Inżynieria danych i modelowanie interdyscyplinarne

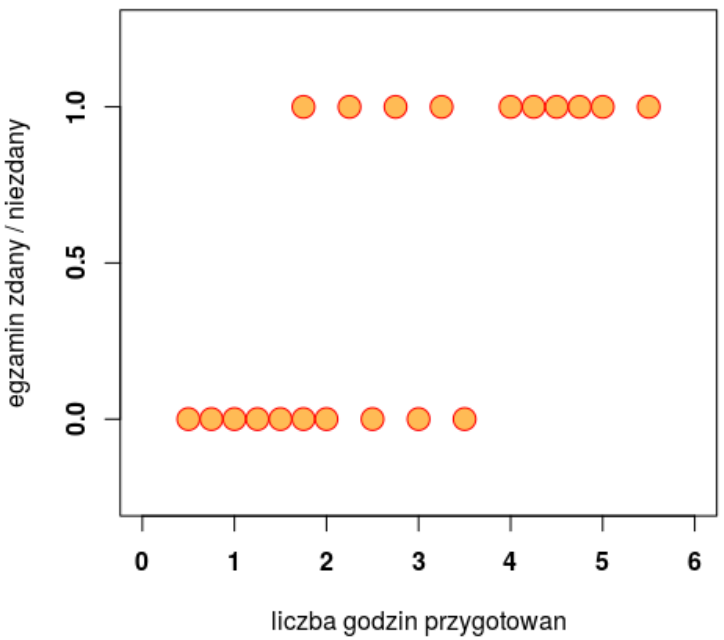
Wykład 9: regresja logistyczna, klasyfikatory, klasyfikator bayesowski, analiza składowych głównych, analiza skupień

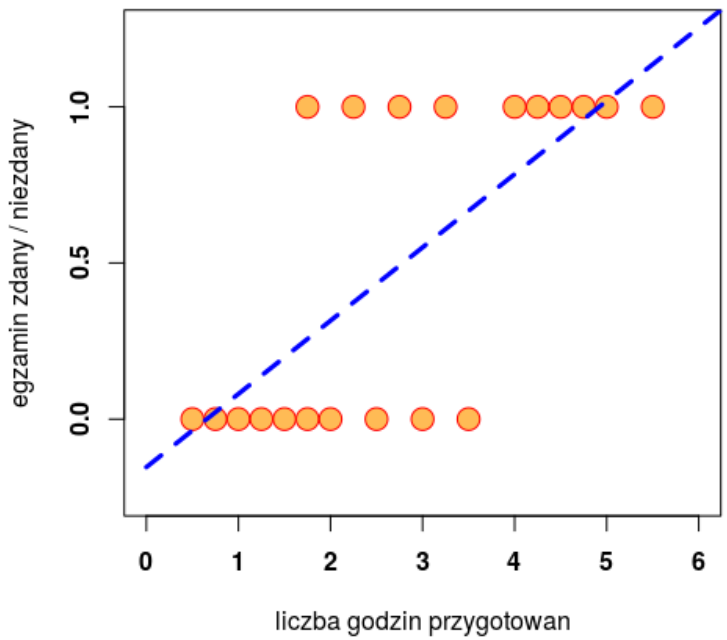
dr hab. inż. Julian Sienkiewicz, prof. ucz.

20 maja 2026

Przykład

Rozpatrzmy następujący przypadek: zależność sukcesu podczas egzaminu od liczby godzin spędzonych na przygotowaniach.





Jak widać, regresja liniowa nie jest adekwatnym narzędziem do modelowania tego typu zależności.

Zmienne

- zakładamy, że Y jest **binarną** zmienną odpowiedzi (zmienną objaśnianą),
- $Y = 1$ oznacza występowanie danej cechy, $Y = 0$ — jej brak,
- $\boldsymbol{X} = (X_1, X_2, ..., X_p)$ jest zbiorem zmiennych objaśniających, które mogą być zarówno ciągłe jak i dyskretne

Model

Model może zostać zapisany jako

$$\pi_i = \Pr(Y_i = 1 | X_i = x_i) = \frac{\exp(\boldsymbol{X}_i\beta)}{1 + \exp(\boldsymbol{X}_i\beta)} = \frac{1}{1 + \exp(-\boldsymbol{X}_i\beta)}$$

lub też w tożsamej formie

$$\text{logit}(\pi_i) = \ln \frac{\pi_i}{1 - \pi_i} = \boldsymbol{X}_i\beta$$

Założenia

- dane $Y_1, ..., Y_n$ są niezależne (tzn poszczególne przypadki są niezależne),
- Y_i pochodzi z rozkładu dwumianowego,
- brak założenia o liniowości pomiędzy zmienną objaśnianą a zmiennymi objaśniającymi, zakłada się liniową zależność pomiędzy funkcją logitową odpowiedzi a $\boldsymbol{X}\beta$,
- podobnie jak w regresji liniowej, poszczególnymi zmiennymi mogą być również potęgi lub inne nieliniowe funkcje oryginalnych zmiennych,
- **nie** musi być spełniona **jednorodność wariancji**,
- niepewności (odchyłki) muszą być niezależne, ale nie muszą pochodzić z rozkładu normalnego

Estymator największej wiarygodności

Estymator największej wiarygodności (MLE) otrzymujemy korzystając z rozkładu prawdopodobieństwa zmiennej objaśnianej – w tym przypadku jest to rozkład dwumianowy, stąd łączny rozkład gęstości prawdopodobieństwa to

$$L(\boldsymbol{y}|\beta) = \prod_{i=1}^{N} \frac{n_i!}{y_i!(n_i - y_i)!} \pi^{y_i} (1 - \pi_i)^{n_i - y_i}.$$

Funkcja wiarygodności ma taką samą postać, jedynie interesuje nas w tym momencie wyrażenie nieznanych wartości β jako funkcji znanych, ustalonych y_i, tzn $L(\beta|\boldsymbol{y})$. Czynnik, który nie zawiera π_i można pominąć:

$$L(\beta|\boldsymbol{y}) \sim \prod_{i=1}^{N} \left(\frac{\pi_i}{1 - \pi_i} \right)^{y_i} (1 - \pi_i)^{n_i} = \prod_{i=1}^{N} \mathrm{e}^{x_i^T \beta y_i} (1 + \mathrm{e}^{x_i^T \beta})^{-n_i}$$

W efekcie logarytm funkcji wiarygodności to

$$l(\beta) = \sum_{i=1}^{N} x_i^T \beta y_i - \sum_{i=1}^{N} n_i \ln(1 + \mathrm{e}^{x_i^T \beta})$$

Błąd standardowy β_i

Aby obliczyć błąd standardowy β_i korzystamy z następującej relacji

$$\sigma_{\beta_i} = \sqrt{(\boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X})^{-1}_{ii}},$$

przy czym $\boldsymbol{X}$ oraz $\boldsymbol{V}$ są dane jako

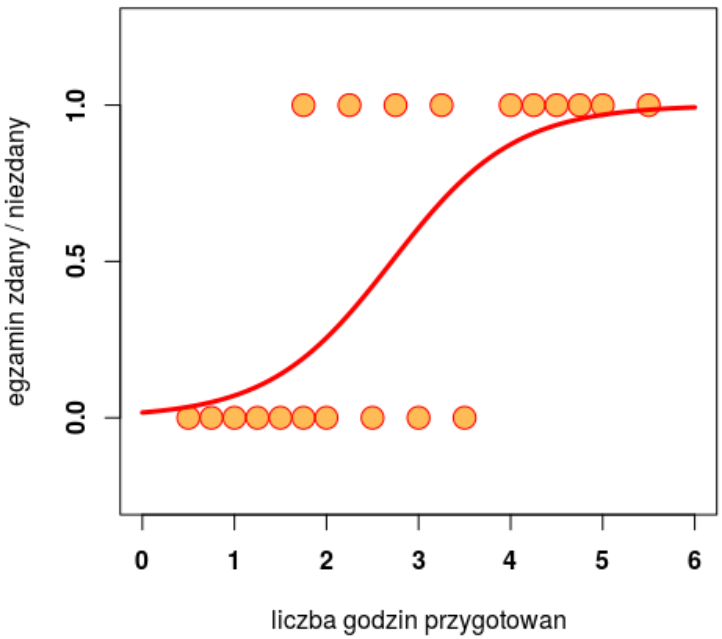
$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{Np} \end{pmatrix}$$

$$\boldsymbol{V} = \begin{pmatrix} \pi_1(1-\pi_1) & 0 & \cdots & 0 \\ 0 & \pi_2(1-\pi_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pi_N(1-\pi_N) \end{pmatrix}$$

Błąd standardowy β_i

W celu oceny istotności współczynników β_i możemy obliczyć statystykę β_i/σ_{β_i} i porównać ją rozkładem normalnym tzn. wyznaczyć p-wartość $p_{\beta_i} = 2[1 - F_z(|\beta_i/\sigma_{\beta_i}|)]$, gdzie F_z to dystrybuanta rozkładu normalnego.

Wracamy do poczatkowego przykładu.



$$\boldsymbol{\beta} = \begin{pmatrix} -4.08 \\ 1.50 \end{pmatrix}$$

$$\boldsymbol{\sigma}_{\boldsymbol{\beta}} = \begin{pmatrix} 1.76 \\ 0.63 \end{pmatrix}$$

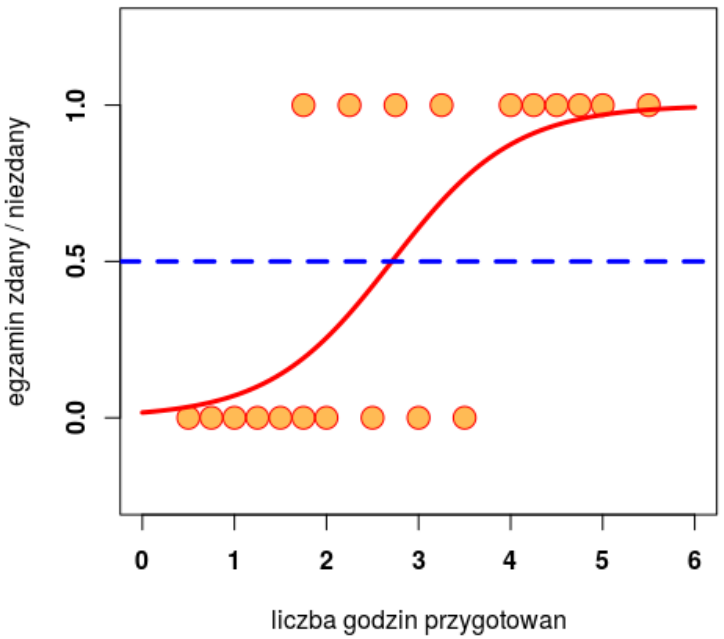
$$\boldsymbol{p}_{\boldsymbol{\beta}} = \begin{pmatrix} 0.021 \\ 0.017 \end{pmatrix}$$

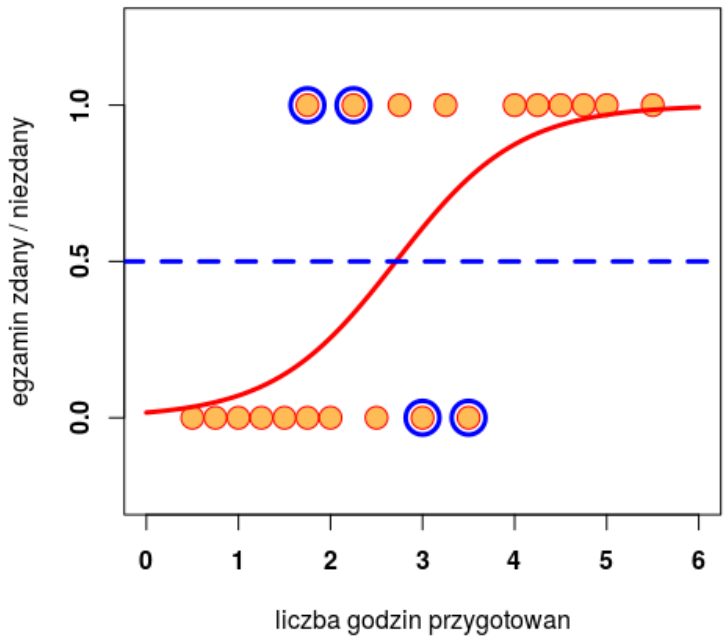
Kolorem czerwonym jest oznaczone $\pi_i = 1/[1 + \exp(-\beta_0 - \beta_1 x_i)]$.

Przykład

W jaki jednak sposób powinniśmy sprawdzić, czy przewidywania modelu pokrywają się z zaobserwowanymi danymi? Ponieważ π_i są wartościami **prawdopodobieństwa**, możemy po prostu przyjąć:

- $\pi_i \geq 1/2$ to 1,
- $\pi_i < 1/2$ to 0.





