GDAŃSK UNIVERSITY OF TECHNOLOGY

Eksploracja danych

KLASTERYZACJA I SEGMENTACJA cz. 2

Wojciech Waloszek

wowal@eti.pg.gda.pl

Teresa Zawadzka

tegra@eti.pg.gda.pl

Katedra Inżynierii Oprogramowania
Wydział Elektroniki, Telekomunikacji i Informatyki
Politechnika Gdańska

GDAŃSK UNIVERSITY OF TECHNOLOGY

Inne ważne metody grupowania

- K-means:
  - Iteracyjne grupowanie na zasadzie odnajdywania centrów podzbiorów przykładów,
- EM:
  - Próba odnalezienia optymalnego globalnie podziału poprzez poszukiwanie parametrów rozkładów Gaussa.

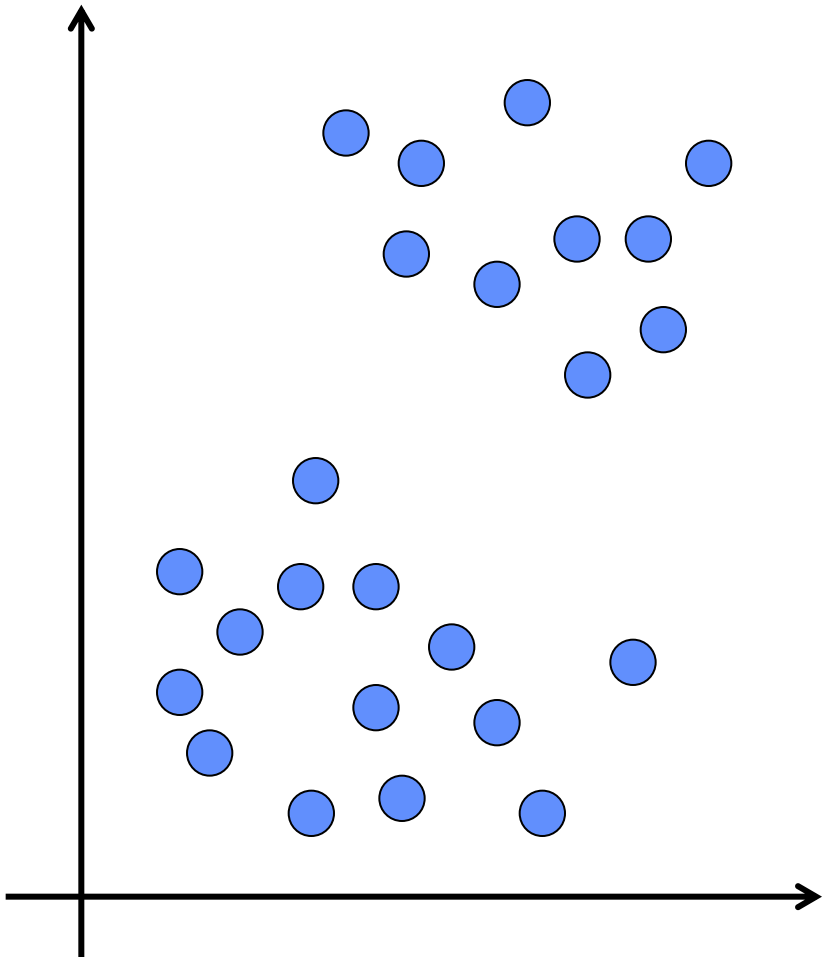
GDAŃSK UNIVERSITY OF TECHNOLOGY

K-means

- Poszukuje „centrów” dla zadanej liczby klastrów,
- Pierwszy dobór centrów odbywa się losowo, kolejne iteracje algorytmu przesuwają „centra”, aby lepiej odzwierciedlały one podział przestrzeni przykładów,
- Kryterium stopu jest zazwyczaj pewien próg odległości, o którą zostały przesunięte „centra” w kolejnej iteracji.

GDAŃSK UNIVERSITY OF TECHNOLOGY

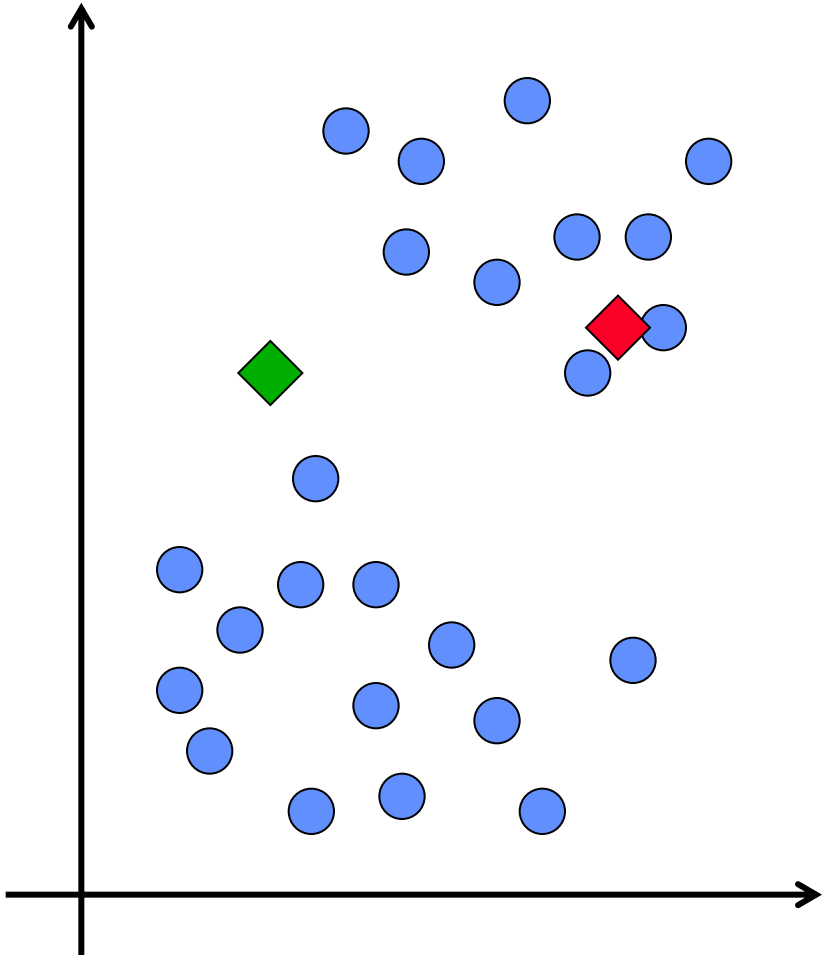
K-means – przykład działania



- Przestrzeń przykładów przedstawiamy sobie jako *n*-wymiarową przestrzeń euklidesową, w której przykłady reprezentowane są jako punkty

