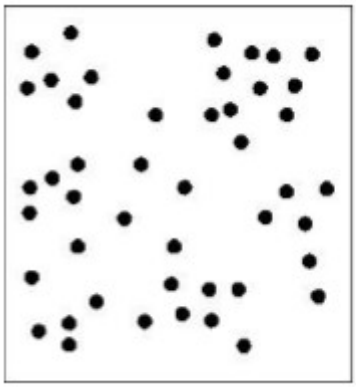
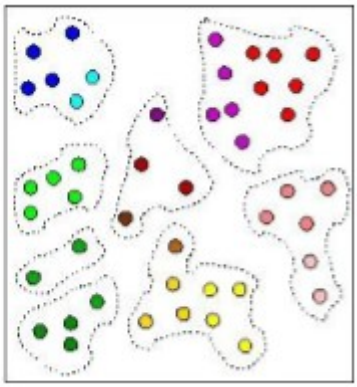
Inteligencja obliczeniowa

Laboratorium 5: Grupowanie.

Tematem dzisiejszych laboratoriów będzie grupowanie (klasteryzacja). W klasteryzacji nie znamy klasy. Algorytmy grupujące same starają się wyodrębnić klasy we wszystkich rekordach starając się jepogrupować, poszukać „skupisk na wykresie”. Rekordy, które są do siebie podobne (leżą blisko siebie) często wpadają do jednego klastra.



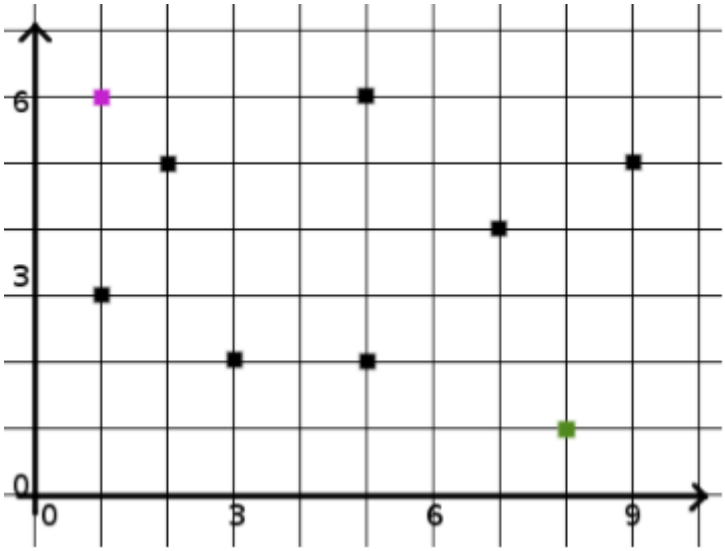


Jedną z metod grupowania jest algorytm k-średnich, który poznamy na tych zajęciach.

Zadanie 1

Algorytm k-średnich działa następująco:
1. Ustal wartość k (liczbę grup),
2. Losowo ustal k poczatkowych srodków grup (centroidy),
3. Dla kazdego rekordu danych znajdz najblizszy centroid (nowy środek grupy)- w ten sposób wszystkie rekordy zostana przydzielone do k grup (klastrów)
4. Dla kazdej z grup znajdz centroid (średnia z rekordów w klastrze) i uaktualnij położenie srodka grupy jako nowa wartosc centroidu
5. Powtarzaj kroki 3-5 dopóki sa zmiany lub nie osiągniemy maks. liczby iteracji.

Dokonaj symulacji powyższego algorytmu na danych z wykresu. Krok 1 został wykonany (wybrano 2 grupy), krok 2 również (losowo wybrano różowy i zielony punkt, nie są one faktycznymi rekordami). Do jakich klastrów zostanie zakwalifikowane 7 rekordów. Dokonaj stosownych obliczeń.



Praca domowa

Zadanie 2
Sprawdzimy, czy algorytm grupowania dobrze podzieli irysy na 3 gatunki, jakimi są w rzeczywistości. Jak sądzisz, uda się? Skorzystaj z tabeli irysów po PCA: 150 rekordów, tylko z dwoma kolumnami PC 1 i PC2.

	PC1	PC2
[1,]	-2.4066389	-0.39695538
[2,]	-2.2235388	0.69018041
[3,]	-2.5811050	0.42754183
[4,]	-2.4508686	0.68600743
[5,]	-2.5368531	-0.50825161
[6,]	-1.8414950	-1.28993811

Do rozwiązania zadania wykorzystaj komendę kmeans:
https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html

Zwizualizuj klastry na wykresie wraz z ich centroidami.

Zadanie 3
Zaimplementuj algorytm k-średnich dla baz danych dwuwymiarowych (dwie kolumny numeryczne). Algorytm powinien wyświetlić zmieniające się centroidy i przyporządkowane im instancje rekordów (klastry). Na koniec algorytm rysuje wykres z klastrami i ostatecznymi centroidami.

Jako dodatkowe wyzwanie można cały przebieg algorytmu zobrazować na wykresie w formacie gif. Każda iteracja to jedna klatka obrazu.

Przetestuj to dla irysów po PCA (dwie kolumny).