Nowoczesne metody analizy skupień

Paweł Strawiński

Zajęcia 11
18 stycznia 2024

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

Złożoność obliczeniowa (czasowa)

Wiele algorytmów analizy skupień działa poprzez obliczenie podobieństwa między wszystkimi parami obserwacji, czas wykonania zwiększa się z kwadratem ich liczby **O(nˆ2)**. Gdzie algorytm k-średnich skaluje się liniowo z liczbą obserwacji **O(n)**.

- k-średnich - O(n * k * i * d) -> O(n)
- DBSCAN - średnio O(nlogn) z górnym ograniczeniem O(nˆ2)
- GMM - O(n * k * d ˆ(2 lub 3)) -> O(n)
- Spectral - O(nˆ3)

Etykiety Miękkie a Twarde

- Twarde (Hard Labels) - przypisanie do jednego skupienia (DBSCAN, k-średnich)
- Miękkie (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 wykorzystywane są rozkłady parametryczne (np. gaussa). Po zastąpieniu każdej ($f_k(x)$ rozkładem normalnym, otrzymuje się tzw.mieszaninę rozkładów gausowskich GMM (ang. *Gaussian Mixture Model*).

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

Podobnie, jeżeli dla $f_k(x)$ przyjmie się rozkład dwumianowy, powstanie Dwumianowa mieszanina rozkładów BMM (ang. *Binomial Mixture Model*).

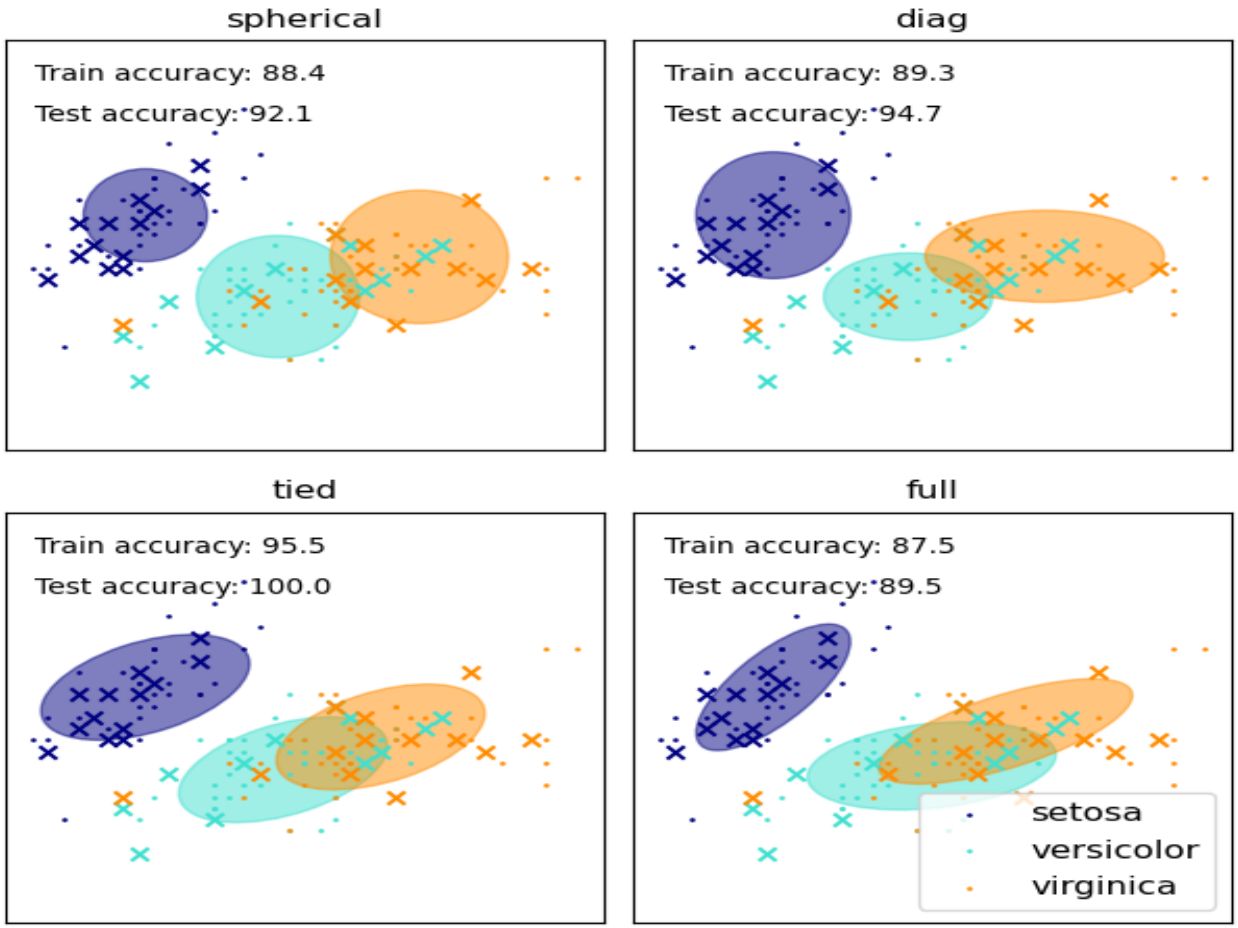
Rozkład normalny jednowymiarowy vs wielowymiarowy