GDAŃSK UNIVERSITY OF TECHNOLOGY

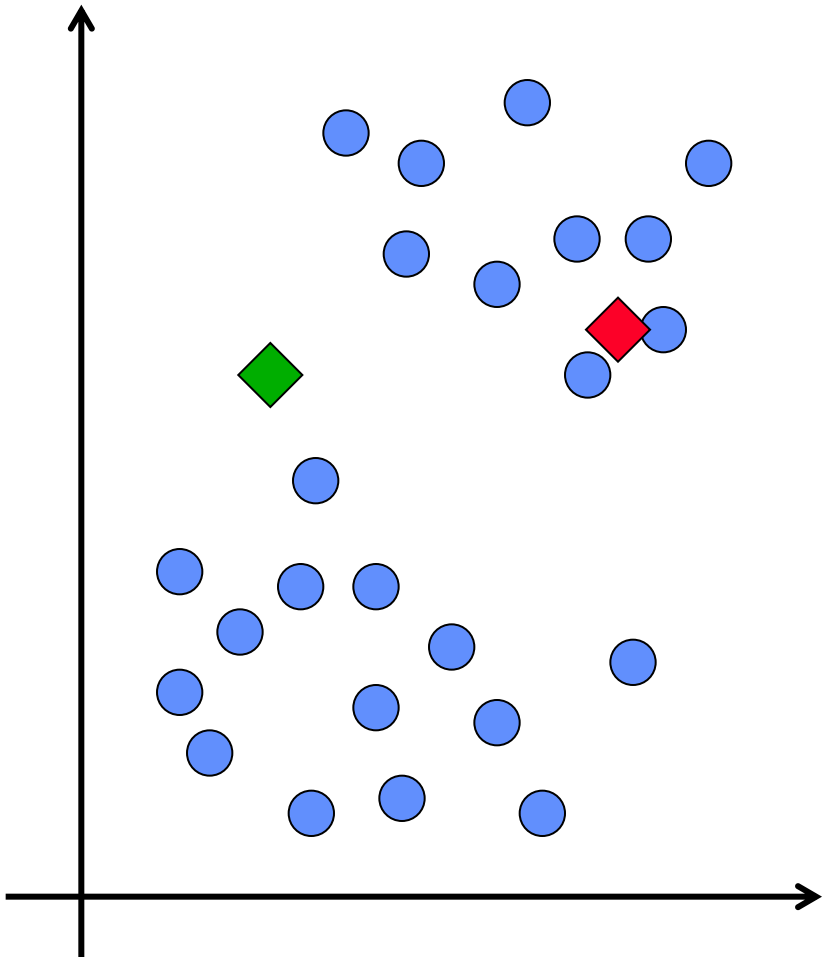
K-means – przykład działania (2)



- Na początek losujemy położenie *k* centrów (tutaj przyjmiemy *k* = 2)

GDAŃSK UNIVERSITY OF TECHNOLOGY

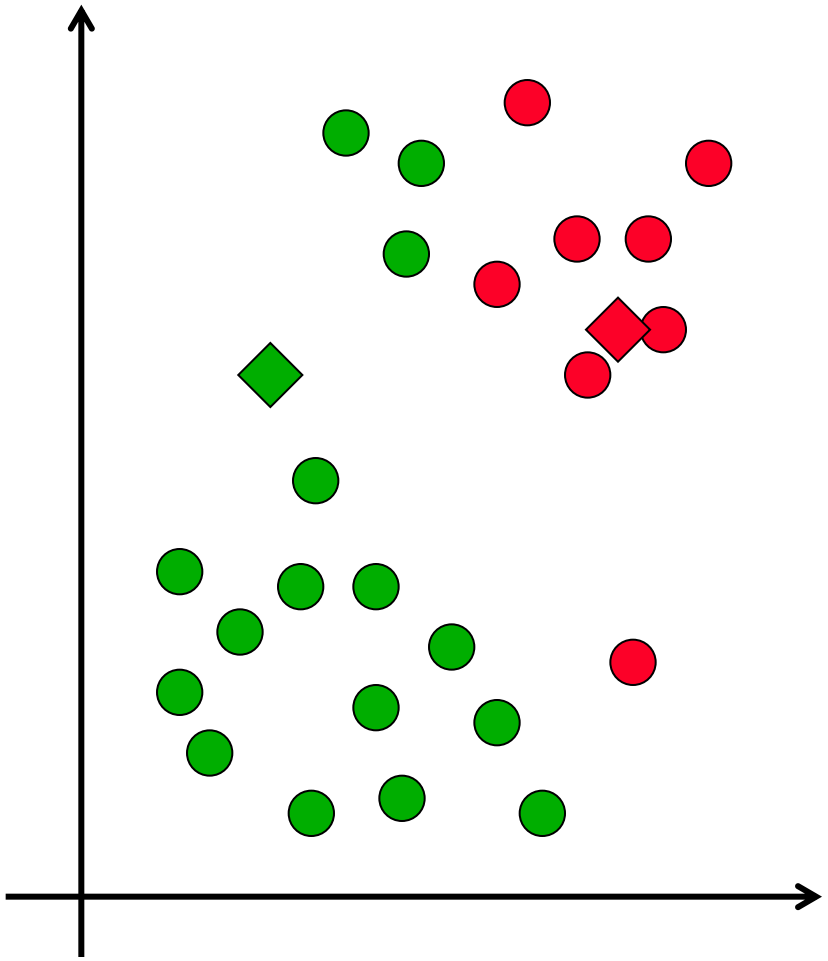
K-means – przykład działania (3)



- Oznaczamy, które przykłady leżą najbliżej którego „centrum”

