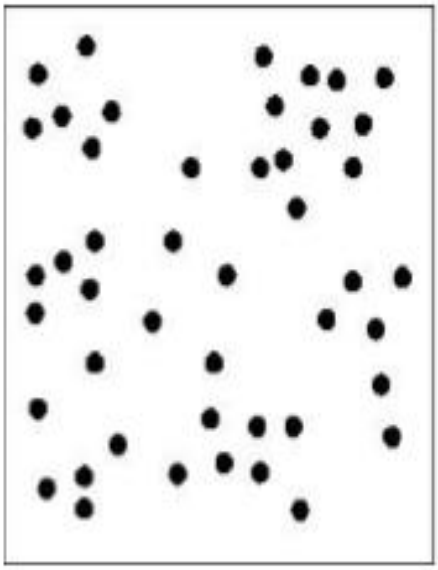
Inteligencja obliczeniowa

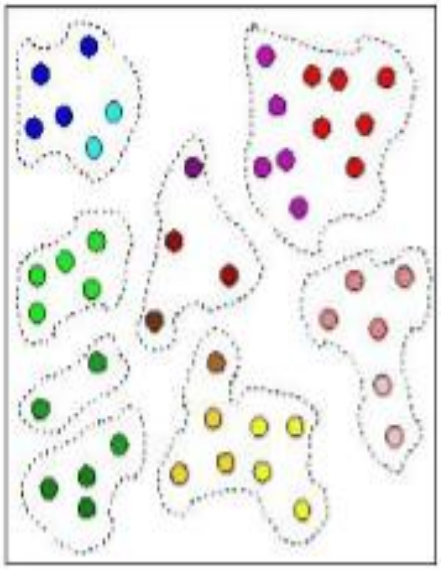
Laboratorium 5: Grupowanie. Reguły asocjacyjne.

Tematem dzisiejszych laboratoriów będą dwie techniki: grupowanie (klasteryzacja) i reguły asocjacyjne.

Grupowanie

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



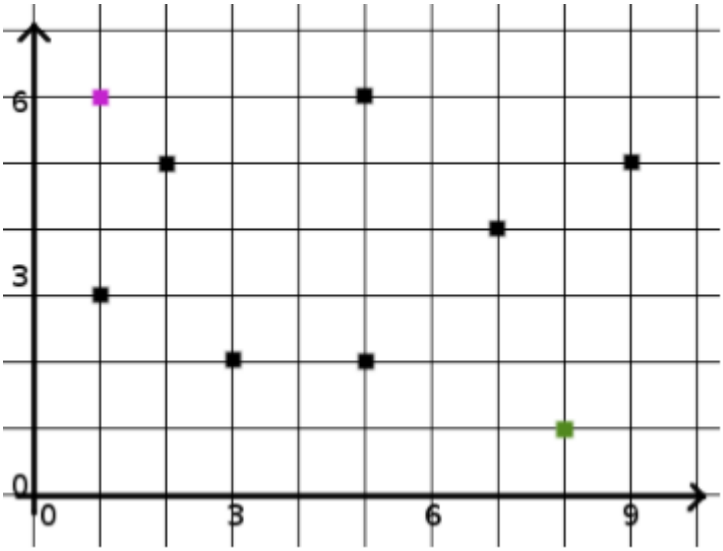


Jedną z metod grupowania jest algorytm k-średnich, który poznamy na tych zajęciach.

Zadanie 1
Algorytm k-średnich działa następująco:

1. Ustal wartość k (liczbę grup),
2. Losowo ustal k początkowych środków grup (centroidy),
3. Dla każdego rekordu danych znajdź najbliższy centroid (nowy środek grupy)- w ten sposób wszystkie rekordy zostaną przydzielone do k grup (klastrów)
4. Dla każdej z grup znajdź centroid (średnia z rekordów w klastrze) i uaktualnij położenie środka grupy jako nowa wartość centroidu.
5. Powtarzaj kroki 3-5 dopóki są 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.



Zadanie 2
Sprawdzimy, czy algorytm grupowania dobrze podzieli irysy na 3 gatunki, jakimi są w rzeczywistości. Jak sądzisz, uda się?

a) Do grupowania 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
```

Przydatne komendy:

```
iris.log <- log(iris[,1:4])
iris.stand <- scale(iris.log, center=TRUE)
iris.pca <- prcomp(iris.stand)
iris.final <- predict(iris.pca)[,1:2]
```

b) Następnie uruchamiamy algorytm k-means dla 3 klastrów i wizualizujemy go na wykresie punktowym. Każdy klaster oznaczony jest innym kolorem i ma zaznaczony swój ostateczny centroid.

Do rozwiązania tego punktu skorzystaj ze strony:
https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html
(na dole strony jest przykład do rozszyfrowania)

c) Czy klastry z faktycznymi gatunkami irysów pokryły się z klastrami wyodrębnionymi przez algorytm k-średnich? Stwórz wykres punktowy, w którym różne kolory będą odpowiadały różnym gatunkom irysów i porównaj z tym z punktu b) powstałym przez grupowanie. Co wstawić w miejsce trzech kropek?

