

Nowoczesne metody analizy skupień

Maciej Nasiński, Paweł Strawiński

Uniwersytet Warszawski

12 stycznia 2023

- 1 Wprowadzenie
- 2 GMM
- 3 DBSCAN
- 4 Spectral Clustering
- 5 Podsumowanie

Metody analizy skupień

- Rozkładu (GMM)
- Centroidowe (k-średnich)
- Hierarchiczne
- Gęstości (DBSCAN)
- Teorii grafów (CLICK, Spectral)
- Fraktalowe
- ...

Szeroki wachlarz metod analizy skupień

- różnorodne cechy danych takie jak liczba wymiarów, rozkład cech, skorelowane cechy, braki danych
- ograniczone zasoby obliczeniowe
- precyzyjny cel lub założenia analizy

Złożność obliczeniowa (czasowa)

Wiele z algorytmów skupień działa poprzez obliczenie podobieństwa między wszystkimi parami obserwacji, czas wykonania zwiększa się z kwadratem ich liczby **$O(n^2)$** . Gdzie algorytm k-średnich skaluje się liniowo z liczbą obserwacji **$O(n)$** .

- k-średnich - $O(n * k * i * d) \rightarrow O(n)$
- DBSCAN - średnio $O(n \log n)$ z górnym ograniczeniem $O(n^2)$
- GMM - $O(n * k * d^{(2 \text{ lub } 3)}) \rightarrow O(n)$
- Spectral - $O(n^3)$

Etykiety Miekkie vs Twarde

- Twarde (Hard Labels) - przypisanie do jednego skupienia (DBSCAN, k-średnich)
- Miekkie (Soft Labels) - przypisanie do wielu skupień (GMM)

Mixture Model

Mixture Model to model probabilistyczny, w którym zakłada się, że wszystkie punkty danych są generowane z mieszaniny skończonej liczby rozkładów o nieznanych parametrach.

$$f(x) = \sum_{k=1}^K \alpha_k f_k(x)$$

α_k reprezentuje wagę k -tego składnika/rozkładu, gdzie $\sum_{k=1}^K \alpha_k = 1$. Składowe $f_k(x)$ to dowolny rozkład.

Gaussian Mixture Model

W praktyce często stosuje się rozkłady parametryczne (np. gaussa). Jeśli zastąpisz każde $f_k(x)$ rozkładem gaussowskim, otrzymasz tak zwany GMM (Gaussian Mixture Model).

$$f(x) = \sum_{k=1}^K \alpha_k f_k(x; \theta_k)$$

Podobnie, jeśli dla $f_k(x)$ użyje się rozkładu dwumianowego, otrzymasz BMM (Binomial Mixture Model).

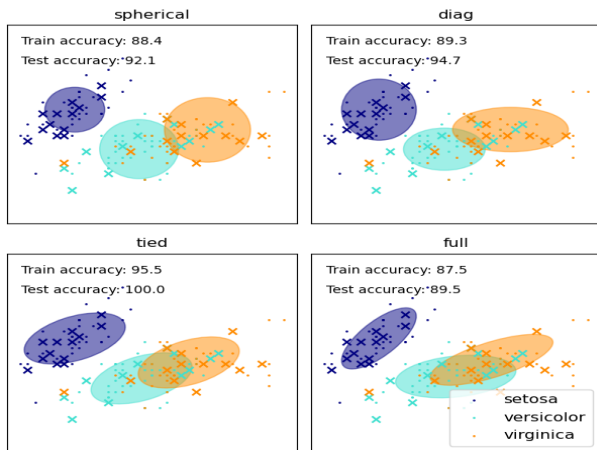
Rozkład normalny jednowymiarowy vs wielowymiarowy

$$N(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x - \mu)^2}{2\sigma^2}}$$

$$N(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right)$$

Dla przypadku wielowymiarowego konieczne jest policzenie macierzy kowariancji Σ .

GMM - algorytm - Macierz Kowariancji



GMM - algorytm EM

- 1 Wybierz liczbę skupień K .
- 2 Wybierz początkowe wartości μ_k i Σ_k dla każdego składnika. (k-średnich) “Hard Labels (etykiety twarde)”.
- 3 Wykonaj **krok-E** (przypisanie punktów do skupień). “Soft Labels (etykiety miękkie)”.
- 4 Wykonaj **krok-M** (dopasowanie parametrów).
- 5 Powtarzaj kroki 3 i 4, aż do osiągnięcia kryterium zatrzymania.

Bayes Theorem

Bayes Theorem:

$$P(e \cap h_i) = P(e|h_i)P(h_i) = P(h_i|e)P(e)$$

$$P(h_i|e) = \frac{P(e|h_i) * P(h_i)}{P(e)}$$

$$P(e) = \sum_{i=1}^N P(e|h_i)P(h_i)$$

Posterior: $P(h_i|e)$

Likelihood: $P(e|h_i)$ Proporcjonalne do Posterior

Prior: $P(h_i)$ informative or uninformative

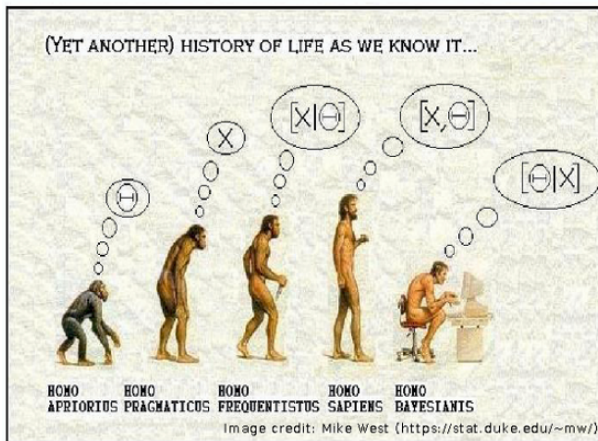
Twierdzenie o prawdopodobieństwie całkowitym