GDAŃSK UNIVERSITY OF TECHNOLOGY

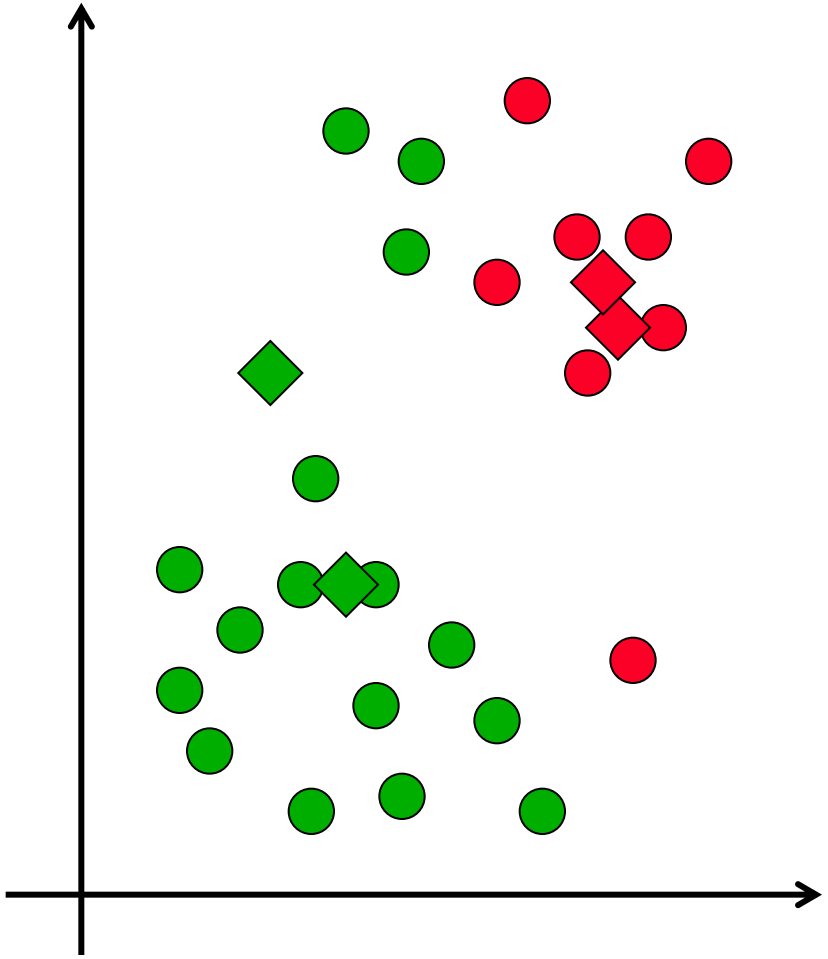
K-means – przykład działania (3)



- Oznaczamy, które przykłady leżą najbliżej którego „centrum”

GDAŃSK UNIVERSITY OF TECHNOLOGY

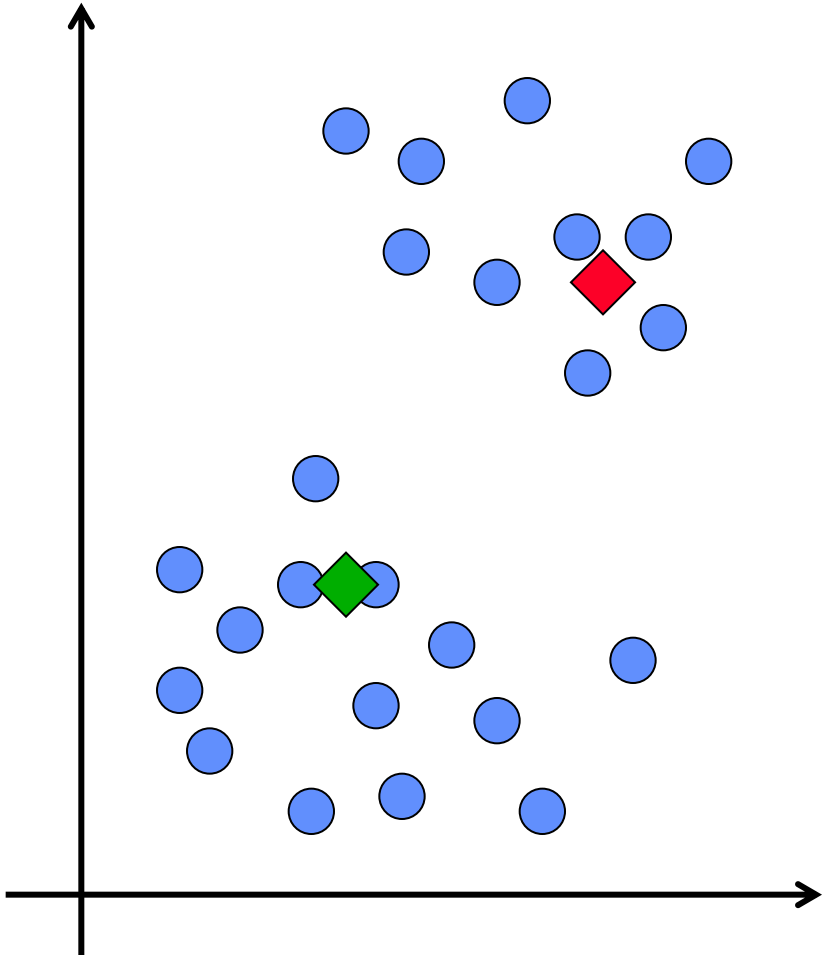
K-means – przykład działania (4)



- Przenosimy dotychczasowe „centra” do „środków” przypisanych im zbiorów przykładów

