

Analiza skupień

ML by ML

Mateusz Lango

7 czerwca 2019

Czym jest grupowanie?


Clustering is the task of grouping a set of objects such that similar objects end up in the same group and dissimilar objects are separated into different groups. Clearly, this description is quite *imprecise* and possibly *ambiguous*. Quite surprisingly, it is not at all clear how to come up with a more rigorous definition^a.

^aUnderstanding Machine Learning: From Theory to Practice

- Przechodniość grupowania i nieprzechodniość podobieństwa
- Brak złotego standardu
- Grupuj recenzje: po temacie? po wydźwięku? po autorze?
- Patrz materiały dodatkowe: "Impossibility Theorem" (odporność na skalowanie, bogactwo reprezentacji, zgodność)

Metody połączeniowe: AHC

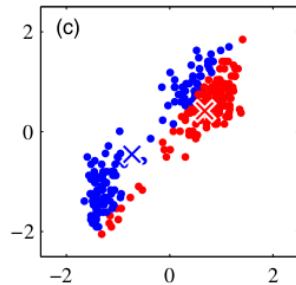
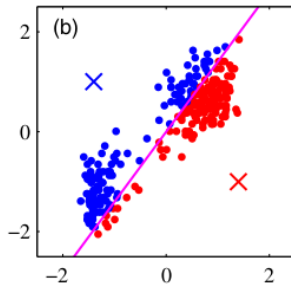
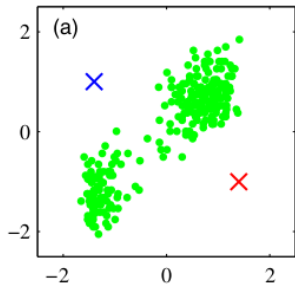
- Iteracyjnie łączymy najbliższe pary przykładów
- Ale co zrobić jak przykład jest już w grupie? Jak policzyć odległość pomiędzy przykładem a grupą? (Albo dwiema grupami?)
 - Single-link: odległość między najbliższymi elementami w skupieniach
 - Complete-link: odległość między najdalszymi elementami w skupieniach
 - Average-link: średnia odległości wszystkich par między skupieniami
 - Ward: połączenie, które minimalizuje całkowitą wariancję skupienia¹
- Dendogram
- Warunek stopu?
- Złożoność?

¹suma wariancji na poszczególnych cechach, ślad macierzy kowariancji skupienia 

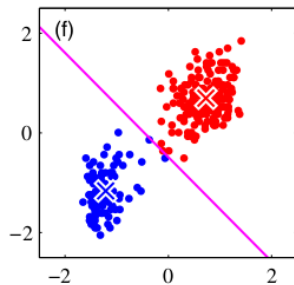
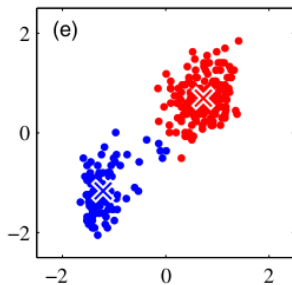
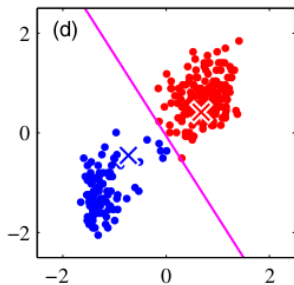
Metody optymalizacyjne: K-średnich

- Inicjalizuj k centroidów
- Przypisz przykłady do najbliższego centroidu (każdy centroid tworzy grupę)
- Policz nowy centroid (średnia arytmetyczna) każdej grupy
- Wróć do punktu 2, aż do zbieżności/wyczerpania budżetu czasowego/...

