Grupowanie

Główną ideą grupowania jest podział zbioru próbek na kilka grup (klastrów) i ew. wyznaczenie środka tych grup. Wszystkie próbki w danym klastrze powinny znajdować się w miarę blisko siebie (blisko tzw środka grupy) oraz jak najdalej od innych grup (i ich środków).

Przygotowanie zbioru próbek

W ćwiczeniu będziemy operować na zbiorze próbek zapisanych w macierzy "probki". Każda próbka posiada 2 atrybuty i zawarta jest w pojedynczym wierszu macierzy "probki". Można je wygenerować za pomocą poniższego skryptu napisanego z języku Octave, ale najlepiej użyć gotowej tablicy z próbkami podanej niżej.

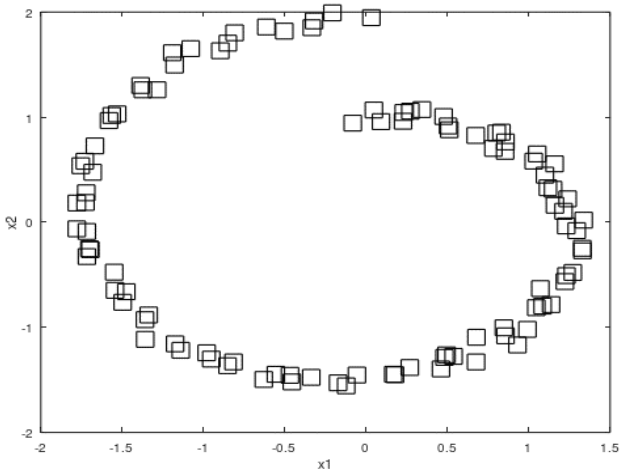
```
rand('seed', 123);
n = 2; %liczba atrybutow
M = 101; % liczba probek
m = 4; % liczba grup
%tworzenie zbioru probek
t= 1:0.01:2;
probki = [  t.*sin(t*2*pi)-0.1+0.2*rand(1,length(t))  ; ...
   t.*cos(t*2*pi)-0.1+0.2*rand(1,length(t))   ]' ;
 clear t;
plot(probki(:,1), probki(:,2), 'sk'); xlabel('atrybut 1'); ylabel('atrybut 2');
```

Baza próbek spiralka

Każda kolumna określa parametr1 (x1) i parametr2 (x2), pojedynczy wiersz to pojedyncza próbka i jednocześnie pojedynczy punkt:

```
  -0.0824    0.9435
   0.0913    0.9575
   0.0476    1.0683
   0.2311    1.0452
   0.2269    0.9615
   0.2717    1.0570
   0.3430    1.0724
   0.4765    1.0060
   0.5039    0.9162
   0.5113    0.8803
   0.6735    0.8263
   0.8030    0.8494
   0.8318    0.8573
   0.7825    0.7032
   0.8566    0.7646
   0.8544    0.6752
   1.0519    0.6502
   1.0302    0.5818
   1.1587    0.5559
   1.0992    0.4486
   1.1175    0.3286
   1.1493    0.3144
   1.1607    0.1619
   1.2388    0.2232
   1.2136    0.1041
   1.3375    0.0204
   1.2301   -0.0344
   1.2946   -0.0799
   1.3301   -0.2716
   1.3287   -0.2480
   1.2707   -0.4779
   1.2219   -0.5640
   1.2314   -0.5076
   1.0710   -0.6311
   1.0461   -0.8127
   1.0850   -0.7972
   1.1371   -0.7834
   0.9918   -1.0194
   0.8467   -1.0041
   0.9301   -1.1675
   0.8569   -1.0805
   0.6782   -1.0943
   0.6759   -1.3292
   0.5377   -1.2760
   0.4944   -1.2617
   0.4819   -1.2857
```

```
 0.4601   -1.3957
 0.2684   -1.3820
 0.1669   -1.4465
 0.1774   -1.4508
-0.0566   -1.4538
-0.1226   -1.5564
-0.1725   -1.5277
-0.3374   -1.4758
-0.4568   -1.5191
-0.4682   -1.4563
-0.5552   -1.4464
-0.6295   -1.4956
-0.8523   -1.3649
-0.8152   -1.3287
-0.9520   -1.3021
-0.9805   -1.2422
-1.1398   -1.2181
-1.1746   -1.1570
-1.3583   -1.1144
-1.3366   -0.8821
-1.3595   -0.9266
-1.4969   -0.7615
-1.5441   -0.6479
-1.4743   -0.6608
-1.5488   -0.4747
-1.7169   -0.3264
-1.7011   -0.2524
-1.6954   -0.2604
-1.7168   -0.0879
-1.7790   -0.0625
-1.7251    0.1885
-1.7787    0.1831
-1.7208    0.2806
-1.7544    0.5381
-1.6810    0.4760
-1.7296    0.5807
-1.6673    0.7270
-1.5808    0.9680
-1.5305    1.0313
-1.5638    1.0163
-1.3728    1.2606
-1.2852    1.2611
-1.3866    1.3016
-1.1773    1.4962
-1.1918    1.6133
-1.0784    1.6521
-0.8967    1.6324
-0.8501    1.7047
-0.8084    1.8031
-0.6135    1.8584
-0.5031    1.8191
-0.3340    1.8515
-0.3200    1.9158
-0.2074    1.9929
 0.0322    1.9473
```