GDAŃSK UNIVERSITY OF TECHNOLOGY

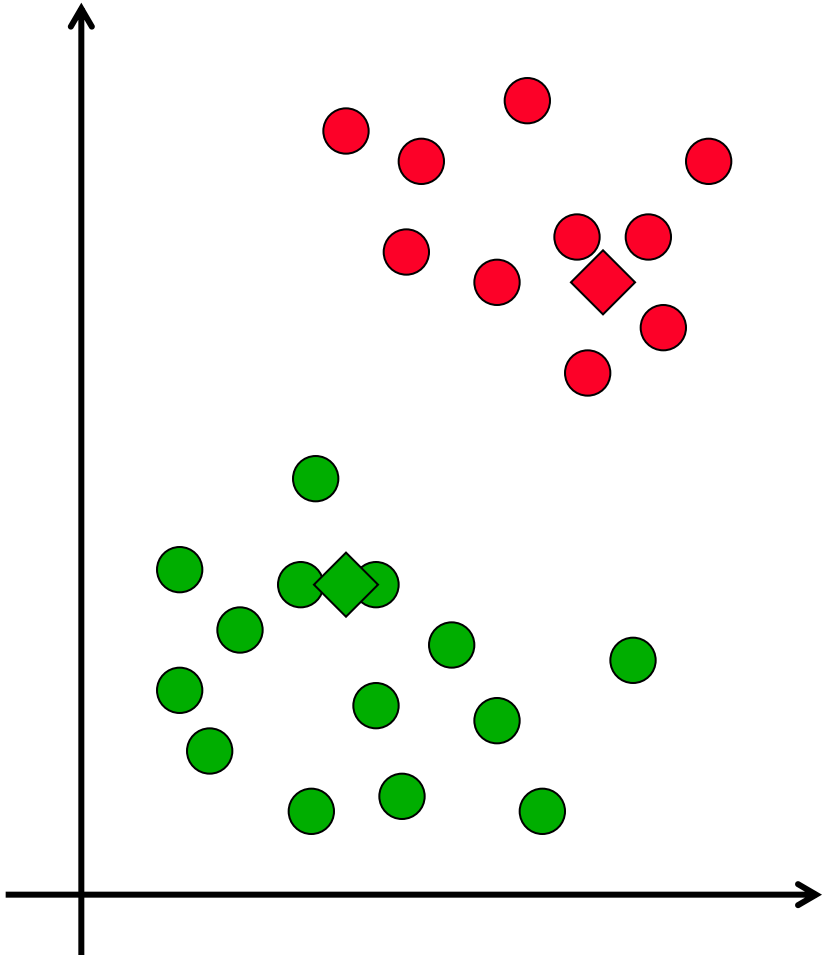
K-means – przykład działania (5)



- Ponownie przypisujemy przykłady „centrom”

GDAŃSK UNIVERSITY OF TECHNOLOGY

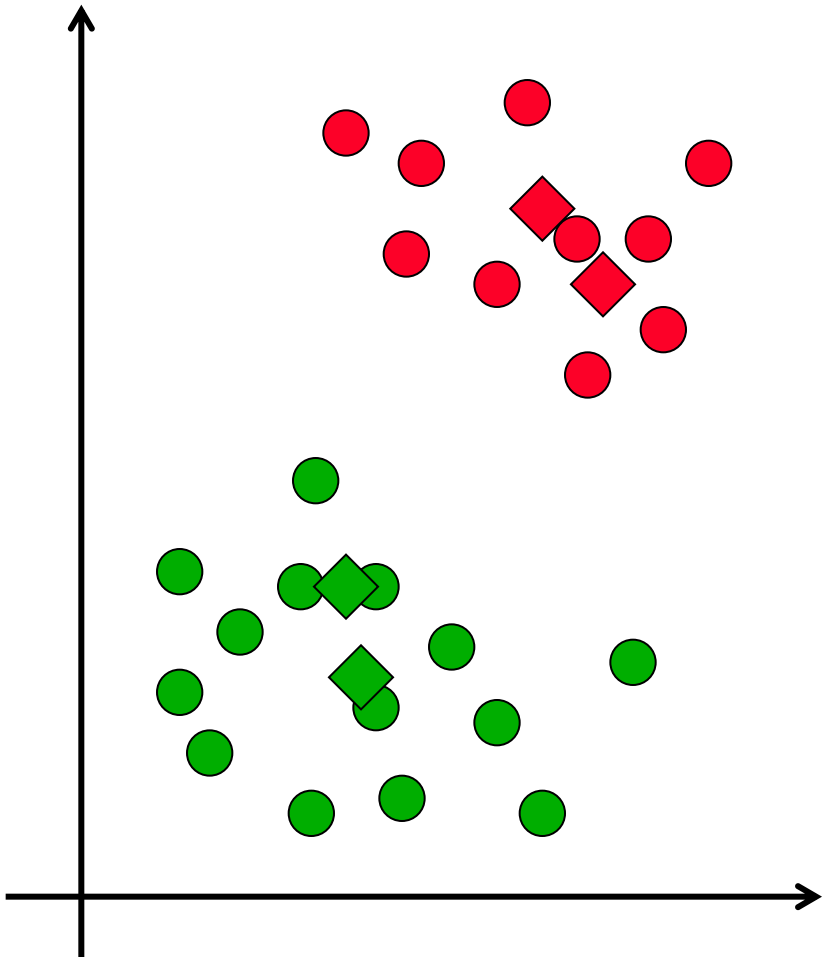
K-means – przykład działania (5)



- Ponownie przypisujemy przykłady „centrom”

GDAŃSK UNIVERSITY OF TECHNOLOGY

K-means – przykład działania (6)