Twierdzenie o prawdopodobieństwie całkowitym:

$$P(e) = \sum P(e \cap h_i) = \sum_{i=1}^N P(e|h_i)P(h_i)$$

$$P(e) = \int_{-\infty}^{\infty} P(e|X = x)f_X(x)dx$$

Ewolucja



GMM - algorytm EM - E-step

Cel oszacować $P(x_i \in k_j | x_i)$ dla każdego punktu danych x_i i każdego składnika k_j .

$$P(x_i \in k_j | x_i) = \frac{P(x_i | x_i \in k_j) P(k_j)}{P(x_i)}$$

gdzie:

$$P(x_i | x_i \in k_j) = \mathcal{N}(x_i | \mu_{k_j}, \sigma_{k_j}^2)$$

$$P(k_j) = \alpha_{k_j}$$

$$P(x_i) = \sum_{k=1}^K \alpha_k \mathcal{N}(x_i | \mu_k, \sigma_k^2)$$

GMM - algorytm EM - M-step

Cel oszacować μ_{k_j} , $\sigma_{k_j}^2$ oraz α_{k_j} , wykorzystując $P(x_i \in k_j | x_i)$.

$$\mu_k = \frac{\sum_i^N P(x_i \in k_j | x_i) x_i}{\sum_i^N P(x_i \in k_j | x_i)}$$

$$\sigma_k^2 = \frac{\sum_i^N P(x_i \in k_j | x_i) (x_i - \mu_k)^2}{\sum_i^N P(x_i \in k_j | x_i)}$$

$$\alpha_k = \frac{\sum_i^N P(x_i \in k_j | x_i)}{N}$$

GMM - algorytm - Expectation Maximization

Kroki E-step (Expectation) oraz M-step (Maximization) są powtarzane aż do konwergencji.

Potrzebujemy zdefiniować funkcje celu / kosztu aby wiedzieć jak odnaleźć najlepsze rozwiązanie oraz kiedy zatrzymać algorytm.

$$P(X|\mu, \sigma, \alpha) = \sum_{k=1}^K \alpha_k \varphi(X|\mu_k, \sigma_k^2)$$

$$\ln P(X|\mu, \sigma, \alpha) = \sum_{n=1}^N \ln \sum_{k=1}^K \alpha_k \varphi(x_n|\mu_k, \sigma_k^2)$$

Większy log-likelihood oznacza lepsze dopasowanie parametrów w modelu.

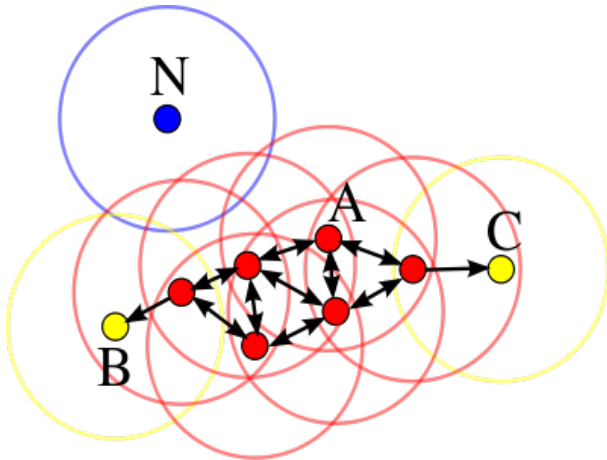
DBSCAN

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to algorytm grupowania oparty na gęstości. Grupuje punkty, które ściśle sasiadują (punkty z wieloma pobliskimi sąsiadami).
- DBSCAN składa się z dwóch parametrów: ϵ i *MinPts*.
- DBSCAN może być używany do znajdowania skupień o dowolnym kształcie, w przeciwieństwie do k-średnich, które zakładają, że skupienia mają kształt wypukły.

DBSCAN - algorytm

- 1 Znajdź wszystkie punkty w odległości ε od każdego punktu.
- 2 Jeśli punkt ma co najmniej *MinPts* punktów w odległości ε , jest to punkt główny.
- 3 Jeśli punkt jest punktem centralnym, wszystkie punkty w odległości ε od niego są częścią tego samego skupienia.
- 4 Jeśli punkt nie jest punktem centralnym, ale znajduje się w odległości ε od centralnego, jest to punkt graniczny.
- 5 Wszystkie inne punkty to szum.

DBSCAN - algorytm - wizualizacja



Spectral Clustering

Spectral clustering to algorytm grupowania oparty na wektorach własnych macierzy Laplace'a grafu (macierzy podobieństwa).

Algorytm opiera się na następujących krokach:

- Skonstruować graf z punktów danych (macierz podobieństwa)
- Oblicz macierz Laplace'a grafu
- Oblicz wektory własne macierzy Laplace'a
- Grupuj punkty danych na podstawie wektorów własnych
- Przypisz punkty danych do skupień na podstawie wektorów własnych (k-średnich)

Podsumowanie

- Pośród tak wielu metod analizy skupień kluczowy jest właściwy dobór metody oraz jej parametrów.
- Algorytmy analizy skupień są bardzo podatne na dane wejściowe.

Dodatkowe Źródła

- 1 Xu, D., Tian, Y. A Comprehensive Survey of Clustering Algorithms. Ann. Data. Sci. 2, 165–193 (2015).
<https://doi.org/10.1007/s40745-015-0040-1>
- 2 Allen B. Downey 2012, Think Bayes - Bayesian Statistics Made Simple, O'Reilly Media, Inc.,
<http://greenteapress.com/thinkbayes/>
- 3 URL: <https://tinyheero.github.io/2016/01/03/gmm-em.html>