Macierz pomyłek

Wygodną metodą oceny jakości jest tzw. macierz pomyłek (*confusion matrix*), w której konfrontujemy wyniki działania modelu z rzeczywistością zliczając wszystkie możliwe pary dane–model:

- dane: 0, model: 0,
- dane: 0, model: 1,
- dane: 1, model: 0,
- dane: 1, model: 1,

		Klasa predykowana – wynik testu	
	Populacja	Klasa pozytywna	Klasa negatywna
Klasa rzeczywista	Stan pozytywny	**Prawdziwie pozytywna, TP**	**Fałszywie negatywna** (błąd drugiego rodzaju, **FN**)
	Stan negatywny	**Fałszywie pozytywna** (błąd pierwszego rodzaju, **FP**)	**Prawdziwie negatywna, TN**

Rodzaje błędów

- **Błąd pierwszego rodzaju** – błąd polegający na odrzuceniu hipotezy zerowej, która w rzeczywistości nie jest fałszywa. Oszacowanie prawdopodobieństwa popełnienia błędu pierwszego rodzaju oznaczamy symbolem α i nazywamy poziomem istotności testu
- **Błąd drugiego rodzaju** (błąd przyjęcia) – błąd polegający na nieodrzuceniu hipotezy zerowej, która jest w rzeczywistości fałszywa. Oszacowanie prawdopodobieństwa popełnienia błędu drugiego rodzaju oznaczamy symbolem β, a jego dopełnienie do jedności nazywane jest mocą testu
- **Błąd trzeciego, czwartego rodzaju**... :-)

Nasz przykład

	model: 1	model: 0
dane: 1	**8**	**2**
dane: 0	**2**	**8**

- dokładność (*accuracy*)
  $ACC = (TP + TN)/(TP + TN + FP + FN)$
- czułość (*sensitivity, true positive rate*)
  $TPR = TP/(TP + FP)$
- specyficzność (*specifity, true negative rate*)
  $TNR = TN/(TN + FN)$

Macierz pomyłek

		Predicted condition			
	Total population $= P + N$	**Predicted positive**	**Predicted negative**	Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Real Positive (P)** [a]	**True positive** (TP), hit[b]	**False negative** (FN), miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate type II error [c] $= \frac{FN}{P} = 1 - TPR$
	Real Negative (N)[d]	**False positive** (FP), false alarm, overestimation	**True negative** (TN), correct rejection[e]	False positive rate (FPR), probability of false alarm, fall-out type I error [f] $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{TP + FP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{TN + FN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{TP + FP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{TN + FN} = 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-} = \frac{TP \times TN}{FP \times FN}$
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2\,PPV \times TPR}{PPV + TPR} = \frac{2\,TP}{2\,TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	*phi* or Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$

Sources: [12][13][14][15][16][17][18][19] view · talk · edit

Klasyfikatory

- regresja logistyczna jest przykładem **klasyfikatora**,
- klasyfikacja to jedno z zadań **uczenia pod nadzorem**, gdzie mamy dostęp do poprawnie sklasyfikowanego zbiór danych (tzn. dane, o których wiemy, że należą do konkretnej klasy), czyli tzw. zbioru treningowego
- zadaniem w problemie klasyfikacyjnym jest przypisanie klasy do **nowych, nieoznaczonych obserwacji** na podstawie dostępnych danych,
- istnieje cała gama klasyfikatorów: począwszy od właśnie regresji logistycznej, poprzez metody związane z podejściem bayesowskim, maszyny wektorów nośnych aż do sieci neuronowych,
- klasyfikatory są dobierane na podstawie dopasowania do konkretnego przypadku — zwykle interesuje nas skuteczność (lub też czułość lub specyficzność)

Klasyfikator bayesowski

Zaobserwowany wektor $\mathbf{x}$ klasyfikujemy jako pochodzący z tej klasy k, dla której wartość

$$p(k|\mathbf{x}) \qquad k = 1, ..., g$$

jest największa.

$p(k|\mathbf{x})$ - prawdopodobieństwo, że obserwacja $\mathbf{x}$ pochodzi z klasy k

Twierdzenie Bayesa

$$p(k|\mathbf{x}) = \frac{\pi_k p(\mathbf{x}|k)}{\sum_{r=1}^{g} \pi_r p(\mathbf{x}|r)} = \frac{\pi_k p(\mathbf{x}|k)}{p(\mathbf{x})}$$

- $p(\mathbf{x}|k)$ - rozkład obserwacji $\mathbf{x}$ z klasy k,
- π_k - prawdopodobieństwo *a priori*, że obserwacja pochodzi z klasy k,
- $p(\mathbf{x})$ - prawdopodobieństwo pojawienia się obserwacji $\mathbf{x}$.

Dla dwóch klas ($g = 2$)

$$\frac{p(2|\mathbf{x})}{p(1|\mathbf{x})} = \frac{\pi_2 p(\mathbf{x}|2)}{\pi_1 p(\mathbf{x}|1)}$$

Załóżmy teraz, że składowe wektora $\mathbf{x} = [x^1, x^2, ..., x^p]^T$ są niezależne

$$p(\mathbf{x}|k) = \prod_{i=1}^{p} p(x^{(i)}|k)$$

Wtedy

$$\frac{p(2|\mathbf{x})}{p(1|\mathbf{x})} = \frac{\pi_2 \prod_{i=1}^{p} p(x^i|2)}{\pi_1 \prod_{i=1}^{p} p(x^i|1)}$$

Po zlogarytmowaniu

$$\ln \frac{p(2|\mathbf{x})}{p(1|\mathbf{x})} = \ln \frac{\pi_2}{\pi_1} + \sum_{i=1}^{p} \ln \frac{p(x^i|2)}{p(x^i|1)}$$

Prowadzi to do następującej reguły klasyfikacyjnej:

Naiwny klasyfikator bayesowski dla $g = 2$

Jeżeli

$$\ln \frac{\pi_2}{\pi_1} + \sum_{i=1}^{p} \ln \frac{p(x^i|2)}{p(x^i|1)} > 0$$

sklasyfikuj obserwacje $\mathbf{x}$ do klasy 2, w przeciwnym razie - do klasy 1.

Cechy naiwnego klasyfikatora bayesowskiego:

- **addytywny** wpływ kolejnych składowych wektora obserwacji **x**
- wartość prawdopodobieństw $p(x^i|k)$ są nieznane i estymujemy je jako

$$\hat{p}(x^i = l|k) = \frac{n_{ik}(l)}{n_{ik}}$$

  - n_{ik} - liczba wszystkich obserwacji w próbie uczącej z klasy k, dla których została zmierzona wartość i-tej składowej,
  - $n_{ik}(l)$ - liczba wszystkich obserwacji w próbie uczącej z klasy k, dla których została zmierzona wartość i-tej składowej, która przyjęła poziom l

- podobnie $\hat{\pi}_k$ estymujemy jako $\frac{n_k}{n}$, gdzie n liczba wszystkich obserwacji, n_k - liczba obserwacji z klasy k

Naiwny klasyfikator bayesowski - przykład

Kodowanie

- Wilgotność powietrza:
  - 0-30 % (mała),
  - 31-70 % (średnia),
  - 71-100 % (duża),
- Ciśnienie atmosferyczne:
  - $<$ 990 hPa (niskie),
  - 990-1010 hPa (średnie),
  - $>$ 1010 hPa (duże),
- Klasy:
  - klasa 1 (pogoda),
  - klasa 2 (deszcz),

dzień	wilgotność	ciśnienie	klasa
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			

Naiwny klasyfikator bayesowski - przykład

	wilgotność		ciśnienie	
	☀	🌧	☀	🌧
niskie	2	2	1	4
średnie	2	4	3	2
wysokie	1	1	1	1
Σ	5	7	5	7

	wilgotność		ciśnienie	
	☀	🌧	☀	🌧
niskie	2/5	2/7	1/5	4/7
średnie	2/5	4/7	3/5	2/7
wysokie	1/5	1/7	1/5	1/7
Σ	1	1	1	1

Estymacja prawdopodobieństw a priori: $\hat{\pi}_1 = \frac{5}{12}$, $\hat{\pi}_2 = \frac{7}{12}$.

Pojawia się nowa obserwacja $\mathbf{x} = [H = niska, p = wysokie]^T$. Do której klasy należy ją sklasyfikować?

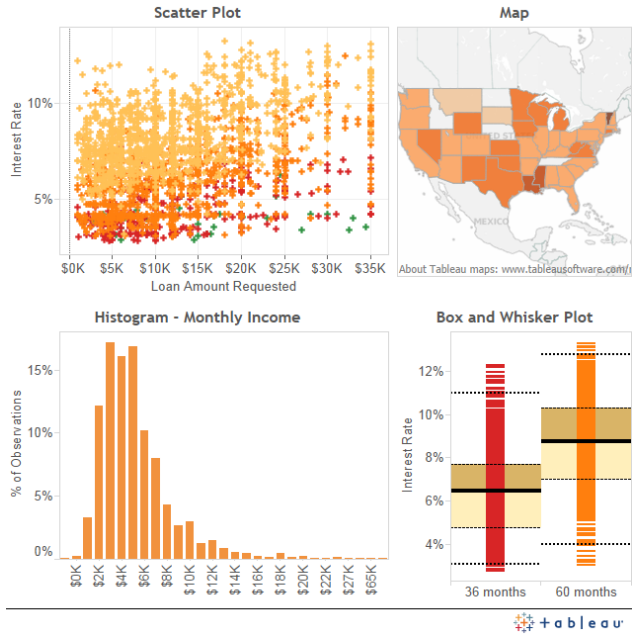
$$\ln \frac{p(2|\mathbf{x})}{p(1|\mathbf{x})} = \ln \frac{\hat{\pi}_2}{\hat{\pi}_1} + \ln \frac{\hat{p}(H = niska|2)}{\hat{p}(H = niska|1)} + \ln \frac{\hat{p}(p = wysokie|2)}{\hat{p}(p = wysokie|1)}$$

$$\ln \frac{p(2|\mathbf{x})}{p(1|\mathbf{x})} = \ln \frac{\frac{7}{12}}{\frac{5}{12}} + \ln \frac{\frac{2}{7}}{\frac{2}{5}} + \ln \frac{\frac{1}{7}}{\frac{1}{5}} = \ln \frac{7}{5}\frac{2}{7}\frac{5}{2}\frac{1}{7}\frac{5}{1} = \ln \frac{5}{7} < 0$$

☀

Na początkowym etapie analizy danych zaczyna się zwykle od najprostszych metod: jest to **eksploracyjna analiza danych** (*exploratory data analysis - EDA*). Składają się na nią m. in.:

- histogramy,
- wykresy punktowe (*scatterplot*),
- wykresy pudełkowe [skrzynkowe] (*box plot*),



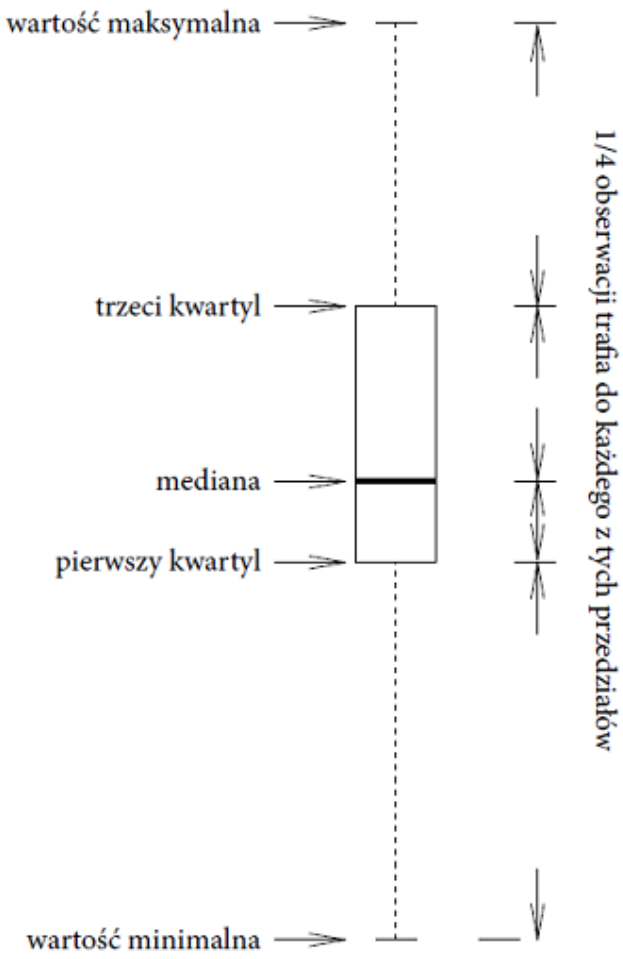
Ogólna definicja Johna Tukeya z lat 60-tych XX w. określa EDA jako

procedury analizowania danych, metody interpretowania wyników takich procedur, sposoby projektowania zbierania danych w celu uczynienia ich analizy łatwiejszą lub bardziej dokładną, oraz wszystkie techniki statystyczne które mogą być stosowane w badaniu danych.