- I ponownie przesuwamy „centra”…

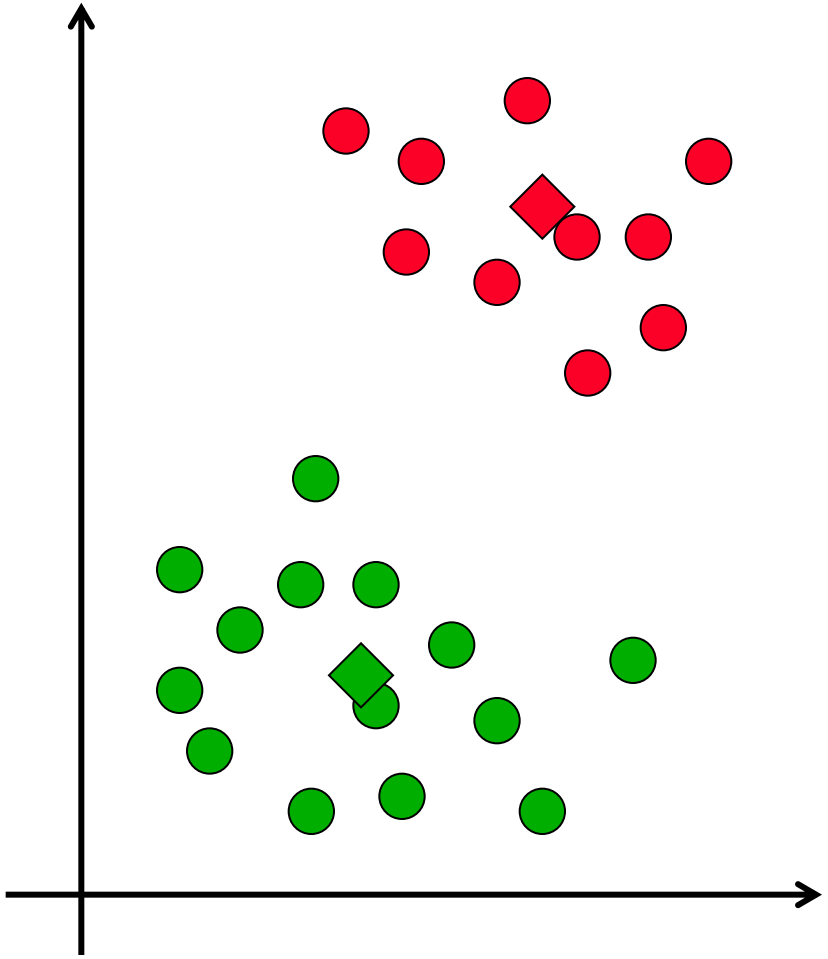
GDAŃSK UNIVERSITY OF TECHNOLOGY

K-means

- K-means może wymagać wykonanie kilku powtórzeń całego algorytmu, jako że sam algorytm ma tendencję do znajdywania „minimum lokalnego”,
- K-means wymaga podanej z góry liczby tworzonych klastrów, zaś podział generowany przez algorytm jest płaski, co w niektórych zastosowaniach może być wadą,
- Z tego względu stosuje się hierarchiczną wersję algorytmu K-means.

GDAŃSK UNIVERSITY OF TECHNOLOGY

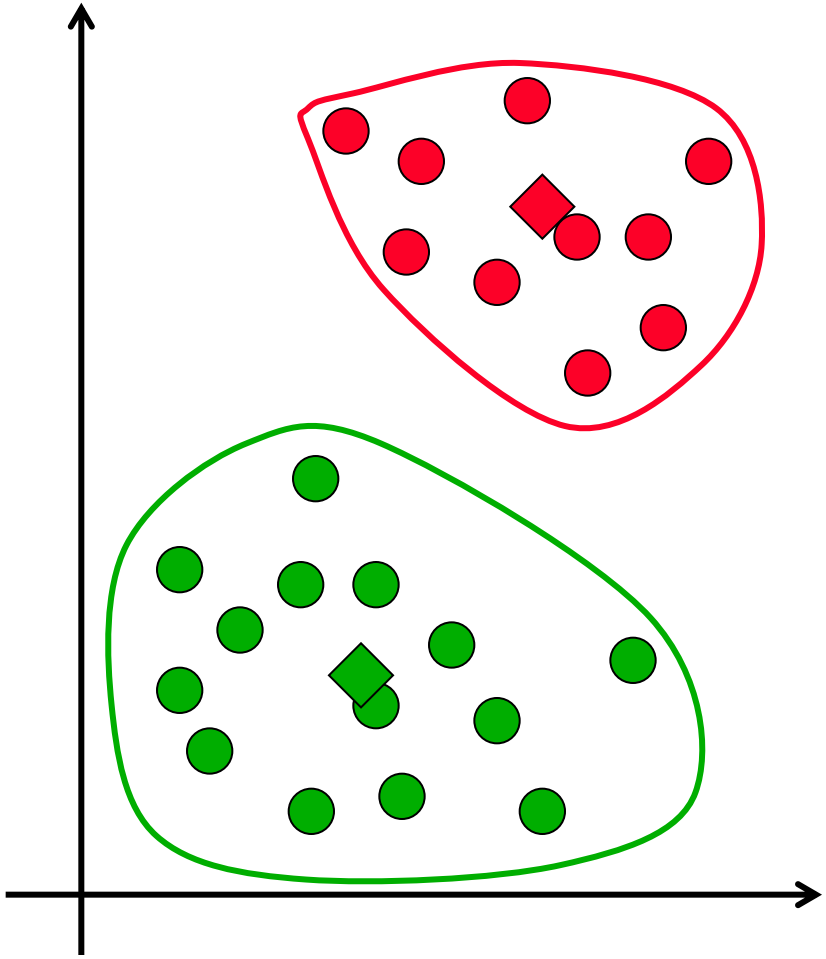
Hierarchiczny k-means



- Algorytm rozpoczyna pracę, jak zwykły k-means dla $k = 2$

GDAŃSK UNIVERSITY OF TECHNOLOGY

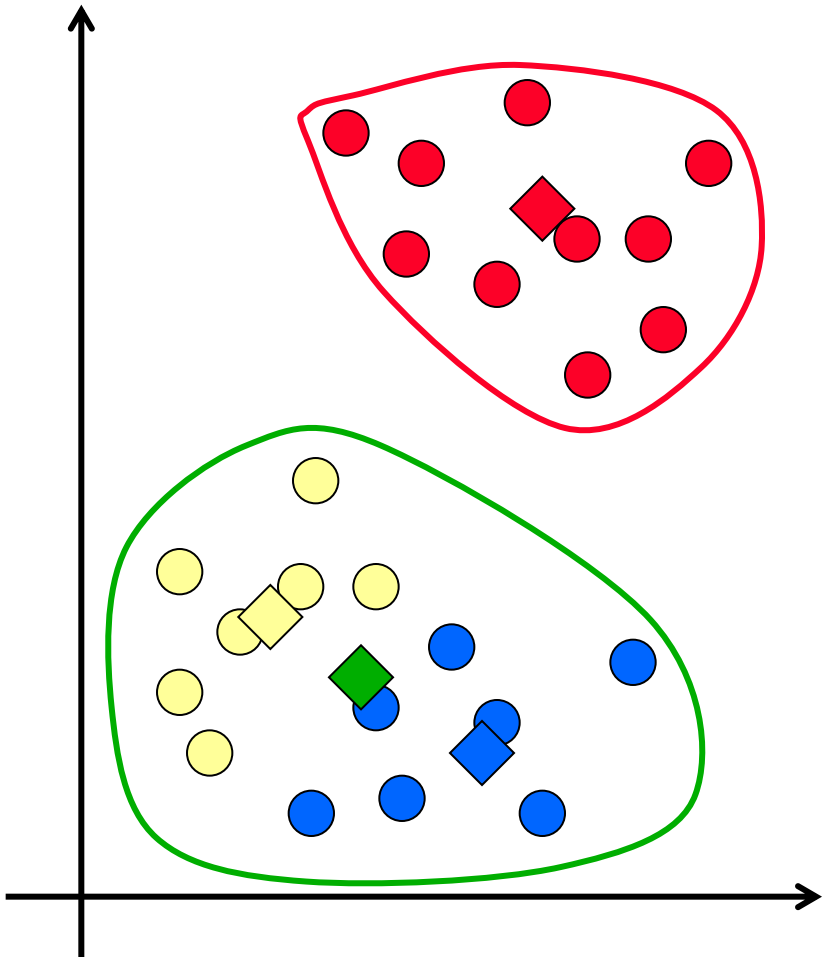
Hierarchiczny k-means (2)