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

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

Dla przypadku wielowymiarowego konieczne jest policzenie macierzy kowariancji Σ.

GMM: 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 stopu.

Twierdzenie Bayesa

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

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

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

Rozkład a-posteriori: $P(h_i|e)$

Funkcja wiarogodności: $P(e|h_i)$ proporcjonalna do rozkładu a=posteriori

Rozkład a-priori: $P(h_i)$ informacyjny lub nieinformacyjny

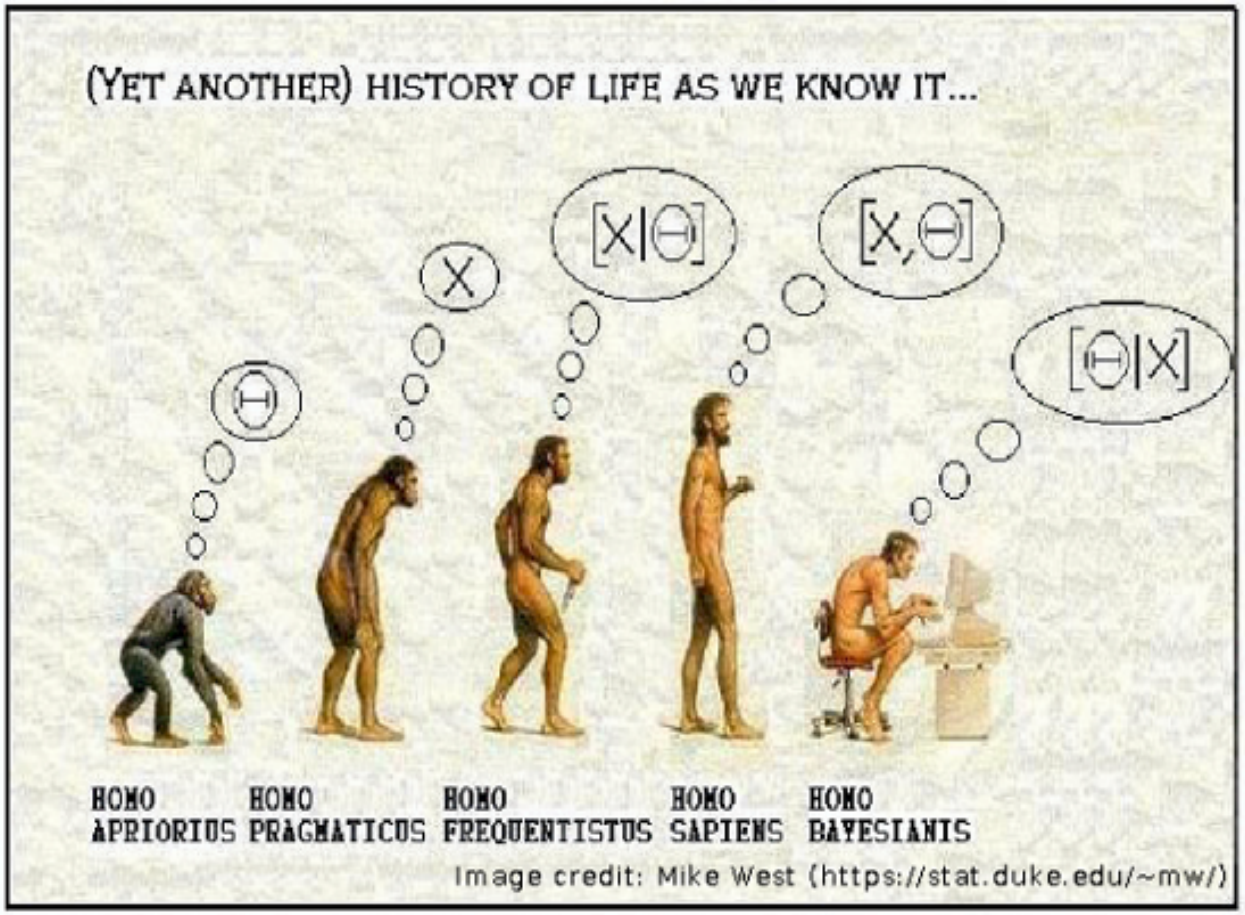
Twierdzenie o prawdopodobieństwie całkowitym