- Sens EDA jest szczególnie widoczny, gdy **nie posiadamy hipotezy badawczej** — często jest tak, że owa hipoteza jest nietrafiona lub też zawęża nasz obszar badań,
- EDA jest, kolokwialnie mówiąc, **grzebaniem się w danych** w celu wytypowania istotnych obserwacji, faktów i cech rozpatrywanego układu,
- zaleca się, aby w trakcie EDA przy korzystaniu ze **statystyk opisowych** odchodzić od średniej i odchylenia na rzecz statystyk odpornościowych (*robust statistics*), takich jak np. mediana,
- Tukey w latach 70-tych XX w. podkreślał, że wywierany jest zbyt duży nacisk na **testowanie hipotez** bezpośrednio na danych,
- postulowane jest wykonanie EDA **w celu** zaproponowania hipotez badawczych

Wykres pudełkowy

- wiek XX przyniósł, prawdziwy rozkwit grafiki statystycznej,
- nowe sposoby pozyskiwania danych, nowe typy danych, nowe problemy badawcze dostarczają nowych wyzwań dotyczących prezentacji danych,
- jednym z nich jest **wykres pudełko z wąsami**, zwany też **wykresem pudełkowym** lub wykresem skrzynkowym, pierwszy raz opublikowany w 1977 r. przez **Johna Tukeya** (1915–2000),
- przedstawia tak zwane pięć liczb Tukeya, czyli brzegi przedziałów dzielących zakres zmienności na cztery równoliczne części

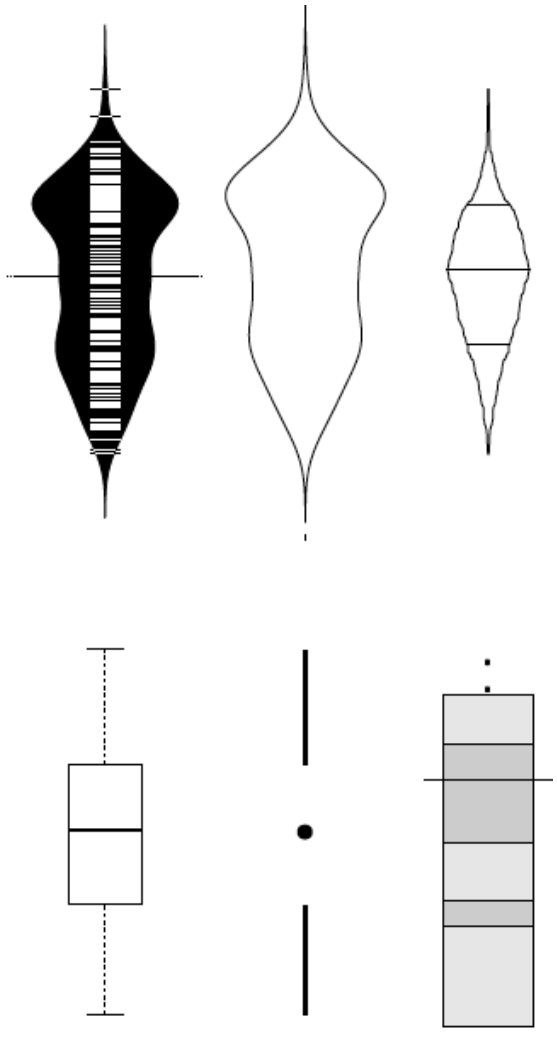


Wykres pudełkowy

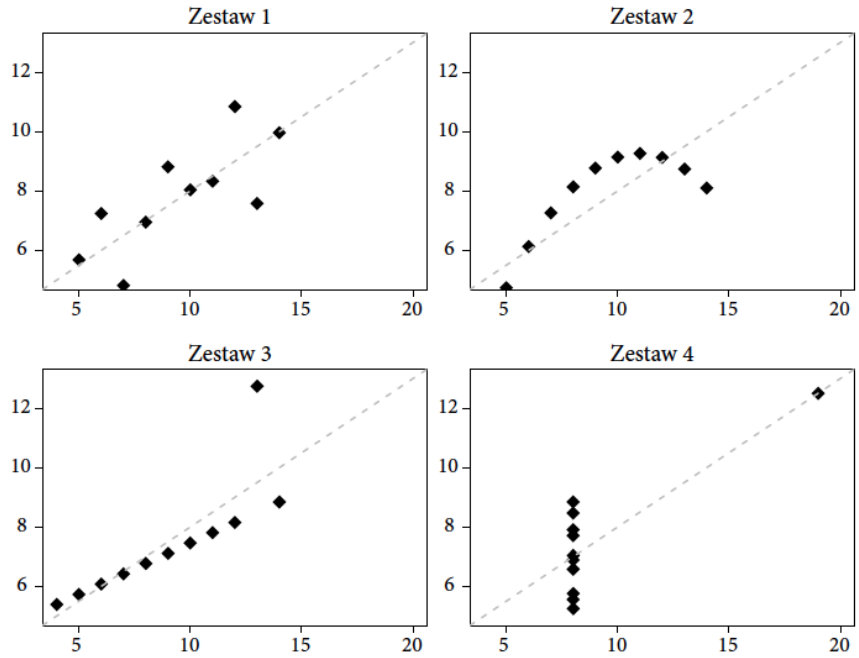
Modyfikacje wykresu pudełkowego

- **wykres fasolkowy** przedstawiający jądrową ocenę gęstości, poszczególne obserwacje i ich średnią (ang. *bean plot*),
- **wykres skrzypcowy** prezentujący jądrową ocenę gęstości (ang. *violin plot*),
- **wykres percentylowy** pokazujący percentyle (ang. *percentile plot*)

- zwykły **wykres pudełkowy** (ang. *boxplot*),
- wersja minimalistyczna wykresu pudełkowego, opracowana przez Edwarda Tuftego
- **wykres natężeniowy HDR** prezentujący obszary o wysokim zagęszczeniu obserwacji (ang. *highest density regions*)



Modyfikacje wykresu pudełkowego



Kwartet Anscombe'a

Średnie na osi poziomej w każdym zestawie danych wynoszą $\bar{x} = 9$, ich wariancje $\sigma_x = 11$. Na osi pionowej $\bar{y} = 7.5$ oraz $\sigma_y = 4.12$. Korelacje zmiennych wynoszą $r = 0.816$, wszystkie dane najlepiej opisuje ten sam model regresji liniowej $y = 3 + x/2$.

Kwartet Anscombe'a

- **Frank Anscombe** (1918–2001) w 1973 roku skonstruował i przedstawił cztery zbiory danych nazywane dziś **kwartetem Anscombe'a**,
- zestawy są dobrane tak, by podstawowe statystyki liczbowe były dla nich identyczne,
- proste przedstawienie graficzne ilustruje oczywiste różnice pomiędzy nimi, co daje wspaniały przykład **wartości wizualizacji danych**

Analiza składowych głównych

Analiza składowych głównych (PCA - *principal component analysis*) polega ona na znalezieniu **nowego kierunku**, który **maksymalizuje wariancję** zrzutowanych na niego obserwacji. Następnie szukamy kolejnego kierunku, również o jak największej wariancji, tyle, że **ortogonalnego** do poprzedniego etc. Okazuje się, że takie cechy odpowiadają wektorom własnym związanym z kolejnymi **wartościami własnymi** (począwszy od największej).

Twierdzenie o składowych głównych

Niech **x** będzie wektorem losowym o wektorze wartości oczekiwanych **m** i macierzy kowariancji **S** i nich wartości własne tej macierzy, λ_i, $i = 1, \ldots, p$, spełniają warunek:

$$\lambda_1 \geq \lambda_2 \geq \ldots \gamma_p > 0.$$

Wektor $\gamma_{(i)}$ i-tej składowej głównej wektora **x**

$$y_i \equiv \gamma_{(i)}^T(\mathbf{x} - \mathbf{m})$$

$i = 1, \ldots, p$, jest równy i-temu wektorowi własnemu macierzy **S**, odpowiadającemu wartości własnej λ_i.

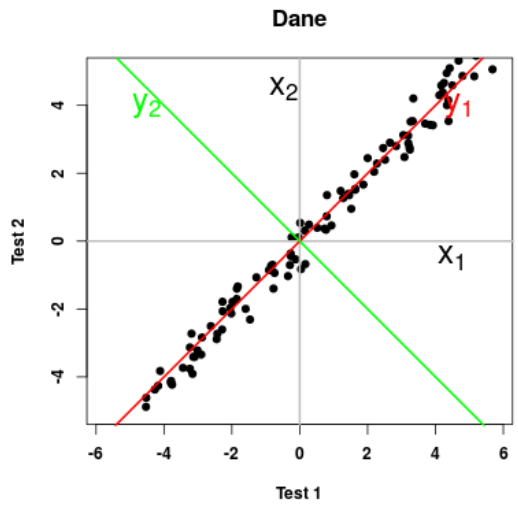
Wszystkie p składowe główne tworzą wektor $\mathbf{y}$ postaci

$$\mathbf{y} = \mathbf{\Gamma}^T(\mathbf{x} - \mathbf{m})$$

czyli jest to przekształcenie wektora losowego $\mathbf{x}$ w wektor $\mathbf{y}$, polegające kolejno na

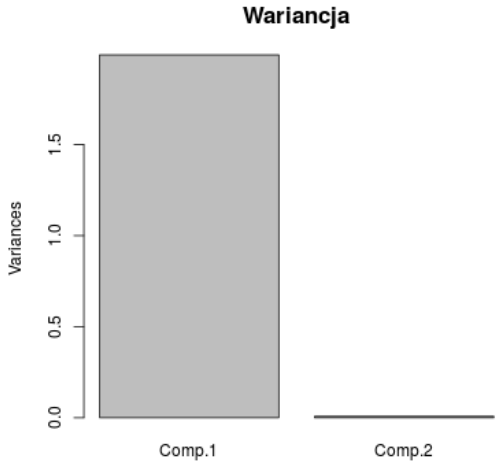
1. przesunięciu wektora $\mathbf{x}$ o jego wartość oczekiwaną (scentrowania wektora),
2. liniowym przekształceniu scentrowanego wektora za pomocą macierzy ortogonalnej $\mathbf{\Gamma}^T$ (geometrycznie jest obrócenie oryginalnego układu współprzędnych o pewien kąt)

Osie nowego układu współrzednych , wyznaczone przez wektory ładunków $\gamma_{(i)}$ są tak dobrane, aby maksymalizować wariancje rzutów oryginalnych wektorów losowych na te osie, pod warunkiem, że kolejne rzuty nie są skorelowane z wcześniejszymi.

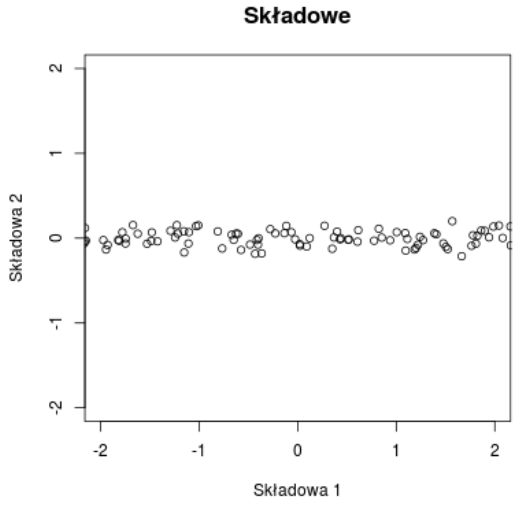


Przykład

Prosty, dwuwymiarowy przykład: obserwacją była para punktów zdobytych przez studenta w dwóch testach. Widać wyraźną zależność pomiędzy jednym a drugim testem.



$$\text{ścieżka} = a\,x + b\,y$$



Do czego przydaje się analiza składowych głównych?

- umożliwia **redukcję wymiaru**: w ten sposób jesteśmy w stanie określić, które składowe (czyli **kolumny** strukturyzowanych danych) są nieistotne,
- kryterium redukcji jest dość proste: suma wszystkich wartości własnych macierzy kowariancji **S** jest równa wariancji poszczególnych współrzędnych wektora **x**; stąd wielkość

$$\frac{\lambda_1 + \ldots + \lambda_k}{\lambda_1 + \ldots + \lambda_p} 100\%$$

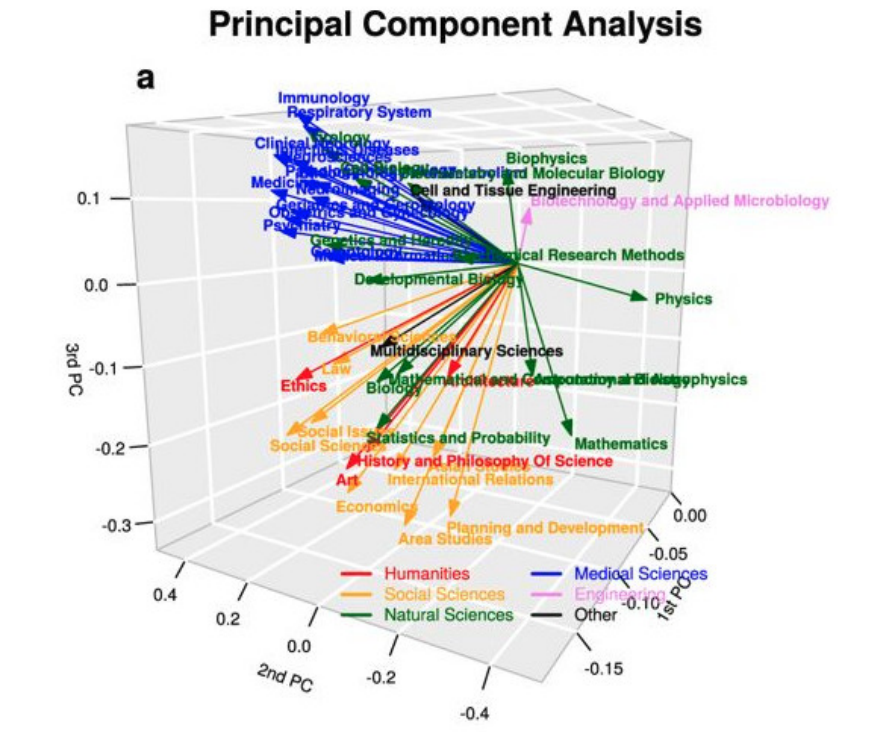
  $k \leq p$ wyraża **procent zmienności** wektora losowego **x** wyjaśniony przez k pierwszych składowych głównych, gdy przez zmienność całkowitą rozumie się sumę wariancji.
- nowe składowe (**składowe główne**) są **kombinacją liniową** oryginalnych kierunków w danych — w ten sposób możemy określić sens (interpretację) nowego kierunku