- Dla „zbyt dużych” klastrów przeprowadzany jest proces ponownego podziały na zasadzie analogicznej, jak dla algorytmu k-means

GDAŃSK UNIVERSITY OF TECHNOLOGY

Hierarchiczny k-means (3)



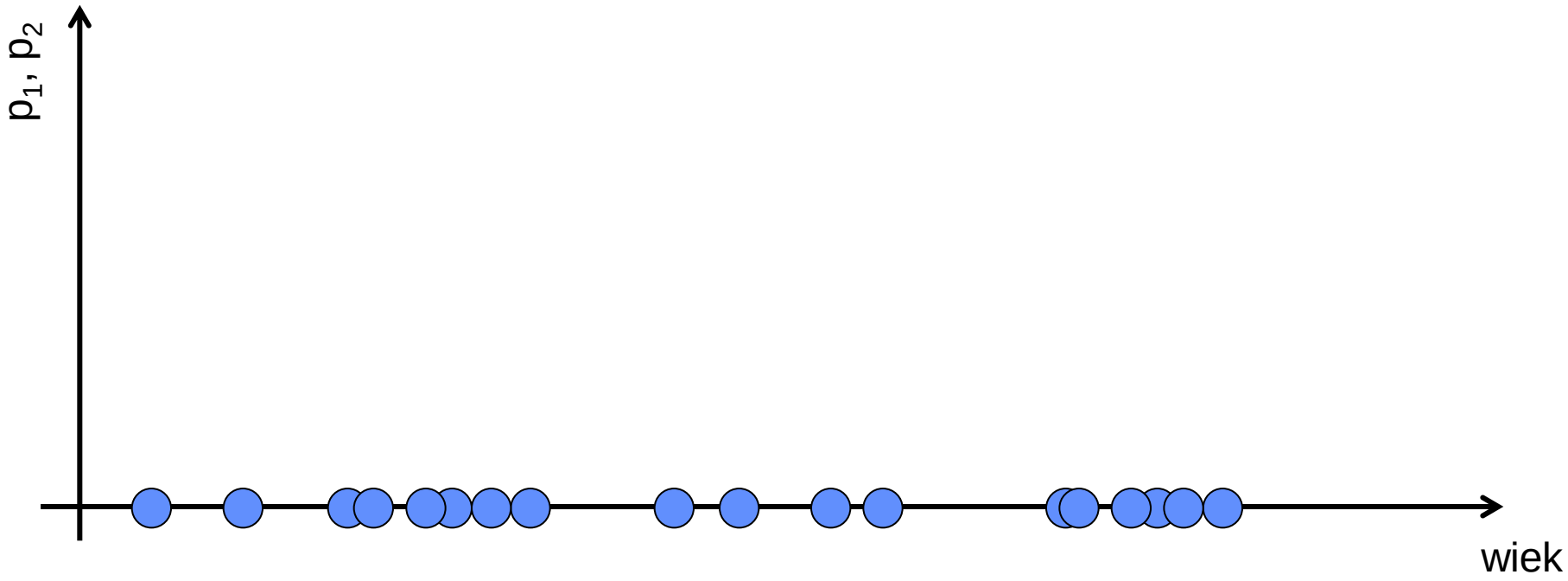
- Wynikiem działania algorytmu jest hierarchiczny podział przestrzeni przykładów na grupy

GDAŃSK UNIVERSITY OF TECHNOLOGY

EM (expectation-maximization)