Miara odległości

Istnieje wiele różnych miar liczących odległość między próbkami. Oto przykład najpopularniejszych z nich:

- euklidesowa - $odleglosc(pktA, pktB) = \sqrt{\sum_{i=1..n}(pktA_i - pktB_i)^2}$, gdzie i to liczba wymiarów, $pktA_i$ to wartość i-tego wymiaru punktu pktA;
- manhattan (taksówkowa) - $odleglosc(pktA, pktB) = \sum_{i=1..n}|pktA_i - pktB_i|$;
- Czebyszewa - $odleglosc(pktA, pktB) = max_{i=1..n}|pktA_i - pktB_i|$;
- Minkowskiego - $odleglosc(pktA, pktB) = \left(\sum_{i=1..n}|pktA_i - pktB_i|^{mink}\right)^{1/mink}$, gdzie mink to parametr sterujący, mink > 0.

Wszystkie te miary zakładają, że oba punkty są liczbami jedno/wielowymiarowymi.

Algorytm K-średnich

Algorytm tworzy *k* grup, każda z tych grup posiada środek (średnią z grupy). Algorytm działa w ten sposób, iż przez zadaną liczbę iteracji na przemian określa dla każdej próbki przynależność do każdej z grup (czyli do którego z *k* środków ma najbliżej) a następnie dla każdej z poprzednio ustalonej grupy wyznacza osobno środek. Algorytm do poprawnego zadania musi zawierać informację n/t środków *k* grup oraz musi posiadać informację n/t tego, które próbki mają najbliżej do danej grupy. Start algorytmu, liczba grup, iteracji i miara odległości nie jest z góry ustalona. W tym przykładzie zaleca się użyć 4 grup (*m*=4), 100 iteracji miary odległości euklidesowej a na początku ustalenie środka grup równe losowo wybranej grupie różnych próbek.

Algorytm zapisany w postaci pseudokodu:

1. Wybierz losowo *m* różnych próbek i uznaj je jako środki grup (V)

2. Pętla wykonywana zadaną liczbę iteracji (*iters*)

2.1. Pętla po wszystkich M próbkach, *s* to indeks aktualnej próbki

2.1.1. Wylicz odległości między próbką *s* a każdym środkiem grupy (V)

2.1.2. Wyznacz u_s równy indeksowi najbliższego środka grupy

2.2. Pętla po wszystkich m grupach, *j* to indeks aktualnej grupy

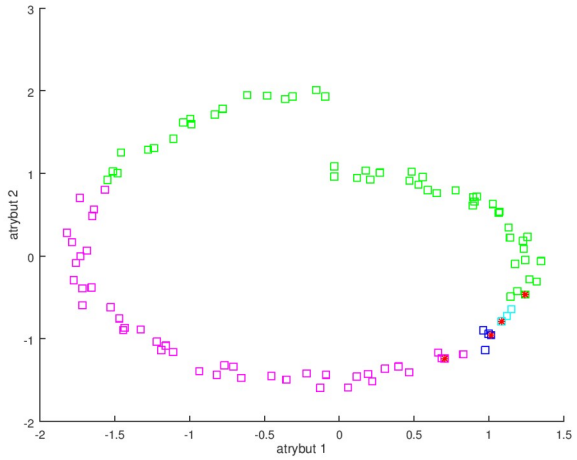
2.2.1. Wybierz próbki, należące do tej grupy (zbiór próbek o indeksach *s,* takich, że u_s == *j),* niech zbiór ten nazywa się X_{gr}

2.2.2. Jeśli zbiór X_{gr} jest pusty, wtedy pomiń wykonanie dalszej części tej pętli.

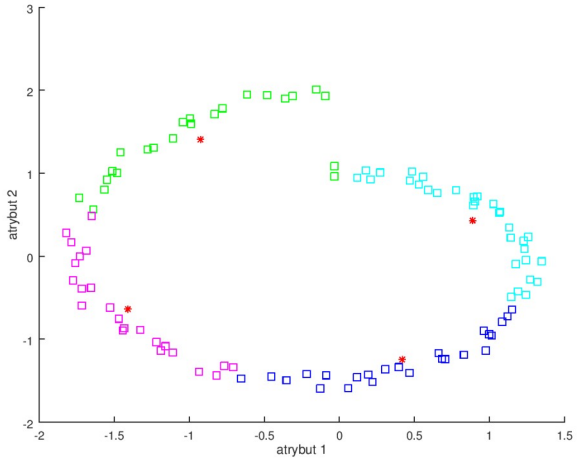
2.2.3. Pętla po wszystkich atrybutach, *i* to index poszczególnego atrybutu

2.2.3.1 Wartość *i*-tego atrybutu grupy *j*-tej to średnia wartość atrybutu *i*-tego wszystkich próbek X_{gr}

Przykład podziału na 4 grupy przed działaniem algorytmu k-średnich, odległość liczona metodą euklidesową:



Przykład podziału na 4 grupy po 11 iteracjach algorytmu k-średnich, odległość liczona metryką euklidesową:



Algorytm Fuzzy c-Means (FCM) **(nieobowiązkowy, na bdb)**

Algorytm jest rozszerzeniem algorytmu k-średnich. Każda próbka nie należy do dokładnie jednej z grup lecz należy do każdej z grup w różnym stopniu (stąd słowo fuzzy - rozmyty w nazwie algorytmu). Stopień rozmywania ustalony jest przez pojedynczy parametr nazwany tutaj fcm_m (fcm_m>1, domyślnie równy 2). Najpierw następuje start algorytmu, następnie wykonywany jest on w wielu iteracjach. W każdej z nich najpierw obliczana jest odległość między każdą próbką i grupą:

$$D_{j,s} = d(x_s, v_j)^2,$$

gdzie j to numer grupy (j=1..m), m to liczba grup, s to numer próbki (s=1..M), M to liczba próbek, d(…) to odległość. Następnie liczona jest przynależność (U) poszczególnej próbki do każdej z grup:

$$U_{j,s} = D_{j,s}^{1/(1-fcm\_m)} / \left(\sum_{j'=1}^{m} D_{j',s}^{1/(1-fcm\_m)} \right)$$

gdzie *fcm_m* to parametr algorytmu FCM określający sposób rozmywania, *fcm_m>1* i domyślnie jego wartość to 2, M to liczba próbek, s to numer próbki.

Warto wspomnieć, iż

$$\forall_s : \sum_{j=1}^{m}(U_{j,s}) = 1$$

czyli suma przynależności pojedynczej próbki do wszystkich grup wynosi 1. Kolejnym krokiem jest obliczenie nowej wartości środków grup używając wzoru

$$V_{j,i}^{(t+1)} = \left(\sum_{s=1}^{M} U_{j,s}^{fcm\_m} x_{s,i}\right) / \left(\sum_{s=1}^{M} U_{j,s}^{fcm\_m}\right),$$

gdzie i to numer atrybutu (i=1..n), n to liczba atrybutów, s to numer próbki (s=1..M), M to liczba próbek, x to wartości próbki, $V_{j,i}^{(t+1)}$ to wartość i-tego atrybutu środka j-tej grupy w kolejnej iteracji algorytmu FCM.