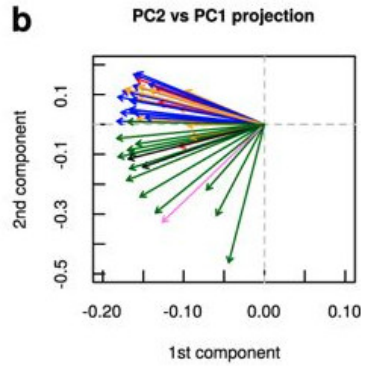
Dane

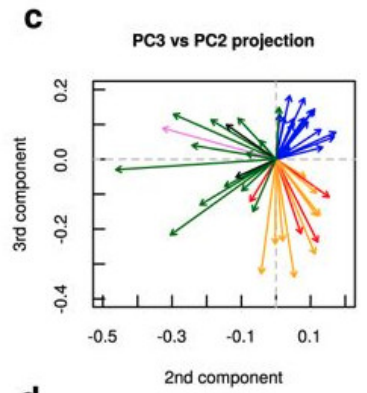
Dane uzyskane z serwisu **Web of Science** (WOS):

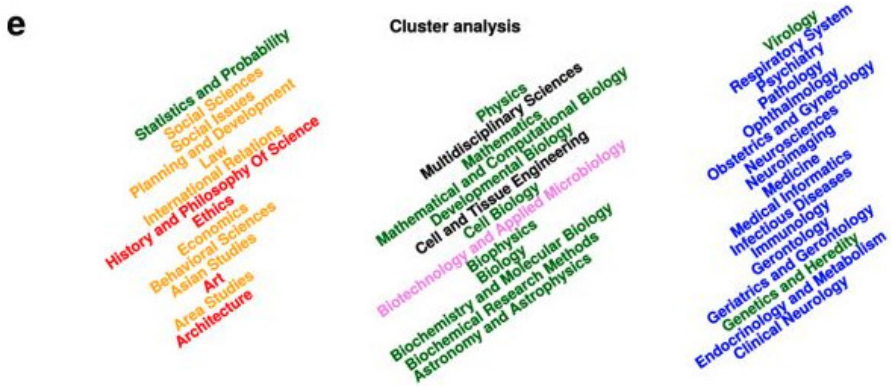
- **lata 2008-2009**,
- około **500.000** rzetelnych **rekordów** (problemy z nazwami uniwersytetów etc),
- dla każdego zebranego artykułu przynależność do **kategorii naukowej** (jednej ze 181),
- w efekcie dla każdej ze **100 najlepszych uczelni** na świecie (m.in. MIT, Stanford etc) liczba publikacji w każdej z dziedzin,
- **ograniczenie** tylko do tych kategorii, które są skorelowane na wysokim poziomie istotności statystycznej,
- macierz 100 x 45 (uniwersytety x istotne kategorie) – wykonanie PCA na takim zestawie danych

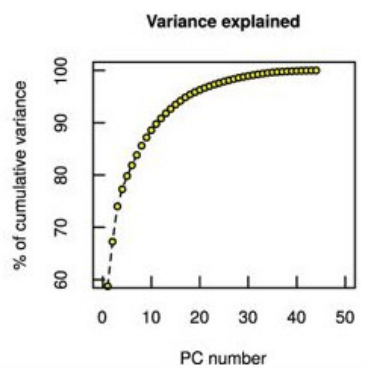
Przykład PCA: publikacje w serwisie Web of Science











Cele i ogólny opis

- **analiza skupień** ma na celu wykrycie w zbiorze obserwacji struktur zwanych **skupieniami**... czyli rozłącznych podziorów zbioru obserwacji, wewnątrz których obserwacje są w jakimś określonym sensie bliskie,
- podzbiory różne są od siebie odległe (w porównaniu z elementami wewnątrz każdego podzbioru),
- jest to przypadek **klasyfikacji bez nadzoru**, czyli **nie** mamy próby uczącej, ani też wiedzy o tym, jak przypisać klasę do obserwacji,

Analiza skupień w przestrzeni euklidesowej R^p

- mamy n-elementowy zbiór obserwacji $\mathbf{x}_i$, $i = 1, ..., n$ o wartościach w R^p,
- chcemy podzielić tę próbę na K skupień,

$$T = \frac{1}{2} \sum_{i=1}^{n} \sum_{i'=1}^{n} d_{ii'}$$

suma kwadratów odległości pomiędzy parami punktów próby

$$d_{ii'} = d(\mathbf{x}_i, \mathbf{x}_{i'})$$

kwadrat odległości pomiędzy obserwacjami $\mathbf{x}_i$ i $\mathbf{x}_{i'}$

- w ten sposób dokonaliśmy arbitralnego podziału obserwacji na K rozłącznych podzbiorów $k = 1, .., K$, gdzie oznaczymy $C(i) = k$ jako przynależność i-tej obserwacji $\mathbf{x}_i$ do k-tego podzbioru.

Analiza skupień w R^p

Sumę kwadratów T można rozłożyć na sumę kwadratów odległości pomiędzy parami **należącymi do tego samego skupienia** oraz parami **należącymi do różnych skupień**

$$T = W + B$$

$$W = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} d_{ii'}$$

W - within the cluster
B - between clusters

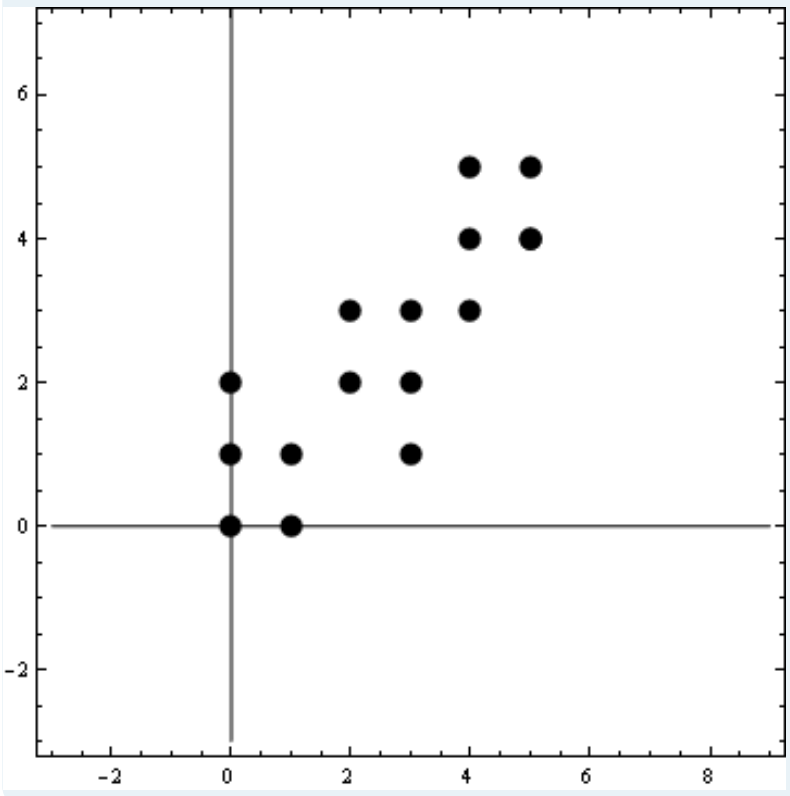
$$B = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')\neq k} d_{ii'}$$

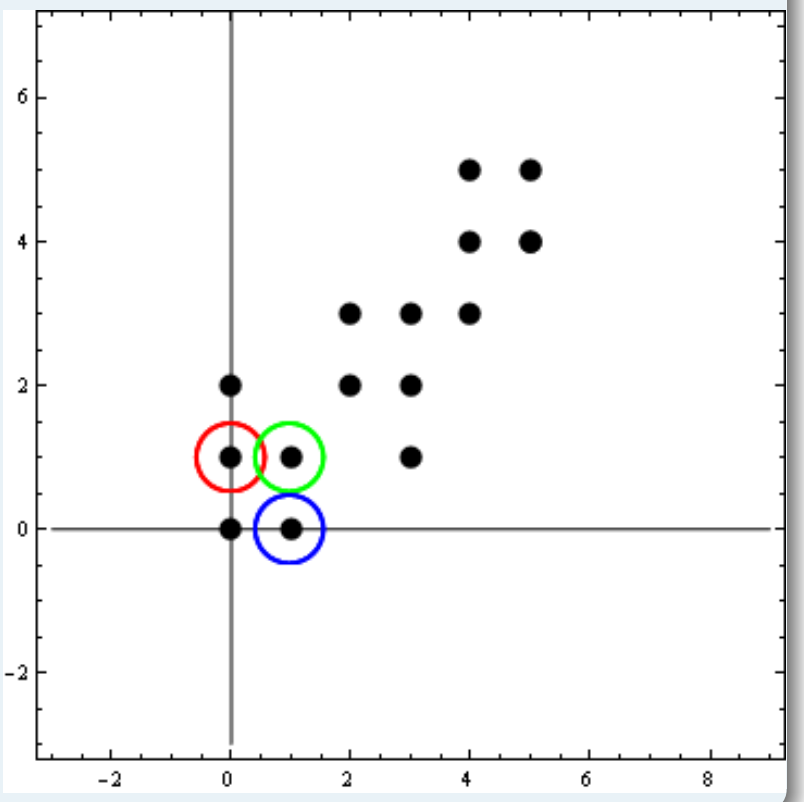
Zmieniając podział punktów na K skupień, zmieniamy W i B (T jest takie samo). Czyli analiza skupień to minimalizacja rozrzutu punktów wewnatrz skupień — minimalizacja W (maksymalizacja B)

Algorytm K-średnich (K-means)

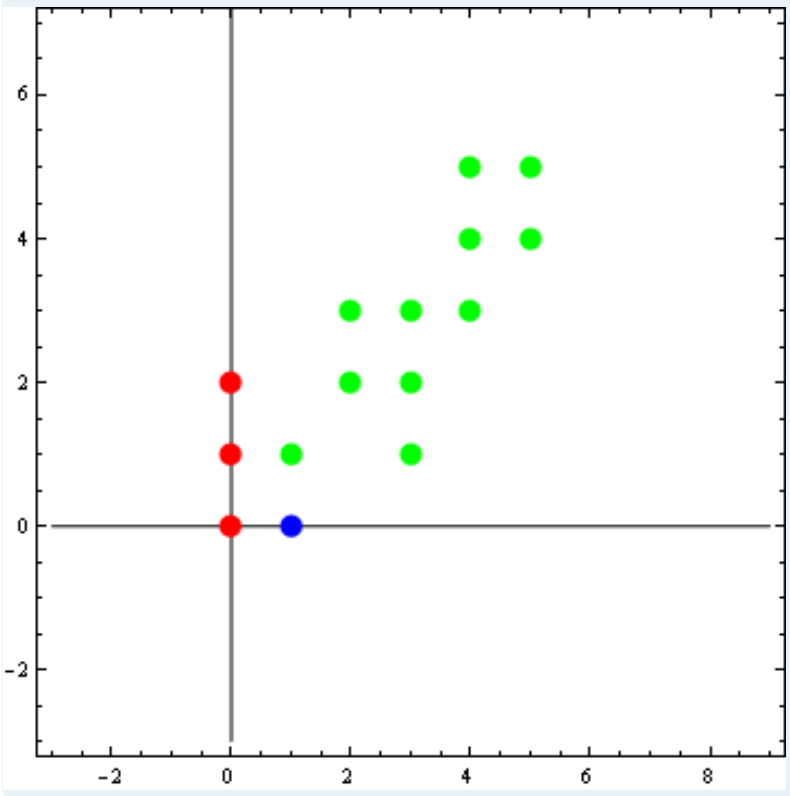
0. inicjalizacja początkowych K środków $\mathbf{m}_K$,
1. w pierwszym kroku przypisujemy punkty do najbliższych środków $\mathbf{m}_k$
2. obliczamy nowe środki skupień i wracamy do kroku 1
3. kontynuujemy iteracje, dopóki żaden punkt nie przeniesie się z jednego skupienia do drugiego

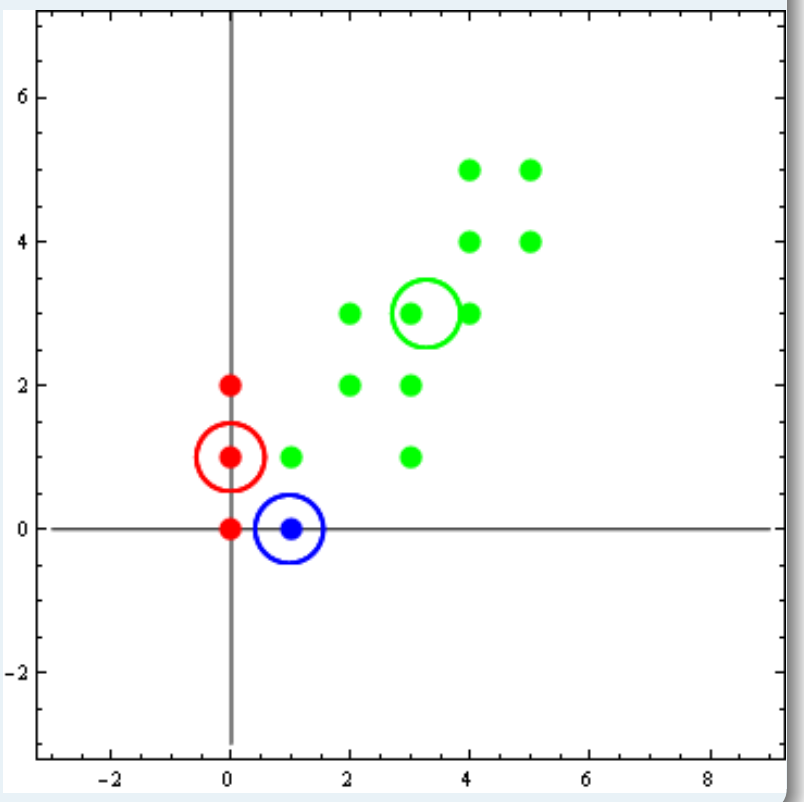
Przykład 1 - 3 klastry, inicjalizacja 1 (kroki 1-2)