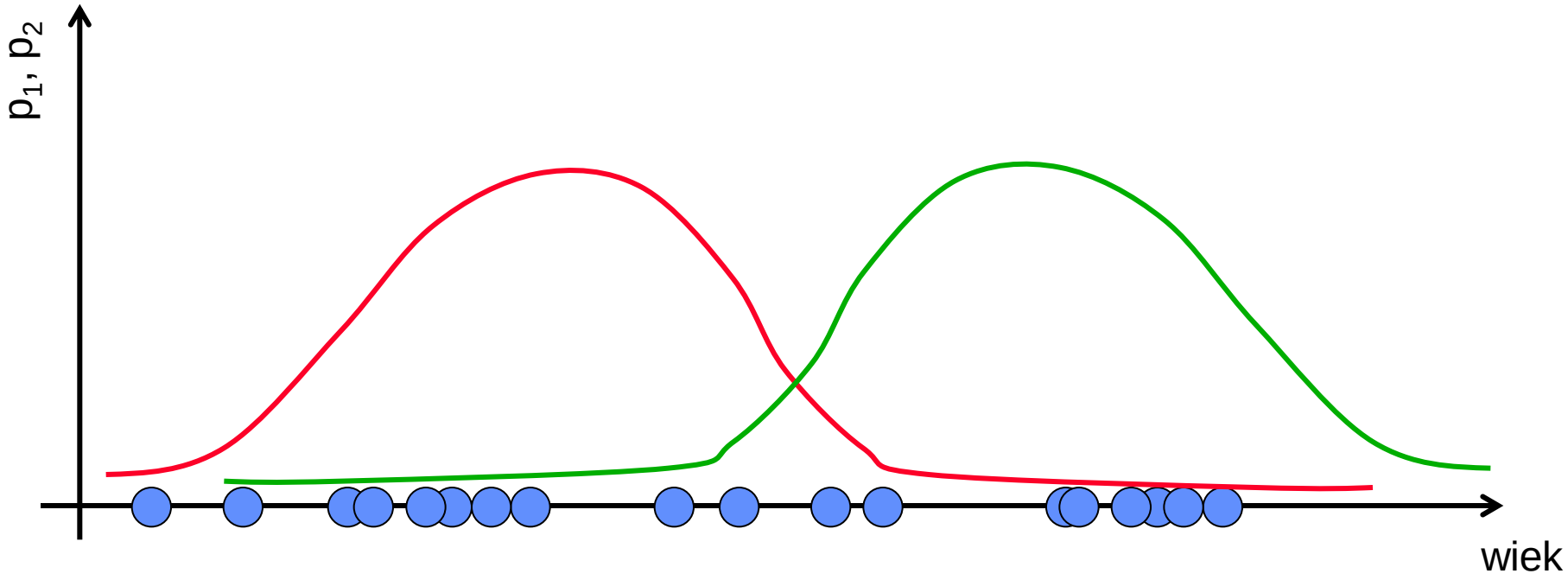
- Algorytm EM zakłada, że zbiór uczący to „mieszanina” przykładów z *k* kategorii,
- EM zakłada, że każda kategoria przykładów charakteryzuje się pewnym rozkładem wartości atrybutów (najczęściej przyjmowany jest rozkład normalny),
- EM, podobnie jak k-means, działa iteracyjnie, dostosowując parametry rozkładów do przykładów obejmowanych przez kategorie na kolejnym etapie przetwarzania.

EM – przykład działania



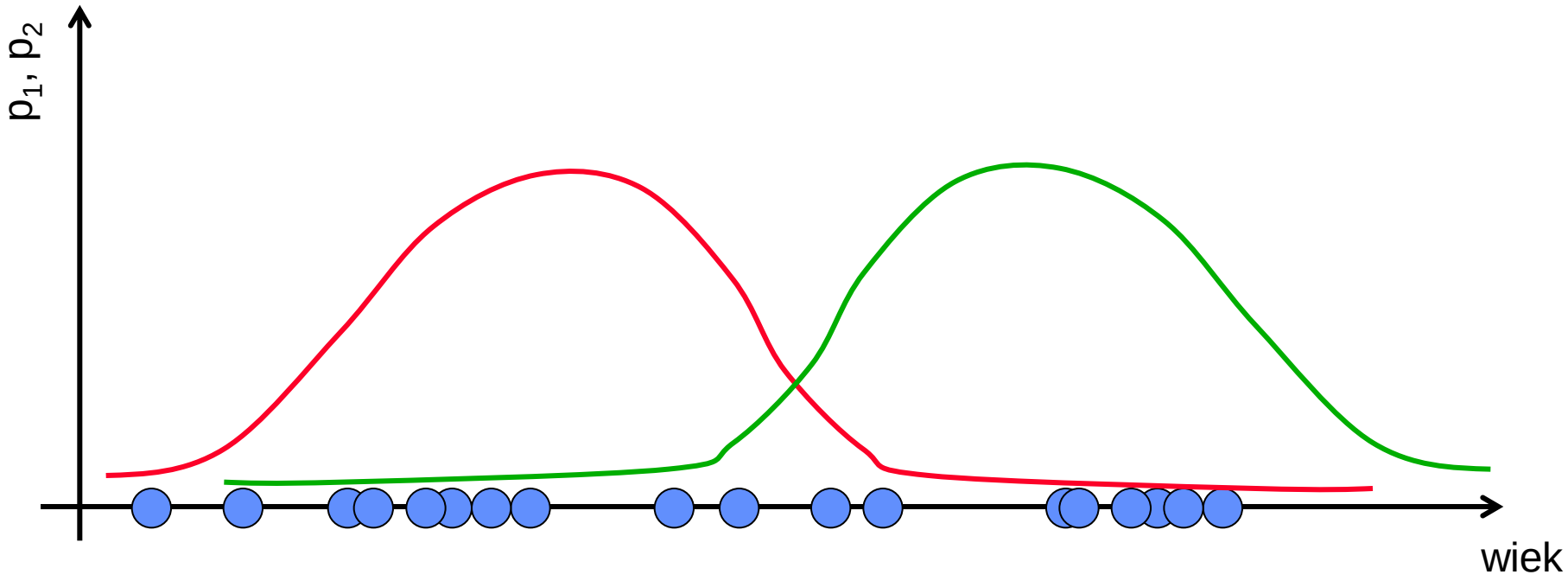
- Przestrzeń przykładów możemy rozumieć podobnie, jak w przypadku k-means, tutaj wyobraźmy sobie 1 wymiar (wartość atrybutu wiek)

EM – przykład działania (2)



