,ij} = \max\{D_{ki}, D_{kj}\} = \frac{1}{2}\left(D_{ik} + D_{kj} + |D_{ik} - D_{kj}|\right)$

Związki rekurencyjne

1. $D_{k,ij} = \min\{D_{ki}, D_{kj}\} = \frac{1}{2}\left(D_{ik} + D_{kj} - |D_{ik} - D_{kj}|\right)$
2. $D_{k,ij} = \max\{D_{ki}, D_{kj}\} = \frac{1}{2}\left(D_{ik} + D_{kj} + |D_{ik} - D_{kj}|\right)$
3. $D_{k,ij} = \frac{n_i}{n_i+n_j} D_{ik} + \frac{n_j}{n_i+n_j}$.

Związki rekurencyjne

1. $D_{k,ij} = \min\{D_{ki}, D_{kj}\} = \frac{1}{2}\left(D_{ik} + D_{kj} - |D_{ik} - D_{kj}|\right)$
2. $D_{k,ij} = \max\{D_{ki}, D_{kj}\} = \frac{1}{2}\left(D_{ik} + D_{kj} + |D_{ik} - D_{kj}|\right)$
3. $D_{k,ij} = \frac{n_i}{n_i+n_j} D_{ik} + \frac{n_j}{n_i+n_j}$.