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".

```
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');
```

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 do której z grup należy (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 polecamy 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

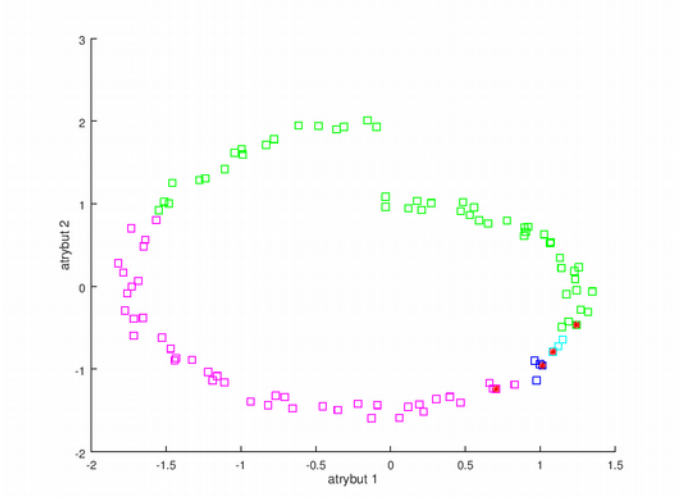
2.2. Pętla po wszystkich m grupach, *j* to indeks aktualnej grupy

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}

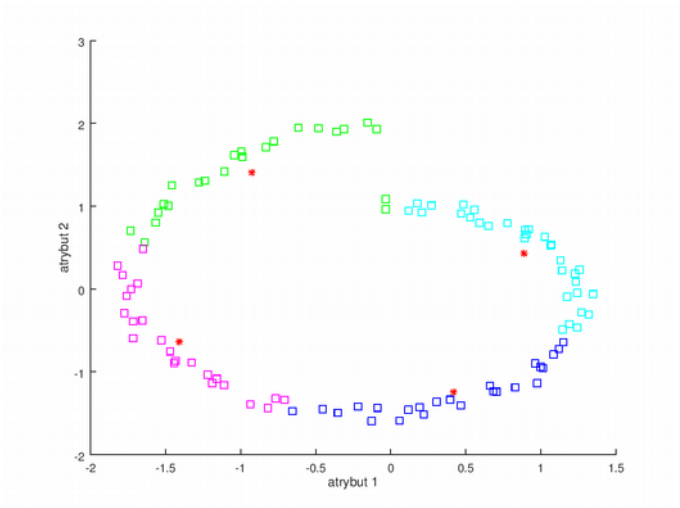
2.2.2. Pętla po wszystkich atrybutach, *i* to index poszczególnego atrybutu

2.2.2.1 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:



Przydatne funkcje: randperm(), rand(), rand('seed', XX)

Algorytm Fuzzy c-Means (FCM) (nieobowiązkowy)

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.

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-10 zastąpić 1e-10).

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).

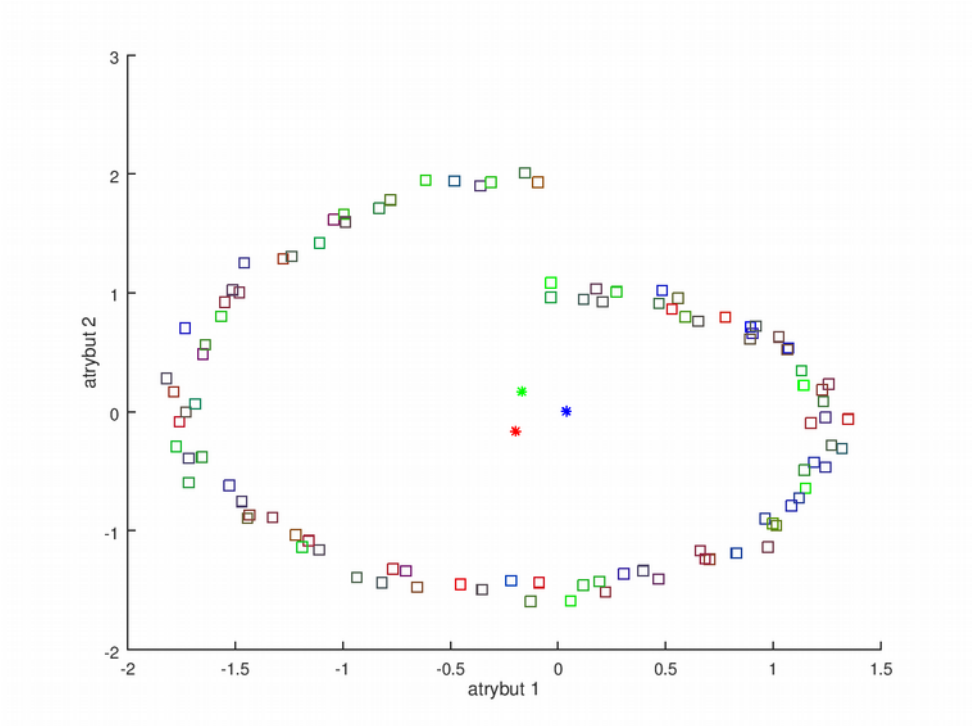
2.4. Należy sprawdzić, czy przypadkiem U nie zawiera wartości nieoznaczonych (w takim przypadku należy przerwać program i wyświetlić komunikat ostrzegawczy)

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

Zadanie do wykonanie:

Proszę pogrupować za pomocą FCM stworzone wcześniej próbki na 3 grupy. Następnie proszę wyświetlić na jednym wykresie próbki oraz środki grup (V). Należy użyć metryki euklidesowej oraz parametru fcm_m=2.

Przykład działania przed wykonaniem głównej pętli:



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