Przykład implementacji algorytmu zapisany w pseudokodzie

1. Inicjalizacja algorytmu

1.1 Stworzenie tablic U i D o rozmiarze m x M, gdzie m to liczba klas, a M to liczba próbek (wartość dowolna)

1.2 Utworzenie tablicy V ze środkami grup o rozmiarze m x n, gdzie m to liczba grup, a n to liczba atrybutów (wartość dowolna).

1.3 Wypełnienie tablicy D losowymi wartościami dodatnimi, na przykład z przedziału (0,1; 1,1).

1.4 Wyliczenie każdej wartości w tablicy $U_{j,s}$ (j=1..m; s=1..M).

1.5 Obliczenie środków grup $V_{j,i}$ (j=1..m; i=1..n; m to liczba grup, n to liczba atrybutów)

2. Główna pętla programu wykonywana przez zadaną liczbę iteracji.

2.1. Obliczenie odległości między każdą próbką a grupą: $D_{j,s}$ (j=1..m; s=1..M, m to liczba grup, M to liczba próbek)

2.2. Należy zadbać, aby wszystkie wartości w tablicy D były większe od ustalonej małej wartości (na przykład wszystkie wartości < 1e-5 zastąpić 1e-5).

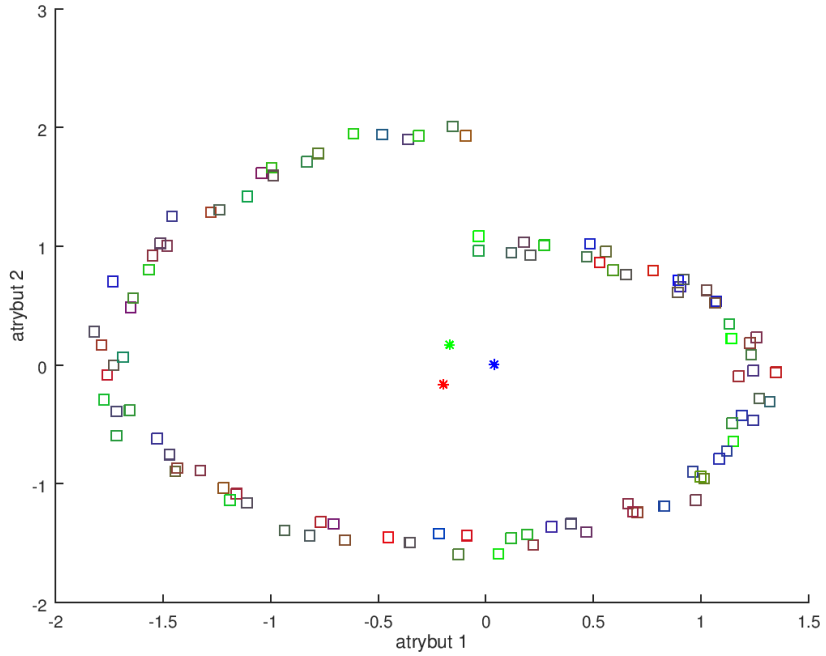
2.3. Wyliczenie stopnia przynależności poszczególnej próbki do każdej grupy: $U_{j,s}$ (j=1..m; s=1..M, m to liczba grup, M to liczba próbek).