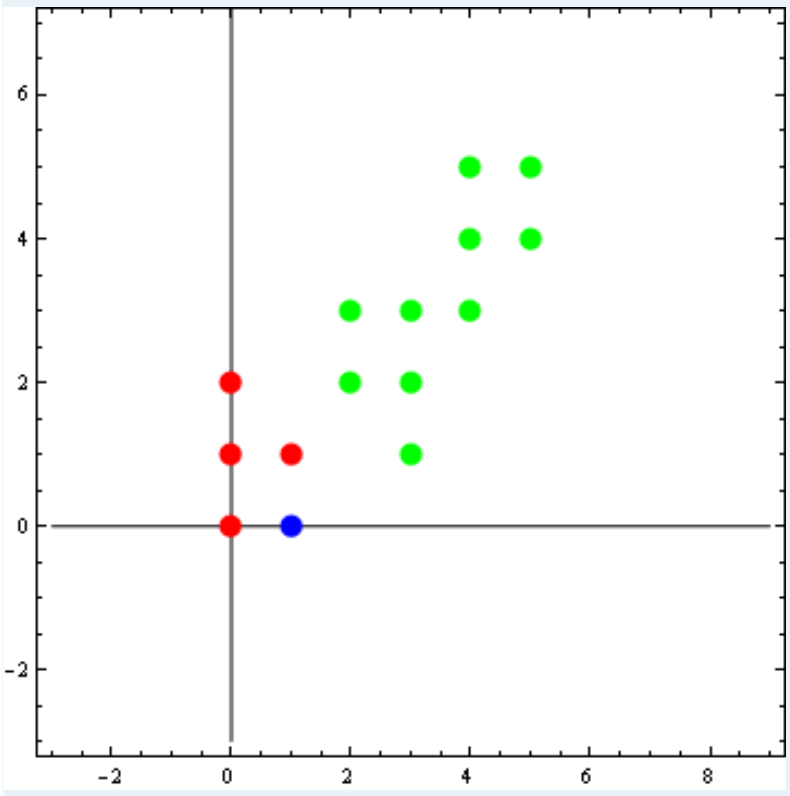


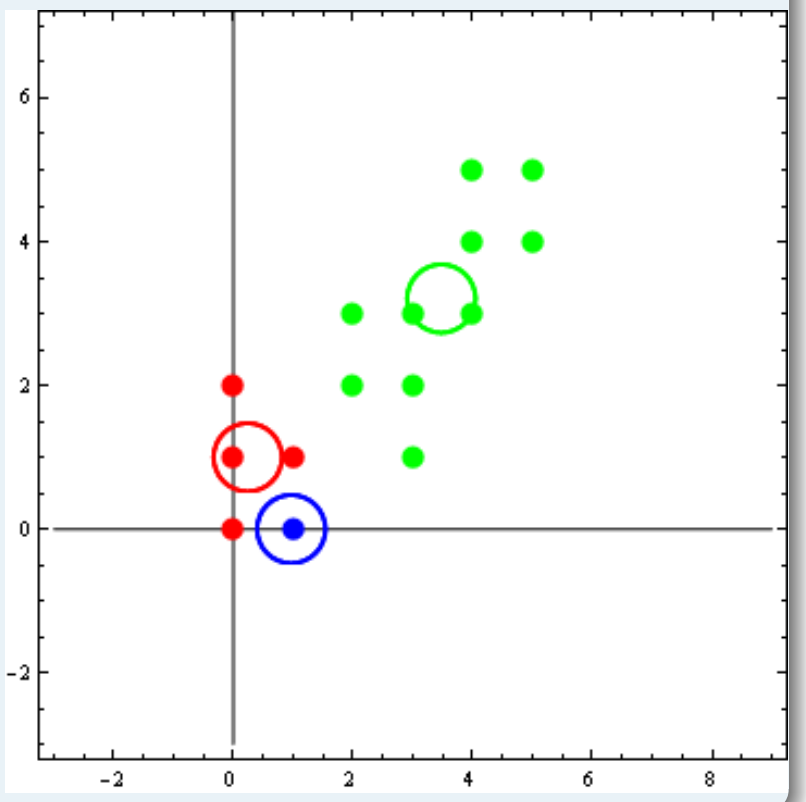
Przykład 1 - 3 klastry, inicjalizacja 1 (kroki 3-4)



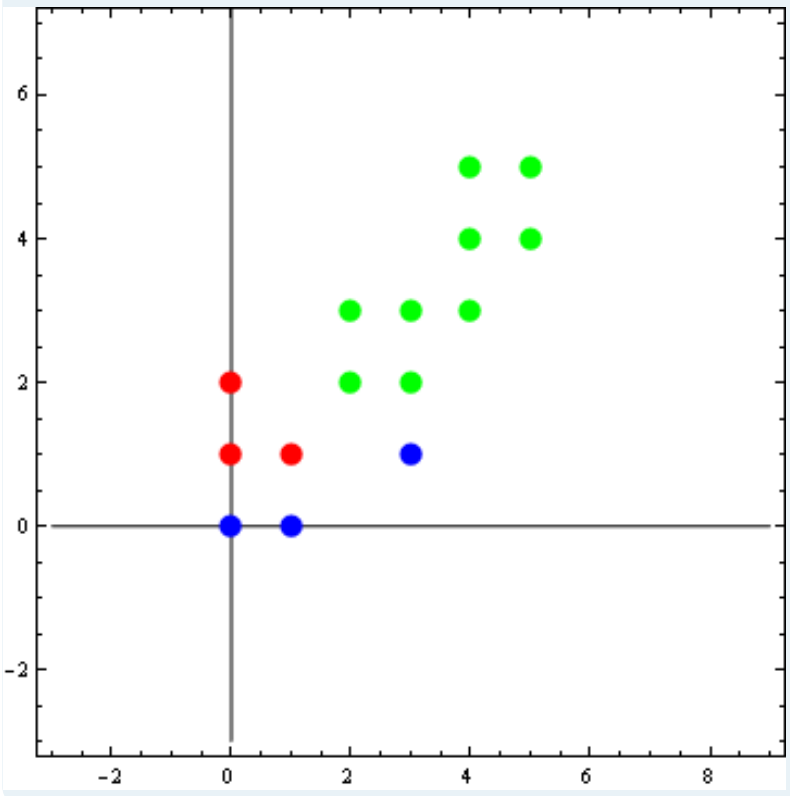


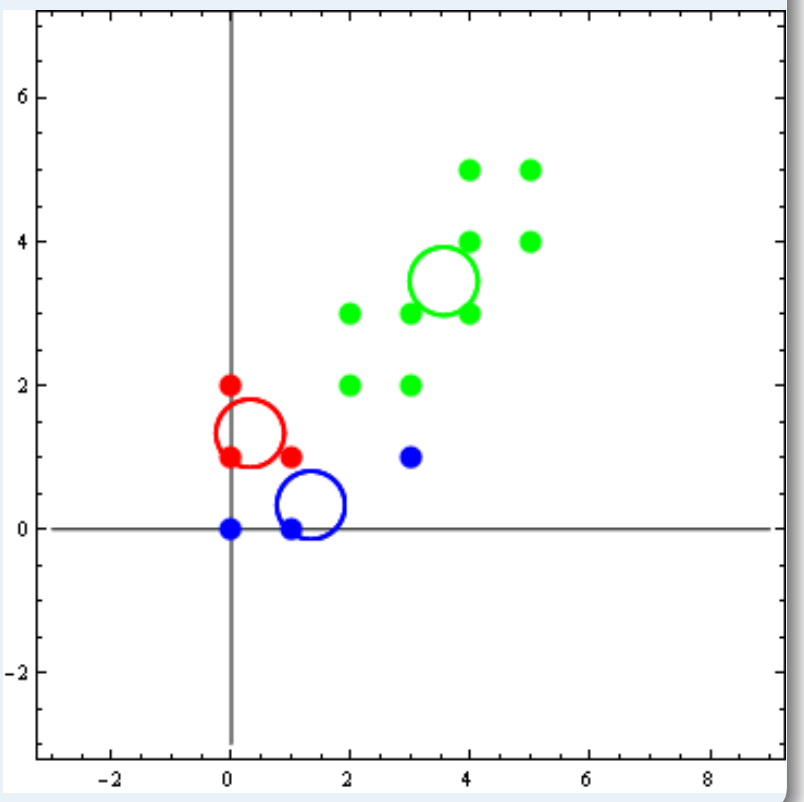
Przykład 1 - 3 klastry, inicjalizacja 1 (kroki 5-6)



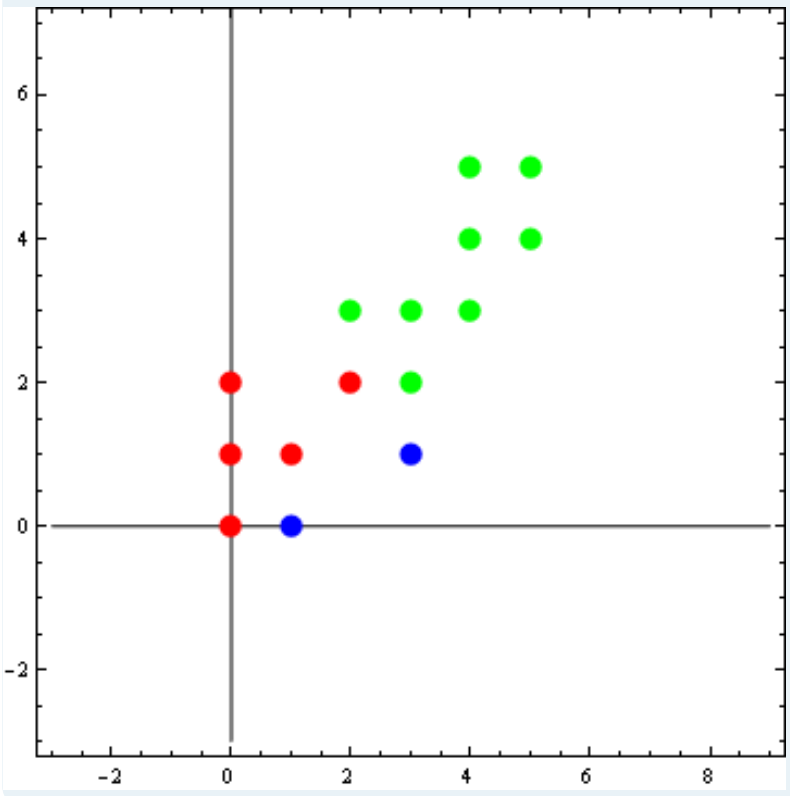


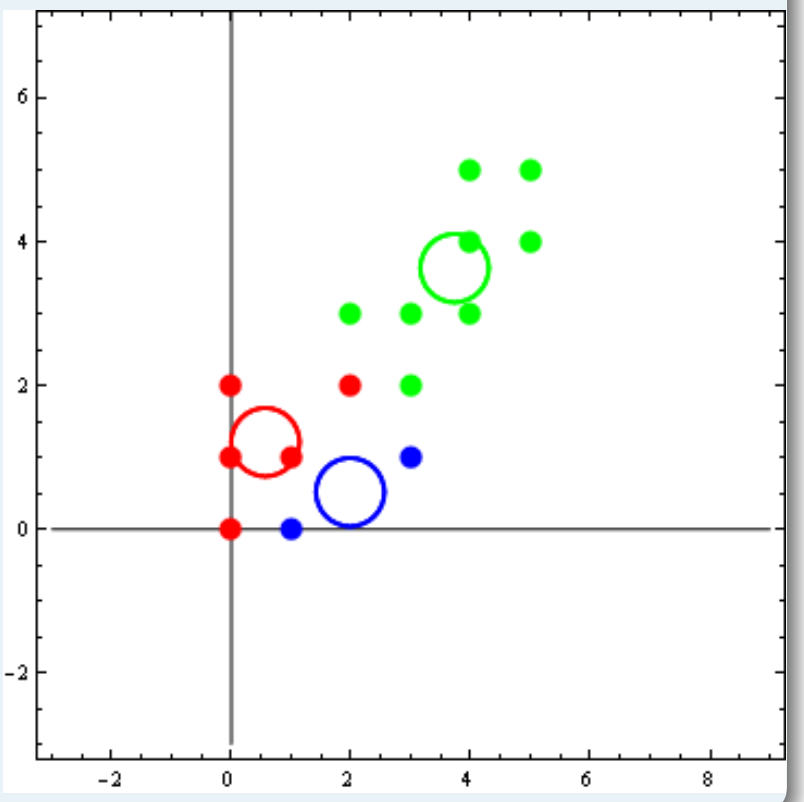
Przykład 1 - 3 klastry, inicjalizacja 1 (kroki 7-8)