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


(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...
[Θ]
[X]
[X|Θ]
[X,Θ]
[Θ|X]
HOMO APRIORIUS
HOMO PRAGMATICUS
HOMO FREQUENTISTUS
HOMO SAPIENS
HOMO BAYESIANIS
Image credit: Mike West (https://stat.duke.edu/~mw/)

GMM - algorytm EM - krok E

Cel oszacowanie ($P(x_i \in k_j|x_i)$ 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:

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

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

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

GMM - algorytm EM - krok M

Cel oszacować μ_{k_j}, $\sigma^2_{k_j}$ 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 EM

- Kroki E (Expectation) oraz M (Maximization) powtarzane są do uzyskania zbieżności.
- Definiowana jest funkcje celu / kosztu by odnaleźć najlepsze rozwiązanie oraz moment zatrzymania algorytmu

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

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

- Im większa wartość funkcji wiarogodności tym lepsze dopasowanie parametrów w modelu.

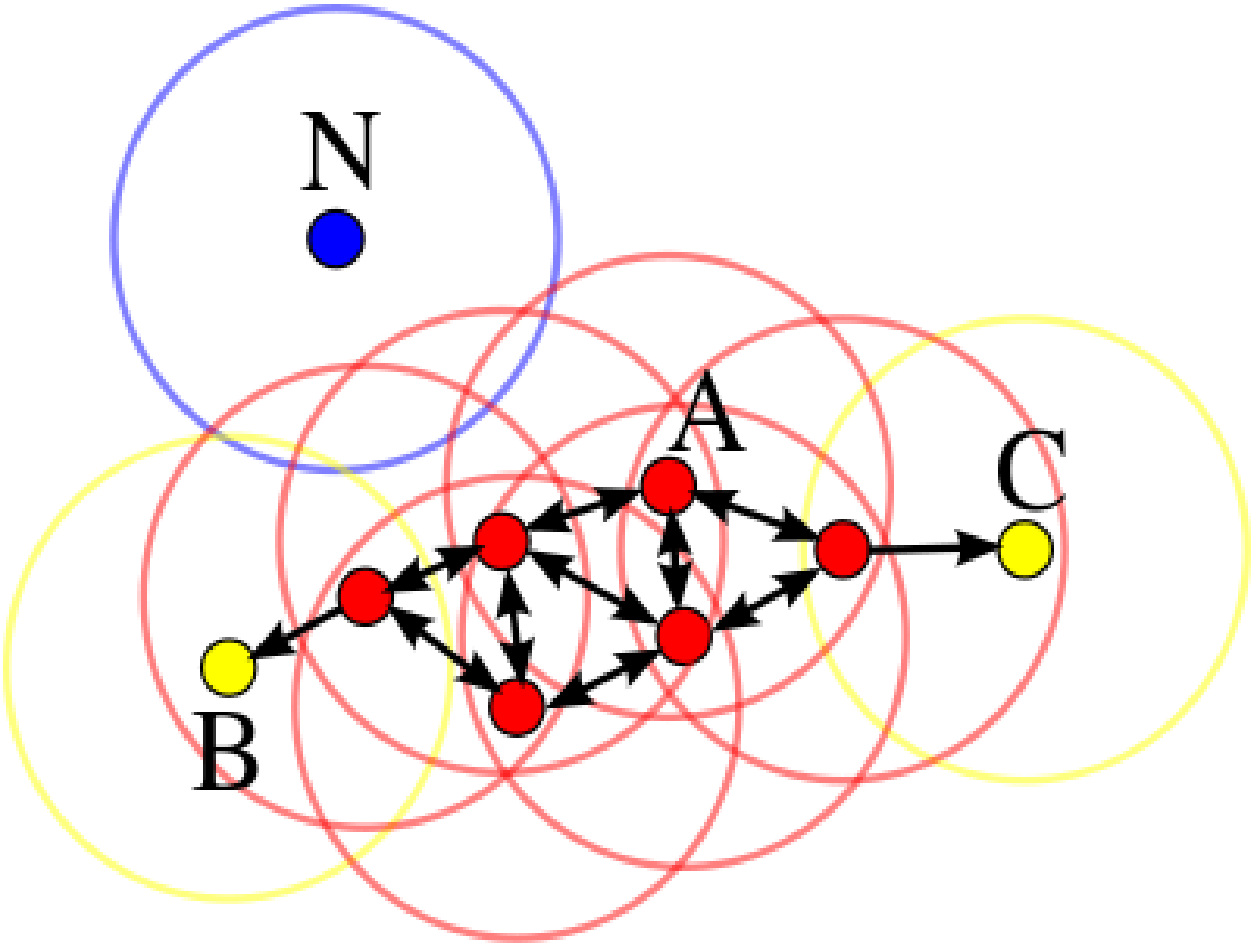
DBSCAN

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to algorytm grupowania oparty na gęstości danych. Grupuje punkty, które ściśle sąsiadują (punkty z wieloma sąsiadami w otoczeniu).