- Na początek przyjmujemy dwa „losowe” rozkłady…

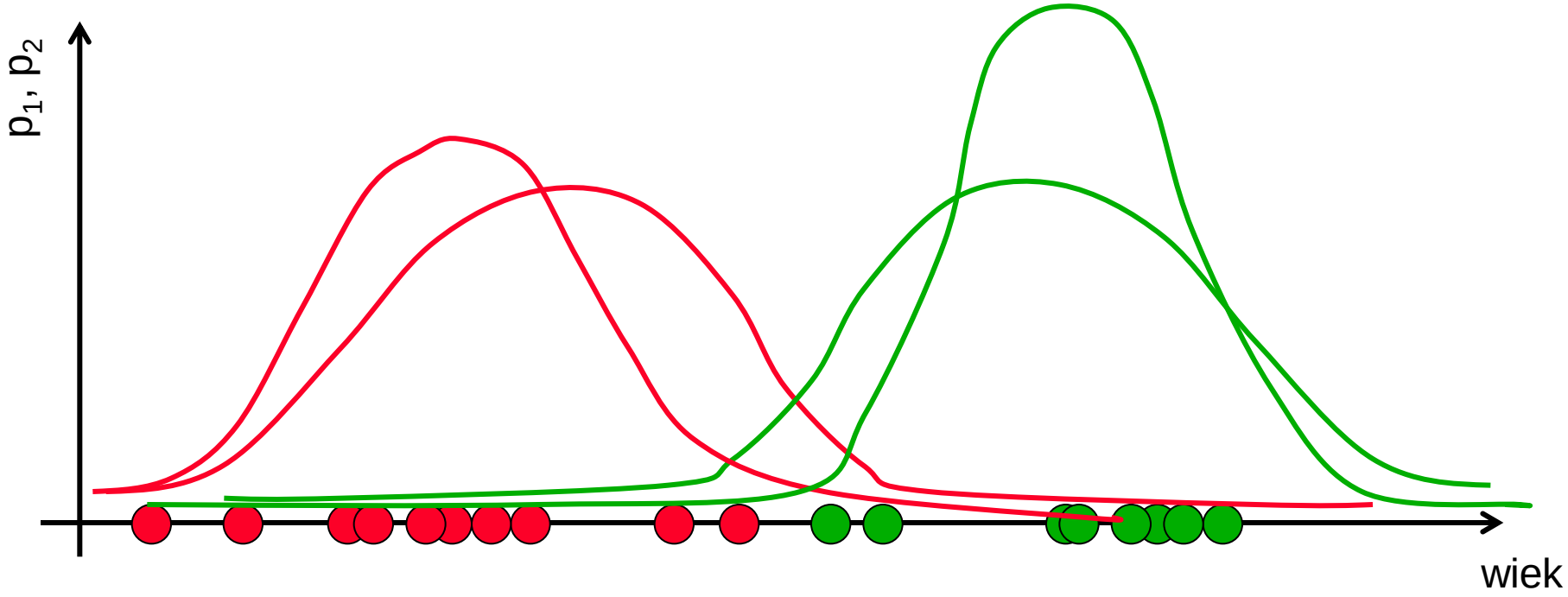
EM – przykład działania (3)



- Następnie przypisujemy przykłady do odp. kategorii, po czym modyfikujemy parametry rozkładów

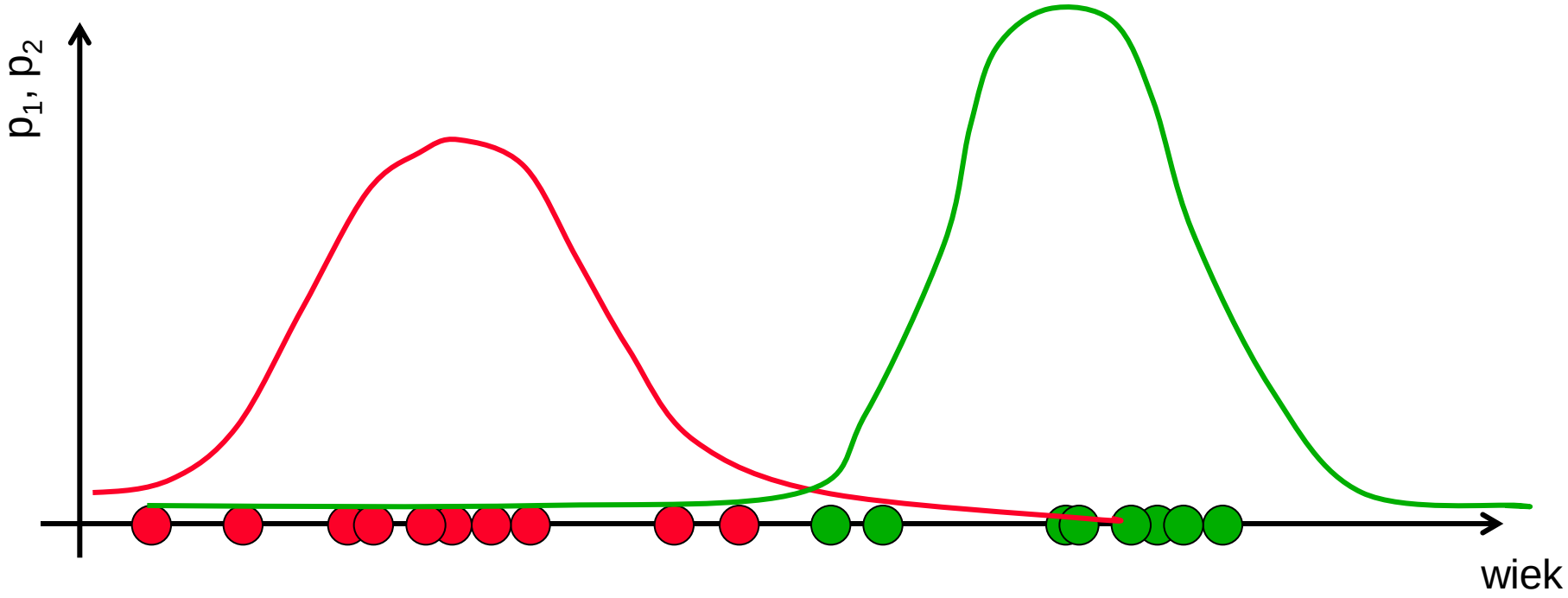
GDAŃSK UNIVERSITY OF TECHNOLOGY

EM – przykład działania (3)



- Następnie przypisujemy przykłady do odp. kategorii, po czym modyfikujemy parametry rozkładów

EM – przykład działania (4)



- Proces powtarzamy iteracyjnie, aż do osiągnięcia kryterium stopu (niewielka zmiana parametrów)

GDAŃSK UNIVERSITY OF TECHNOLOGY

EM – podsumowanie

- EM można rozszerzyć na więcej wymiarów, co jest proste przy założeniu niezależności wartości odpowiednich atrybutów,
- W przypadku odnalezienia zależności między wartościami atrybutów można zastosować rozkład dwu- (lub więcej) wymiarowy i szukać jego parametrów
- Podobnie jak k-means, EM szuka „minimum lokalnego”, stąd konieczne może być wykonanie algorytmu kilkukrotnie

GDAŃSK UNIVERSITY OF TECHNOLOGY

Dziękujemy za uwagę

Zapraszamy na wykład:

REGUŁY ASOCJACYJNE