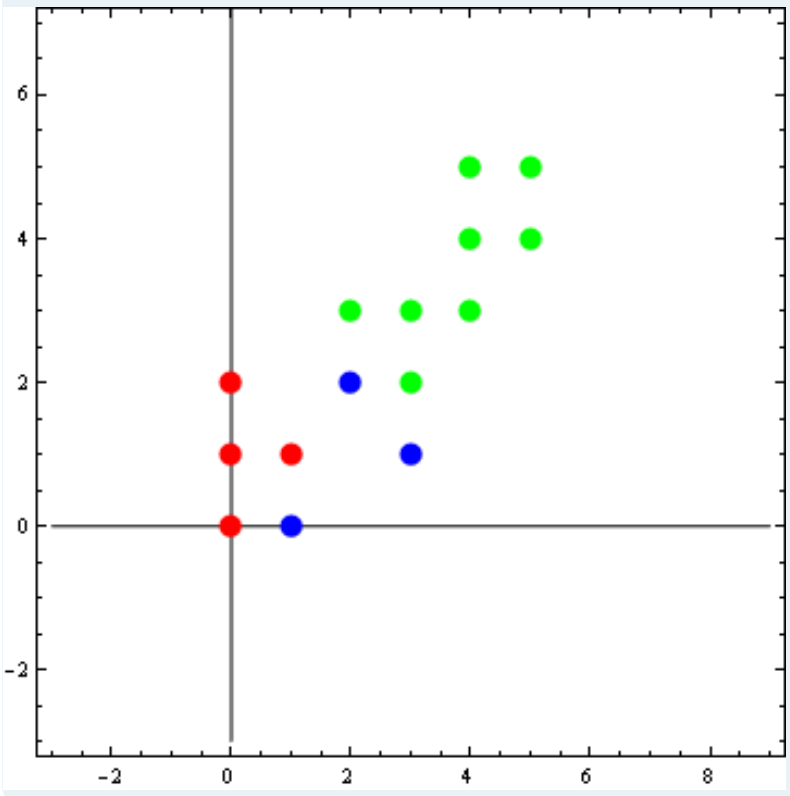


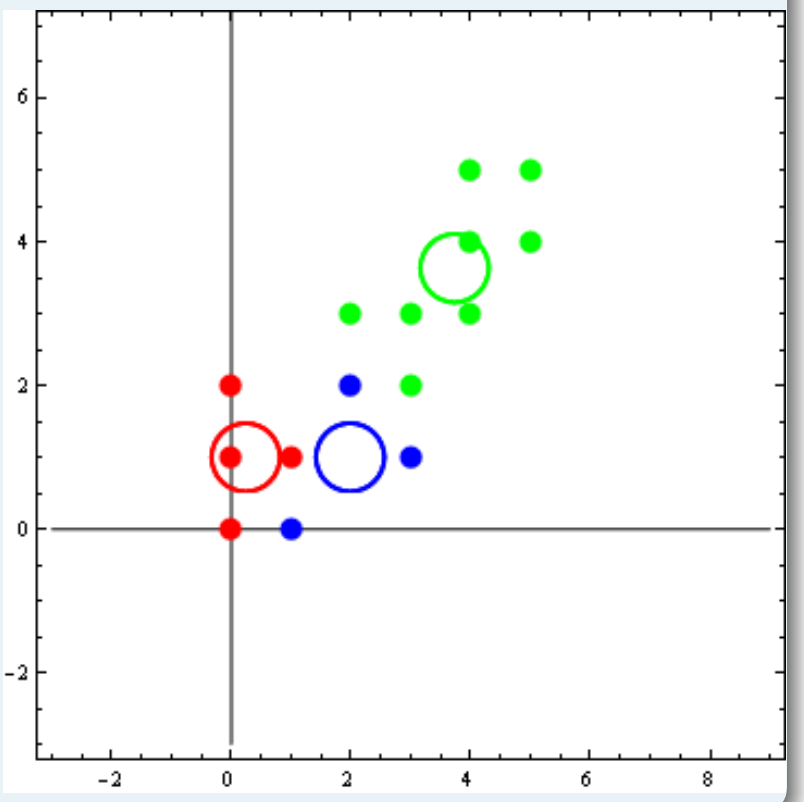
Przykład 1 - 3 klastry, inicjalizacja 1 (kroki 9-10)



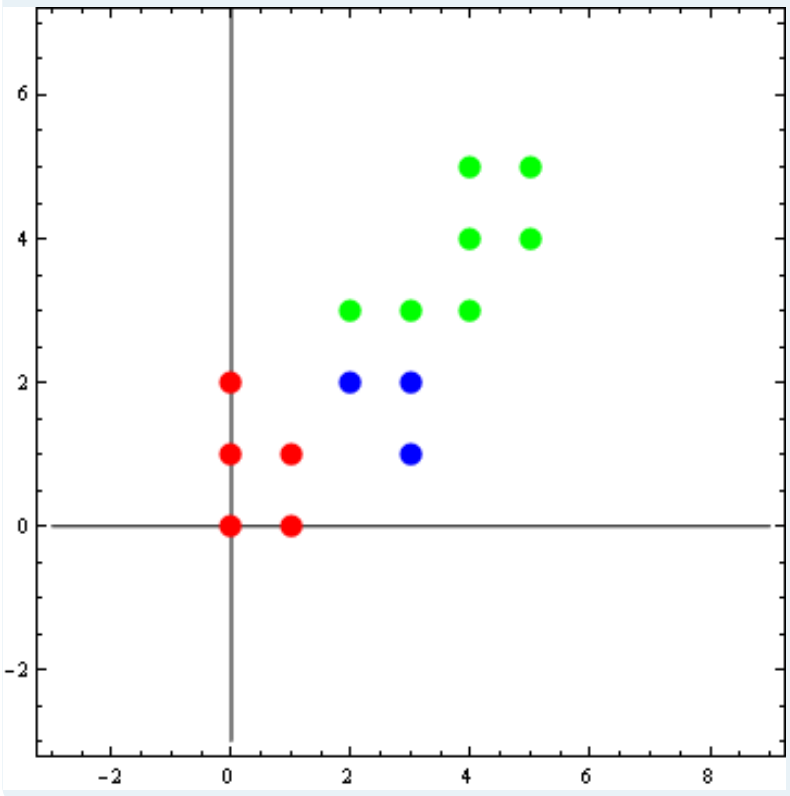


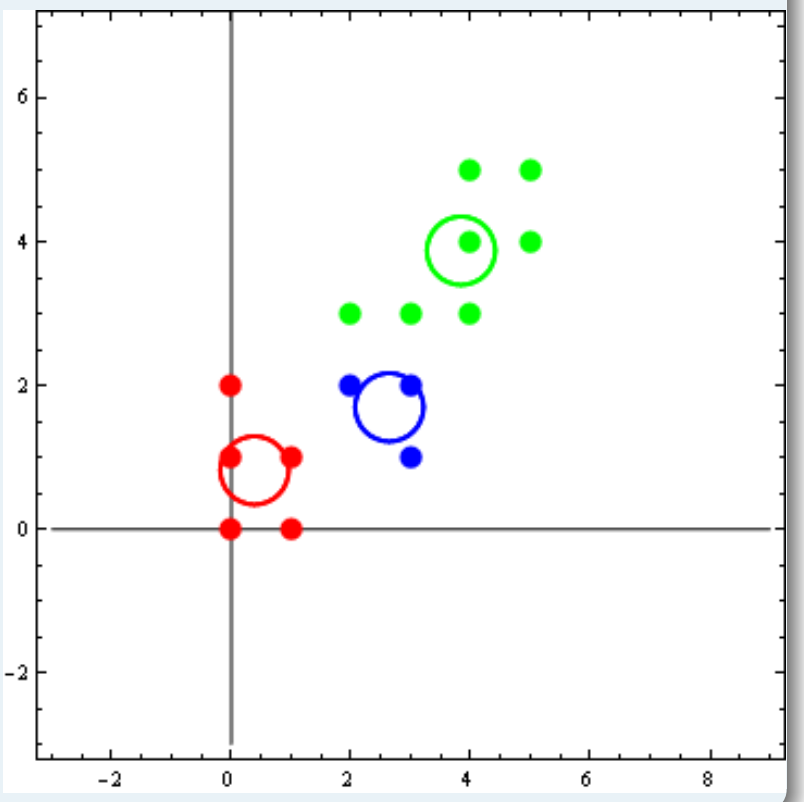
Przykład 1 - 3 klastry, inicjalizacja 1 (kroki 11-12)



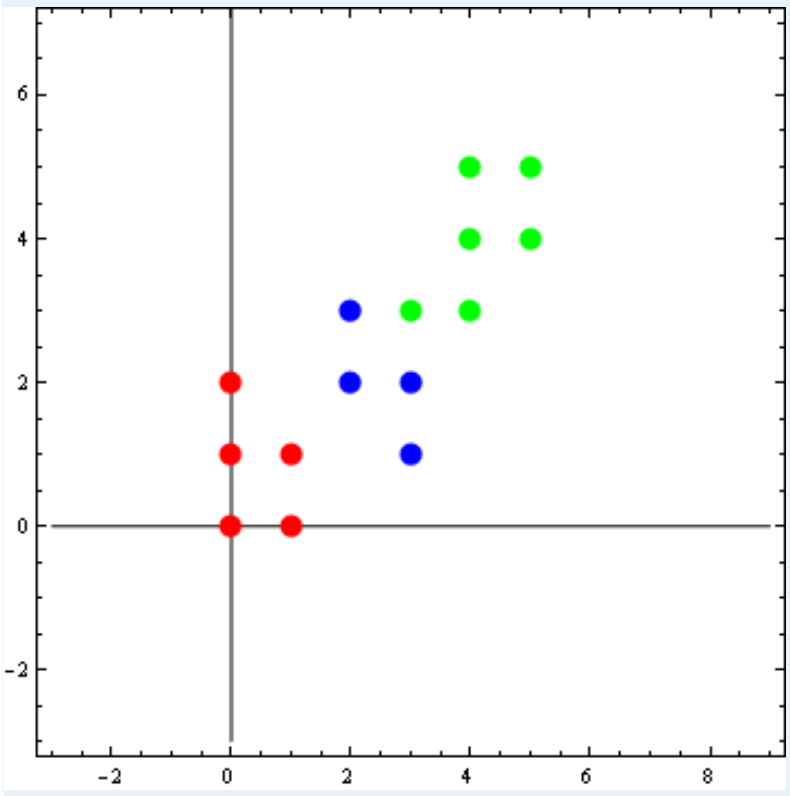


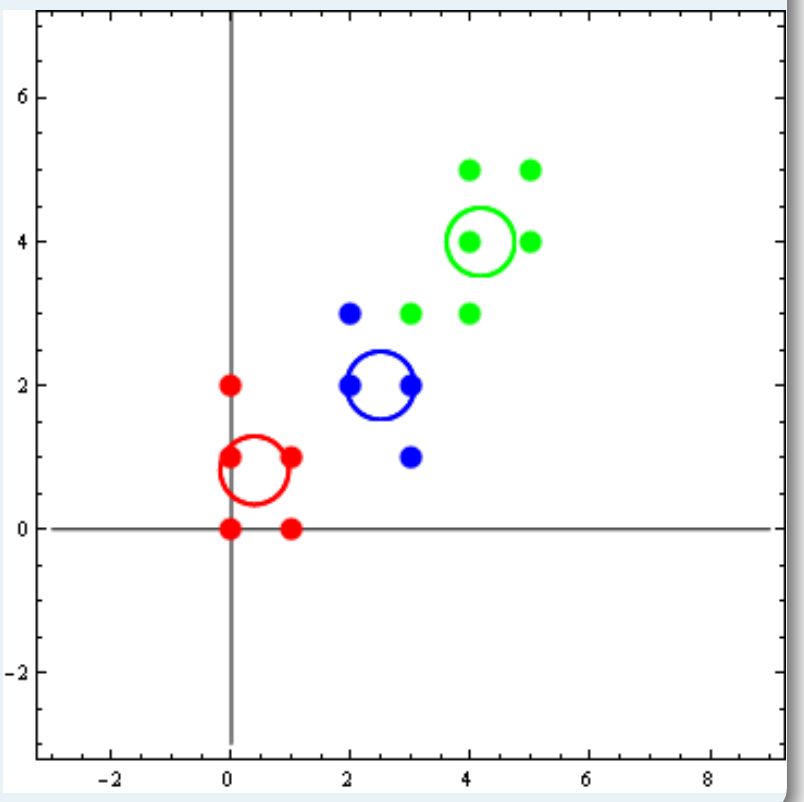
Przykład 1 - 3 klastry, inicjalizacja 1 (kroki 13-14)