2.4. Należy sprawdzić, czy przypadkiem U nie zawiera wartości nieoznaczonych (w takim przypadku należy przerwać program i wyświetlić komunikat ostrzegawczy)

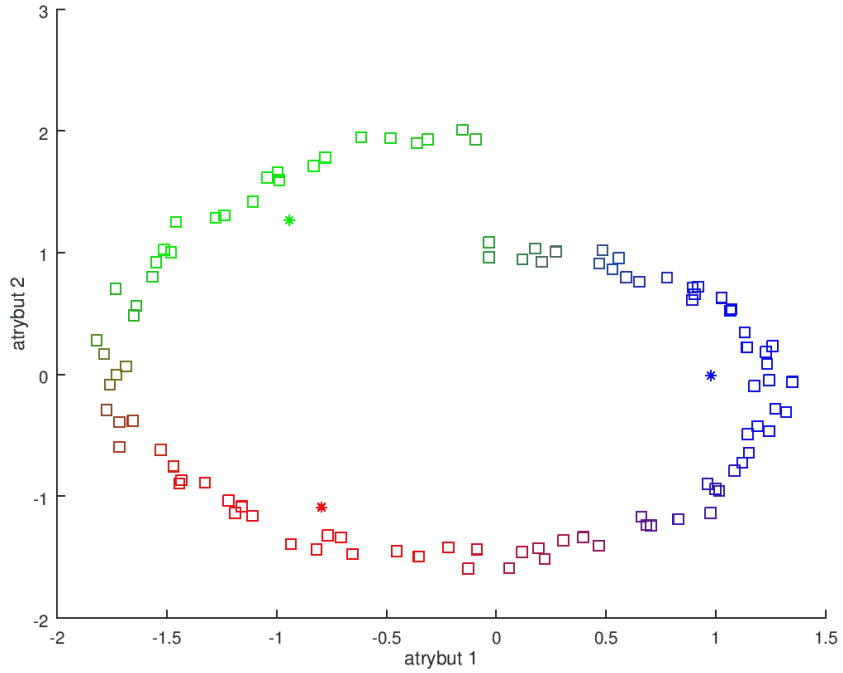
2.5. Obliczenie nowych położeń środków grup $V_{j,i}$ (j=1..m; i=1..n; m to liczba grup, n to liczba atrybutów).

Przykład działania grupowania tych samych próbek przy pomocy FCM na 3 grupy. Na poniższym rysunku widać początkowy rozkład przynależności (przy pomocy kolorów) każdej z próbek (V) do jednej z 3 grup. Użyto to miary odległości euklidesowej oraz parametru fcm_m=2.

Przykład działania przed wykonaniem głównej pętli:



a następnie położenie środków grup i podział na grupy po 10 iteracjach:



Raport

1. Raport (k-średnie)

- Miara odległości euklidesowa,
- liczba grup: 3

- liczba iteracji: 10 (należy podać dane do raportu po 4 i po 10 iteracjach przy pojedynczym uruchomieniu algorytmu)
- zawartość raportu:
  - Położenie każdego ze środków po 4 iteracjach, liczba próbek należących do każdej z tych grup, minimalna i maksymalna wartość obu parametrów (x1, x2) próbek należących dla każdej z tych grup.
  - Wykres zawierający położenie środka oraz zawartość każdej z grup po 4 iteracjach.
  - To samo co uprzednio, ale po 10 iteracjach.

2. Raport (k-średnie)

- Miara odległości następująca: $odleglosc(pktA, pktB) = |pktA_1 - pktB_1|$, czyli brany będzie pod uwagę tylko parametr x1 i ignorowany x2.
- liczba grup: 4
- liczba iteracji: 10
- zawartość raportu:
  - Położenie każdego ze środków po 10 iteracjach, liczba próbek należących do każdej z tych grup, minimalna i maksymalna wartość obu parametrów (x1, x2) próbek należących dla każdej z tych grup.
  - Wykres zawierający położenie środka oraz zawartość każdej z grup po 4 iteracjach.

3. Raport (FCM) **(nieobowiązkowy, na bdb)**

- Miara odległości euklidesowa,
- liczba grup: 3
- współczynnik/parametr fcm_m = 2
- liczba iteracji: 20 (należy podać dane do raportu po 4 i po 20 iteracjach przy pojedynczym uruchomieniu algorytmu)
- zawartość raportu:
  - Stan po 4 iteracjach: położenie każdego ze środków, liczba próbek, które przynależą w stopniu > 0.6 do każdej z tych grup, minimalna i maksymalna wartość obu parametrów (x1, x2) próbek przynależących w stopniu > 0.6 dla każdej z tych grup.
  - To samo co uprzednio, ale po 20 iteracjach.