- DBSCAN posiada parametry: odległość ε i liczba punktów *MinPts*.
- DBSCAN może być używany do znajdowania skupień o dowolnym kształcie, w przeciwieństwie do k-średnich, która zakłada, że skupienia mają kształt wypukły.

DBSCAN - algorytm

1. Znajdź wszystkie punkty w odległości ε od każdego punktu.
2. Jeżeli punkt ma co najmniej *MinPts* punktów w odległości conajwyżej ε, jest to punkt główny.
3. Jeżeli punkt jest punktem centralnym, wszystkie punkty w odległości conajwyżej ε od niego są częścią tego samego skupienia.
4. Jeżeli 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 Laplaca grafu (macierzy podobieństwa). Algorytm obejmuje następujące kroki:

- Skonstruowanie grafu z punktów danych (macierz podobieństwa)
- Obliczenie macierz Laplace'a grafu
- Obliczenie wektorów własnych macierzy Laplace'a
- Grupowanie punktów danych na podstawie wektorów własnych
- Przypisanie punkty danych do skupień na podstawie wektorów własnych (k-średnich)

Podsumowanie

- Pośród wielu metod analizy skupień kluczowy jest właściwy dobór metody oraz jej parametrów.
- Wyniki działania algorytmów analizy skupień zależą od własności danych wejściowych.

Bibliografia

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