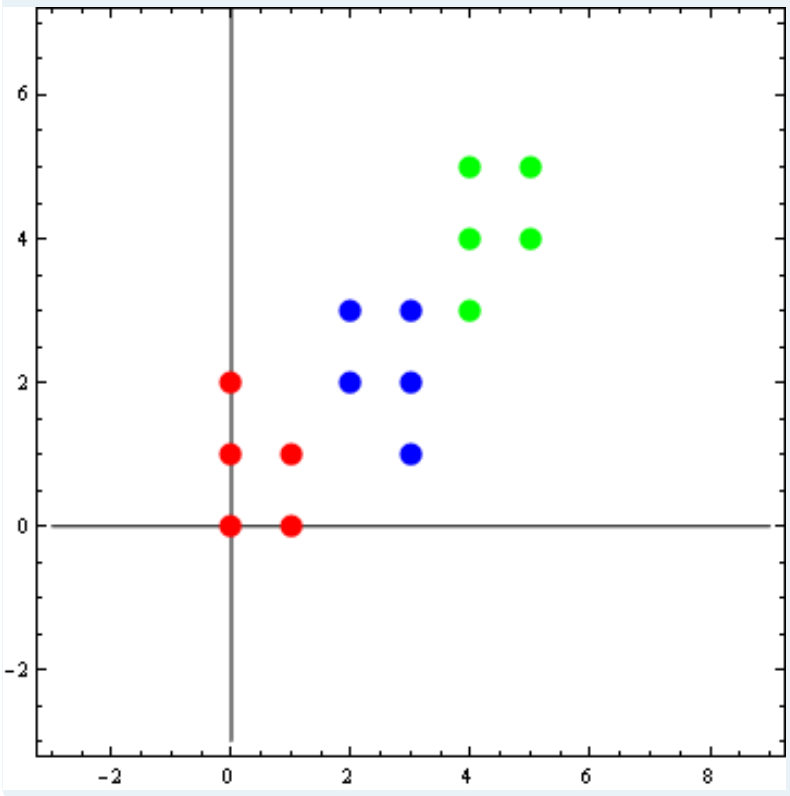


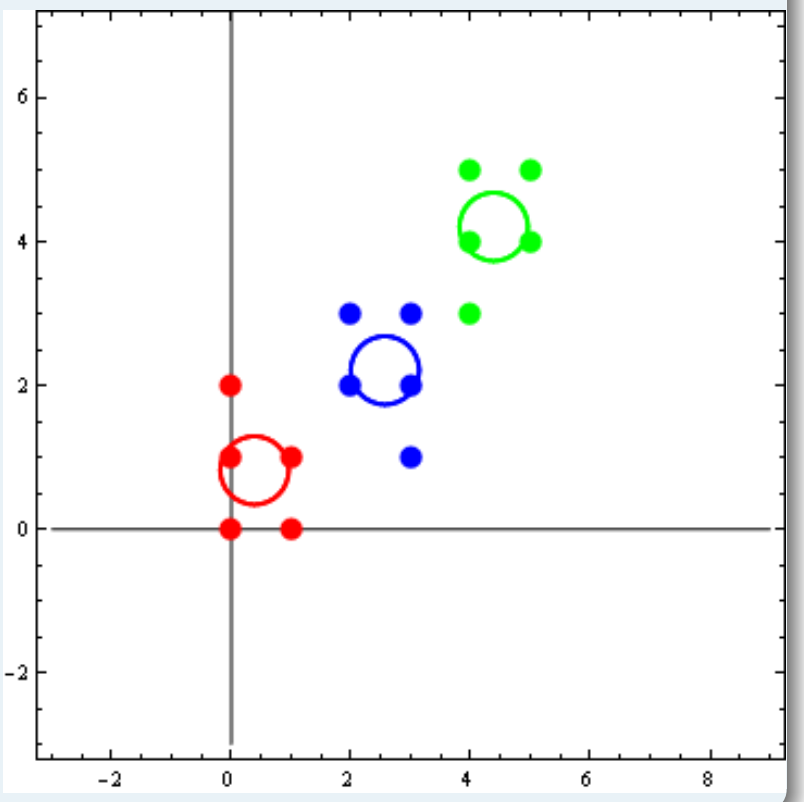
Przykład 1 - 3 klastry, inicjalizacja 1 (kroki 15-16)



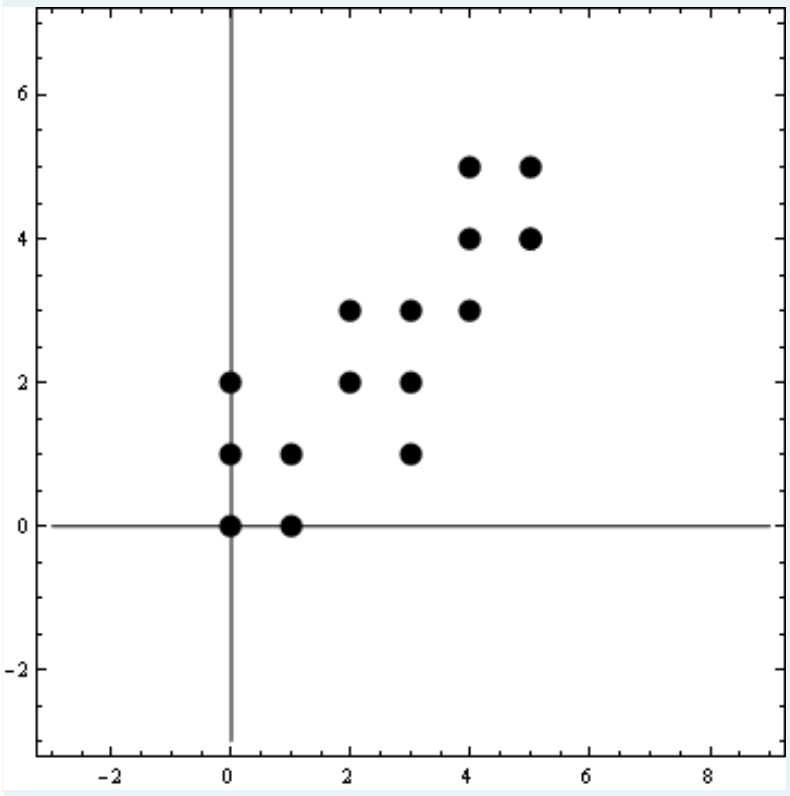


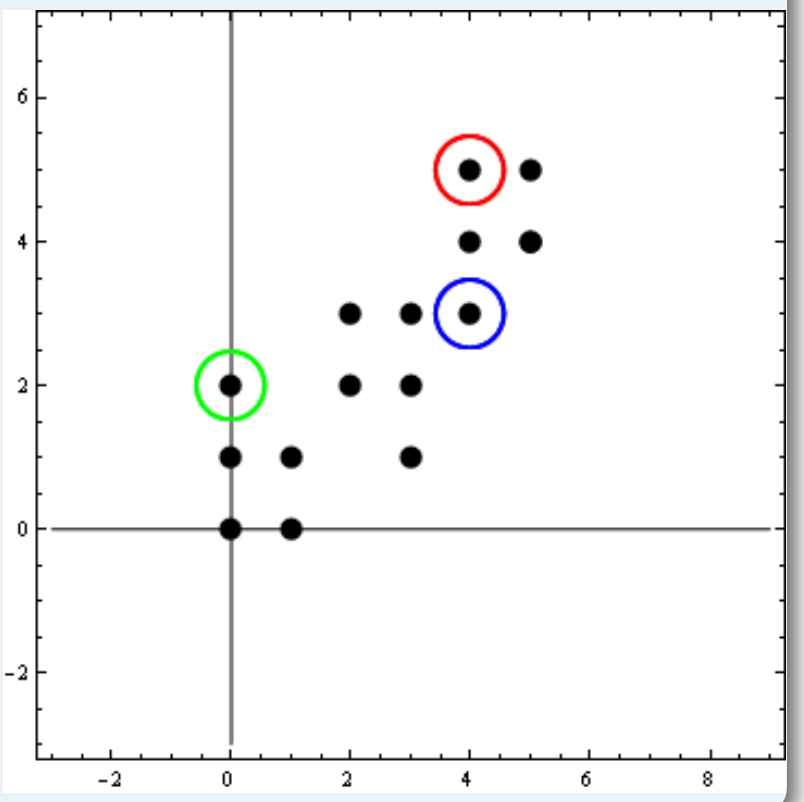
Przykład 1 - 3 klastry, inicjalizacja 1 (kroki 17-18)



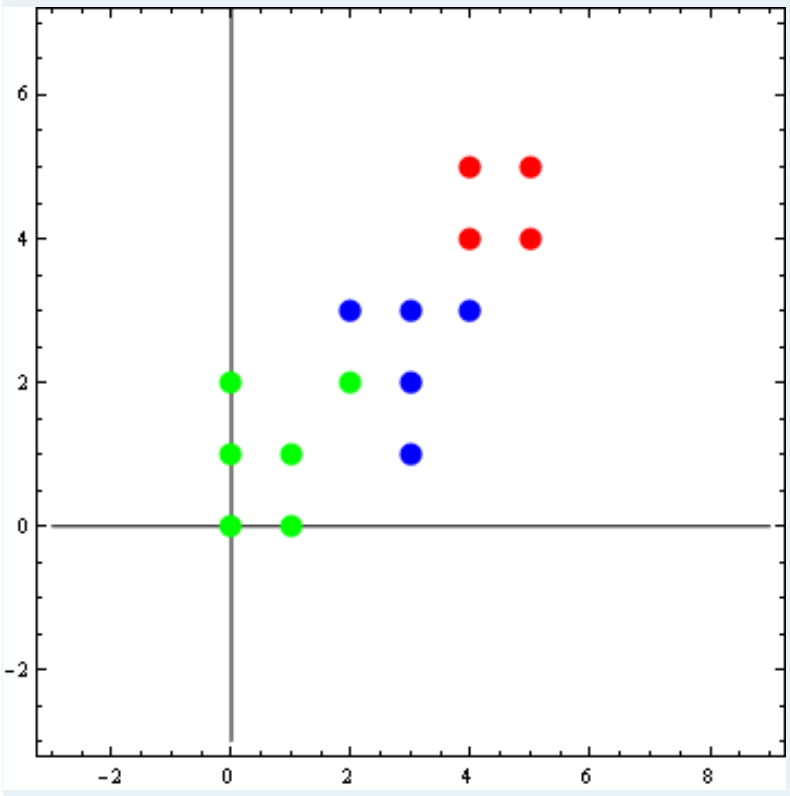


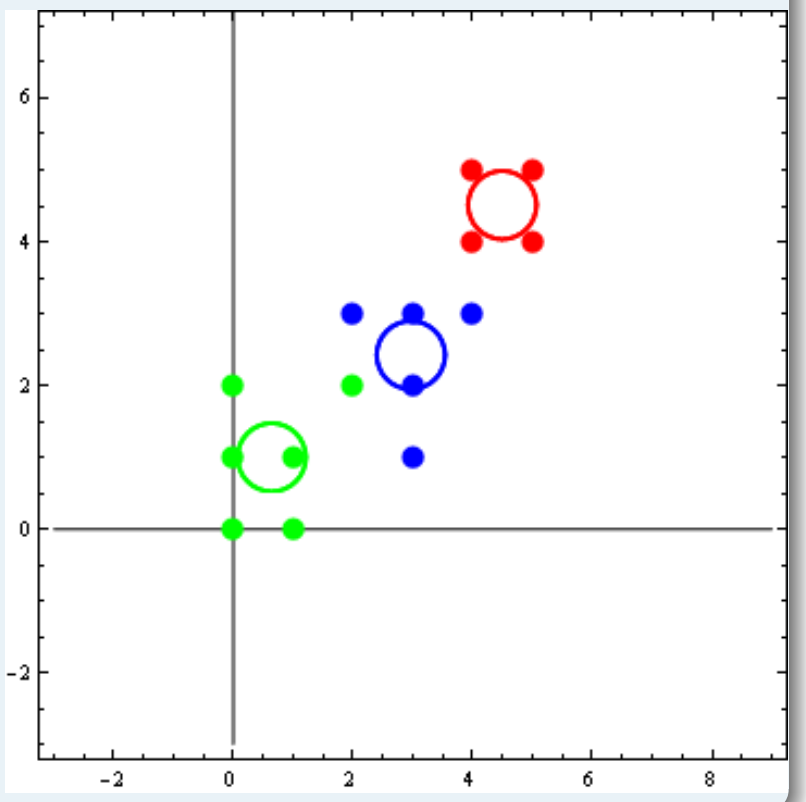
Przykład 2 - 3 klastry, inicjalizacja 2 (kroki 1-2)



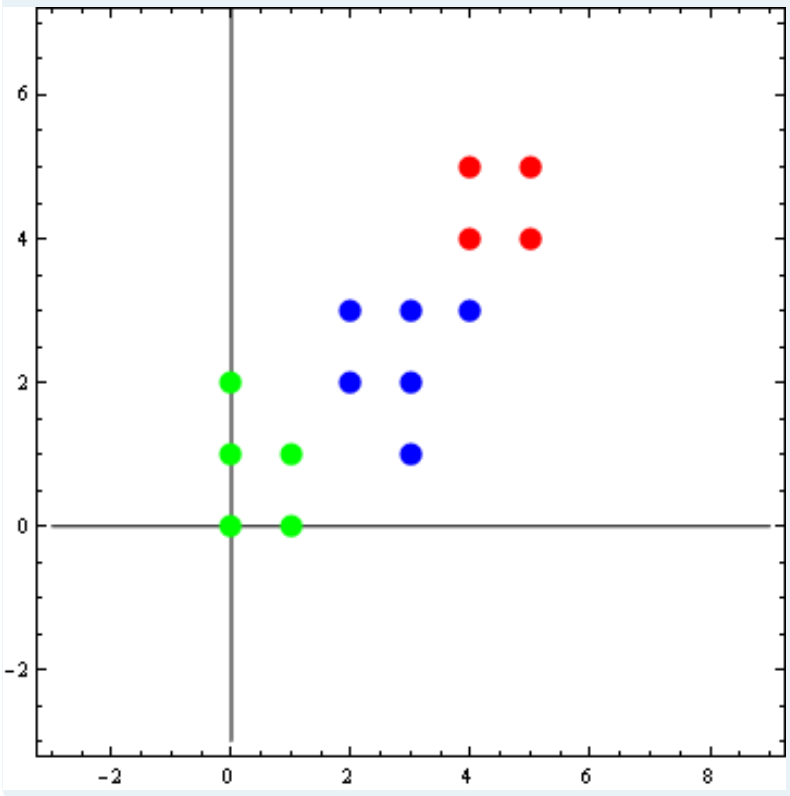


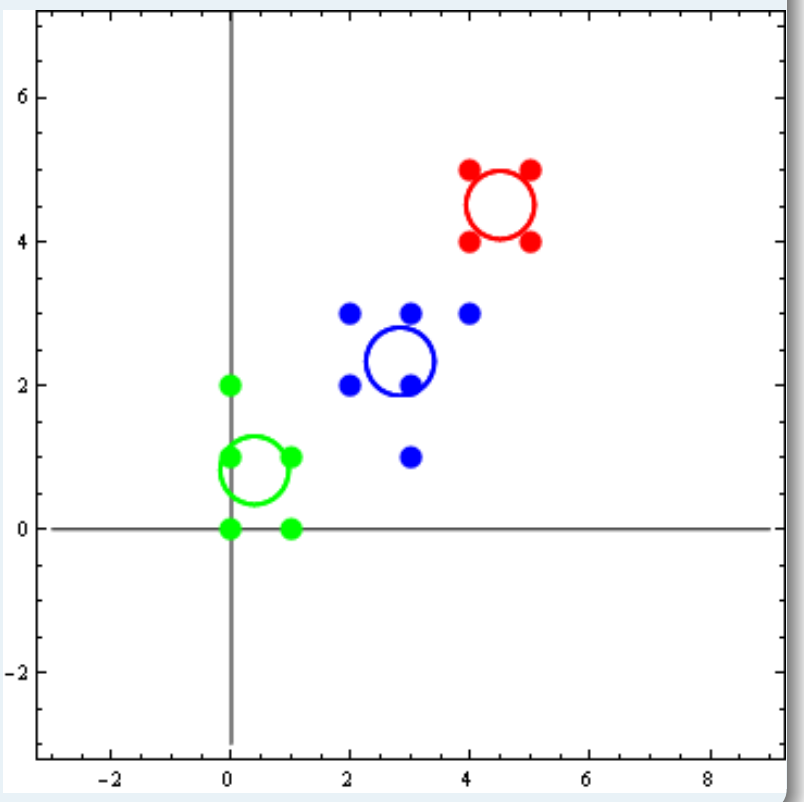
Przykład 2 - 3 klastry, inicjalizacja 2 (kroki 3-4)