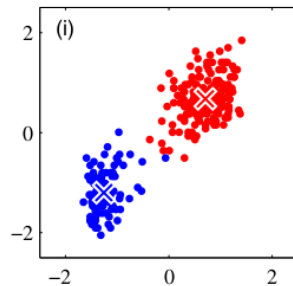
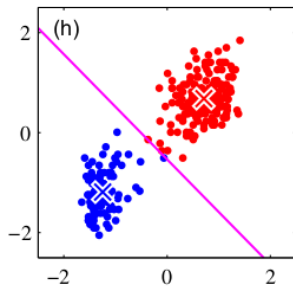
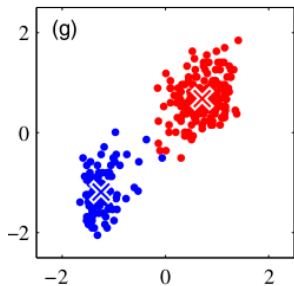
Przykład działania



Przykład działania



Przykład działania



K-średnich

- Dla każdego skupienia odległość przykładów do centroidu to

$$\sum_{i=1}^n \|x_i - c\|^2$$

- Średni błąd kwantyzacji/ suma kwadratów

$$\frac{1}{K} \sum_{i=1}^K \sum_{j=1}^{n_i} \|x_{i,j} - c_i\|^2$$

gdzie $x_{i,j}$ to j -ty element w i -tej grupie o centroidzie c_i

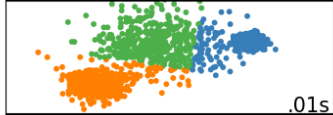
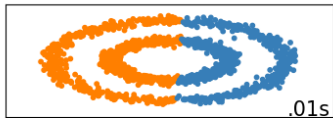
- Złożoność?
- Inicjalizacja?
- Przy złej inicjalizacji algorytm może nie osiągać nawet minimum lokalnego!
- Wybór k ?

Metody gęstościowe: DBSCAN

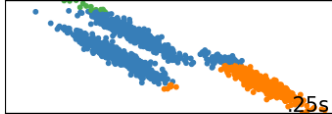
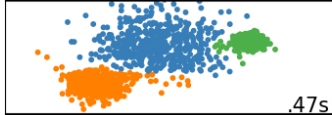
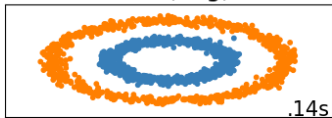
- Parametry: ϵ - maksymalna „bliska” odległość oraz k - minimalna liczba punktów
- Przykład rdzenny: ma co najmniej k przykładów (w tym siebie samego) w odległości maksymalnie ϵ
- Przykłady rdzenne mają zdolność do formowania skupień, które zawierają inne przykłady rdzenne oraz nierdzenne które są osiągalne² z przykładów rdzennych
- Przykłady które nie należą do żadnego skupienia to obserwacje samotnicze
- Wady? Zalety?
- $k = 1$ - każdy jest rdzenny i tworzy grupę, $k = 2$ jedno z rozwiązań AHC single linkage, heurystyka $k = 2d$ (liczba cech).
- *epsilon* - ...

²Są w odległości co najwyżej ϵ

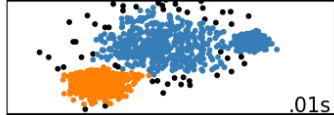
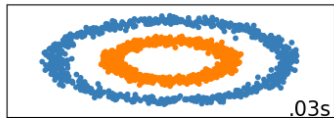
KMeans



AHC (avg)



DBSCAN

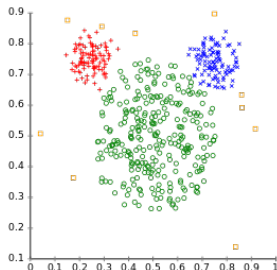


Podsumowanie

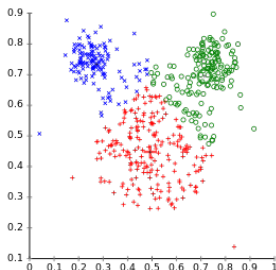
- grupowanie jako problem źle zdefiniowany
- 3 rodzaje metod (ale jest ich więcej)
- kluczowy jest wybór funkcji odległości
- klątwa wymiarowości...

Different cluster analysis results on "mouse" data set:

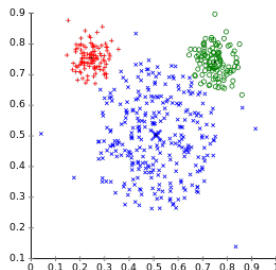
Original Data



k-Means Clustering



EM Clustering



Widzimy się za tydzień!