```
plot(iris.final, col=…)
```

Asocjacje

Bywa, że nie interesuje nas ani klasyfikacja, ani grupowanie danych, a szukanie pewnych reguły wewnątrz bazy danych, które nie potrzebują kolumny zwanej klasą.
Weźmy na przykład małą bazę danych kilku transakcji ze sklepu. Ludzi kupowali różne produkty. Każdy zakup odpowiada jednemu rekordowi tabeli.

masło	chleb	ser	piwo	czipsy
TRUE	TRUE	FALSE	FALSE	FALSE
FALSE	TRUE	TRUE	FALSE	FALSE
TRUE	TRUE	TRUE	FALSE	FALSE
FALSE	FALSE	FALSE	TRUE	TRUE
FALSE	TRUE	FALSE	TRUE	FALSE
FALSE	TRUE	FALSE	TRUE	TRUE
TRUE	TRUE	FALSE	TRUE	TRUE
FALSE	TRUE	FALSE	TRUE	TRUE
TRUE	TRUE	TRUE	TRUE	FALSE
TRUE	TRUE	FALSE	TRUE	TRUE

Czy jesteś w stanie zauważyć, czy są w owej tabeli jakieś ukryte powiązania? Może kupno jednego produktu pociąga za sobą kupno innego produktu?

Takie ukryte reguły nazywamy asocjacjami i mają formę:
„Jeśli zachodzi x1 i x2 i … i xk to mamy też y1 i y2 i … i ym."
lub skrótowo:
{x1, x2, …, xk} => {y1, y2, …, ym}

Przykład:
Weźmy asocjację:
A1: „Jeśli klient kupuje masło i chleb, to kupuje też ser."
A1: {masło=TRUE, chleb=TRUE} => {ser=TRUE}

Czy taka asocjacja jest prawdziwa? Jeśli tak, to w jakim stopniu?
Istnieją dwie główne miary, które sprawdzają czy reguła jest znacząca:

- **Support (czyli: wsparcie)**. Obliczamy jak często wystepują w tabeli wspólnie wszystkie elementy reguły, czyli {x1, x2, …, xk, y1, y2, …, ym}. Elementy A1 wystepują w tabeli tylko dwa razy. Wierszy jest 10.

support(A1)=2/10=0.2
Miara ta określa czy nasza reguła jest częsta. Powyższa może nie jest bardzo częsta, co nie znaczy, że nie należy jej brać pod uwagę.

- **Confidence (czyli: wiarygodność).** Obliczamy wg wzoru:

confidence({x1, x2, …, xk}=>{y1, y2, …, ym}) = support({x1, x2, …, xk, y1, y2, …, ym})/support({x1, x2, …, xk})
czyli:
confidence(A1)=support(A1)/support({masło=TRUE, chleb=TRUE}) = 0.2 / 0.5 = 0.4
Miara ta określa w ilu przypadkach reguła ta jest poprawna. Widać, że w 40% przypadków, co jest słabym wynikiem. W trzech przypadkach osoba z chlebem i masłem nie kupiła sera,

a w dwóch przypadkach osoba z serem i masłem kupiła ser.

Zadanie 3
Dla tabeli z kupnem produktów oblicz wsparcie i wiarygodność następujących reguł:

A2: {chleb=TRUE} => {piwo=TRUE,czipsy=TRUE}
A3: {piwo=TRUE} => {czipsy=TRUE}
A4: {czipsy=TRUE} => {piwo=TRUE}
A5: {masło=TRUE, ser=TRUE} => {chleb=TRUE}

Zadanie jest na tyle proste, że możesz policzyć na kartce (lub w arkuszu kalkulacyjnym).

Zadanie 4
Ponad 100 lat temu zatonął Titanic. Na pokładzie znajdowały się tysiące ludzi. Niektórzy przeżyli, inni nie. Czy były jakieś zależności pomiędzy płcią, zamożnością i przeżywalnością? Na podstawie tutorialu:
http://www.rdatamining.com/examples/association-rules
dokonaj szukania asocjacji.

Po przeprowadzeniu eksperymentów wskaż reguły najbardziej wiarygodne i najbardziej interesujące.