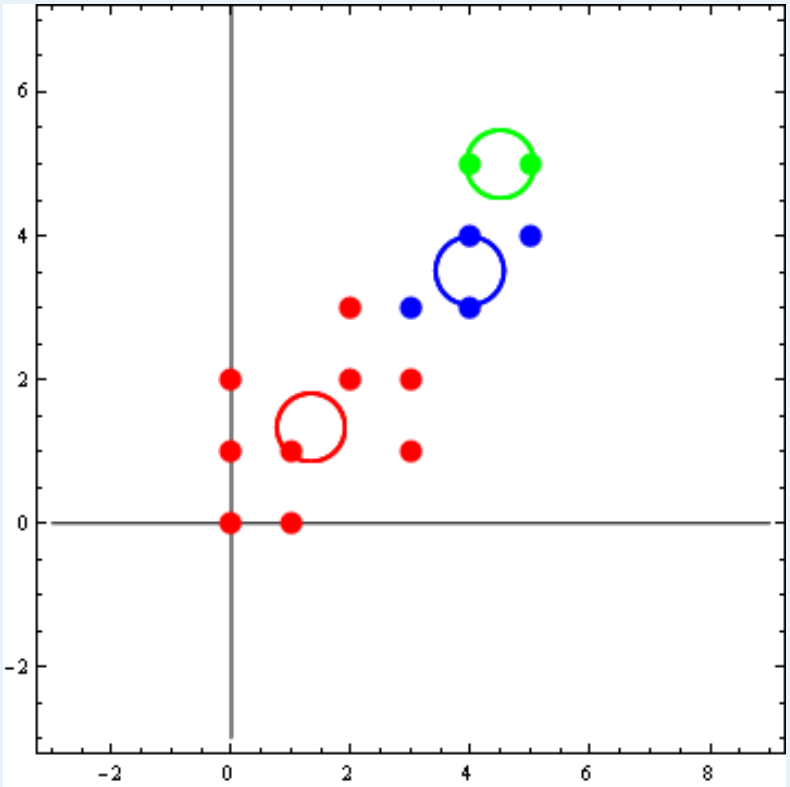


Przykład 2 - 3 klastry, inicjalizacja 2 (kroki 5-6)

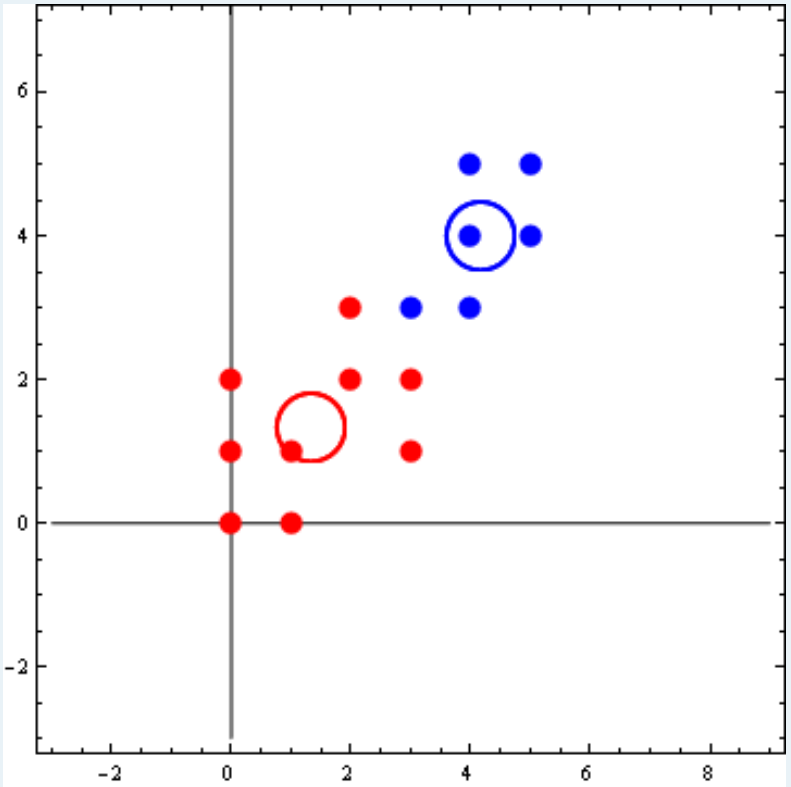




Przykład 3 - 3 klastry

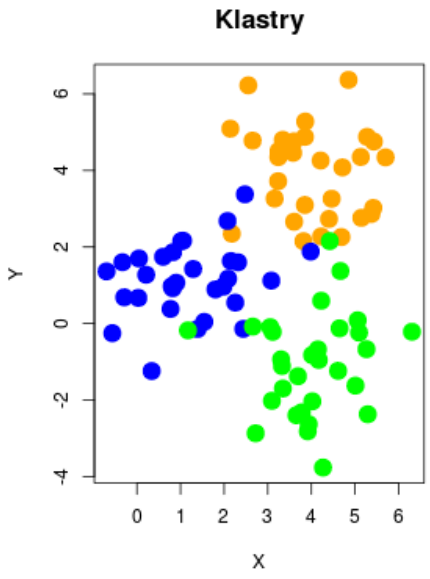


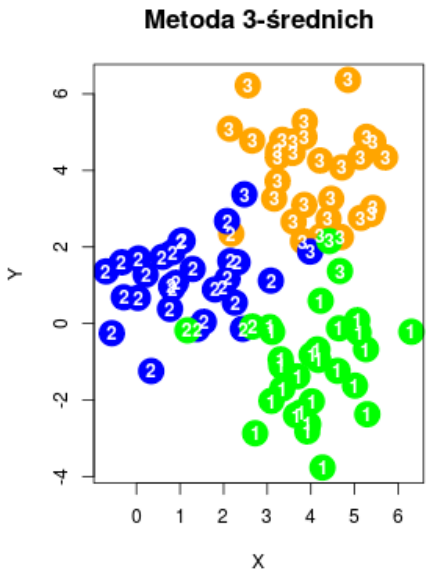
Przykład 4 - 2 klastry



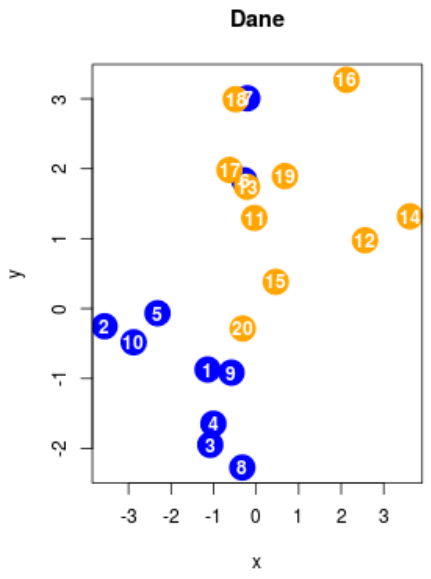
- opierają się na pomiarze uogólnionej odmienności między dwoma dowolnymi zbiorami obserwacji,
- nie wymagają z góry określenia liczby skupień,
- w pierwszym kroku **metody aglomeracyjnej** tworzymy tyle skupień, ile jest obserwacji,
- w następnym kroku w jedno skupienie łączona jest para najmniej odległych obserwacji,
- z kroku na krok skupień jest coraz mniej, aż w ostatnim powstaje cała próba w jednym skupieniu,
- w efekcie otrzymuje się nieskierowane drzewo (dendrogram),
- to samo można otrzymać "idąc od góry"(**metoda oparta na dzieleniu**), ale jest to dużo bardziej złożony obliczeniowo.

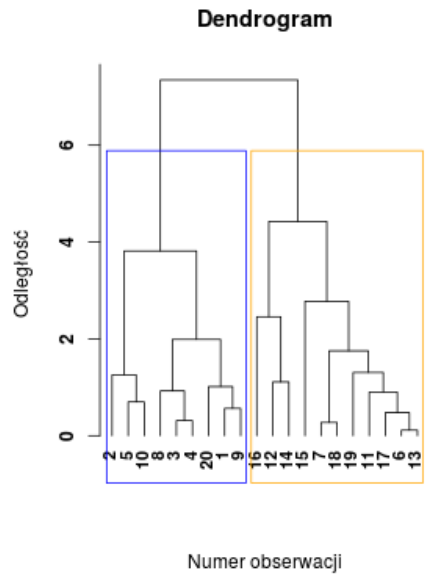
Przykłady



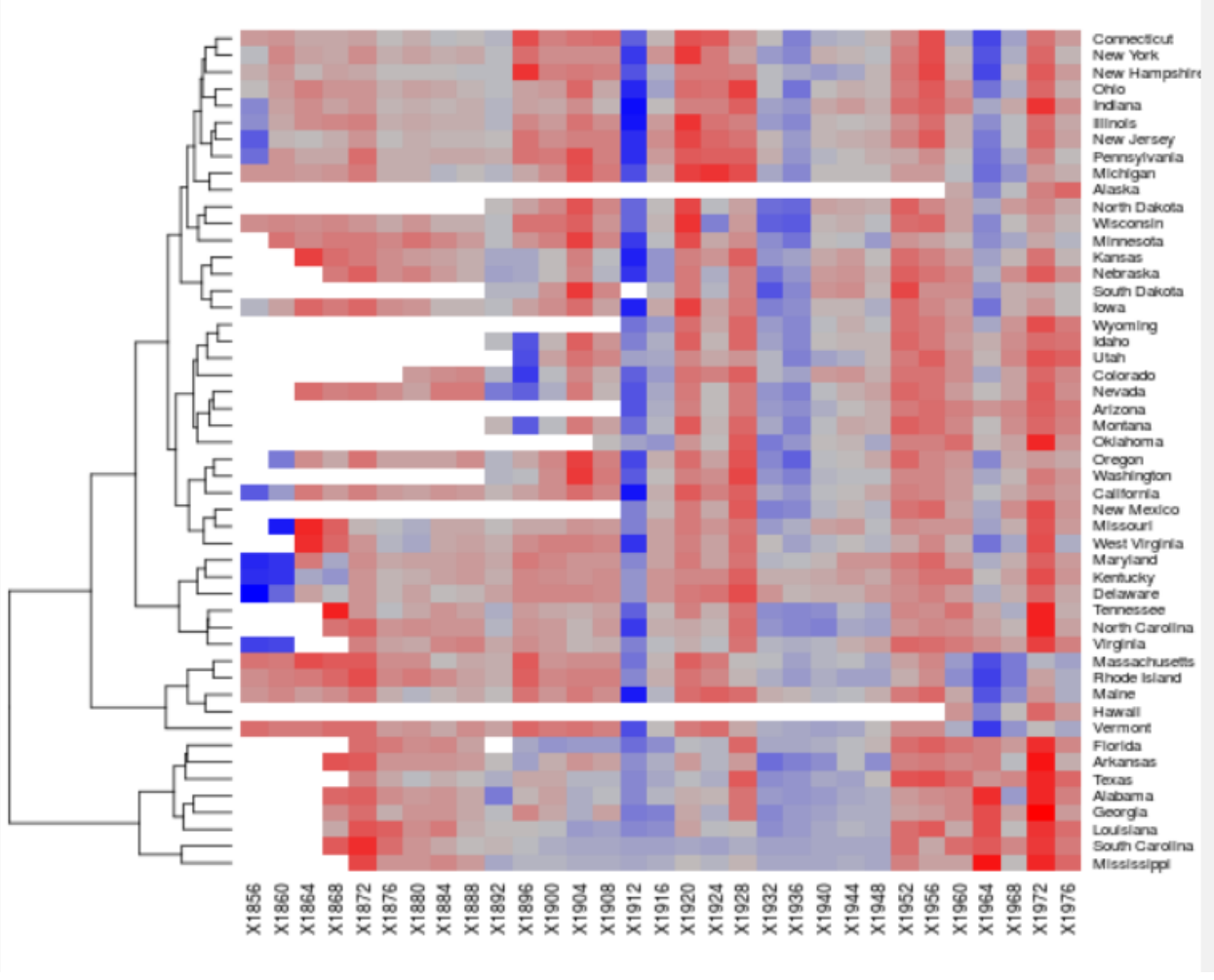


Metoda k-średnich





Metoda hierarchiczna

Przykłady

Metodę można wykorzystać także do bardziej subtelnych danych: poniżej zestawienie głosów oddanych na kandydatów Republikanów w wyborach prezydenckich USA dla różnych lat i różnych stanów.