Analiza Wielowymiarowa

Skalowanie wielowymiarowe i obserwacje nietypowe

Paweł Strawiński

Zajęcia 5
23 listopada 2023

Plan zajęć I

Wprowadzenie 1/2

- Skalowanie wielowymiarowe (ang. *Multidimensional Scaling*) jest to technika statystyczna wykorzystywana do redukcji wymiarowości danych oraz ich wizualizacji
- Jest to zestaw technik porządkowania danych (ang. *ordination techniques*) stosowanych w wizualizacji informacji, w szczególności do wyświetlania informacji zawartych w macierzy odległości
- Miary niepodobieństwa między obserwacjami w przestrzeni wielowymiarowej są reprezentowane w przestrzeni niskowymiarowej (zwykle w dwuwymiarowej), w taki sposób, że miary odległości w przestrzeni niskowymiarowej są zbliżone do miar odległości w przestrzeni wielowymiarowej

Wprowadzenie 2/2

- W praktyce nie oblicza się miar (nie)podobieństwa między obiektami a definiuje je poprzez zbiór cech obiektów (zmiennych)
- Najczęściej wykorzystywana jest odległość L2 (euklidesowa) lub L1 (miejska)
- Zazwyczaj liczba obserwacji przekracza liczbę zmiennych, ale nie jest to warunek konieczny dla przeprowadzenia skalowania wielowymiarowego

Klasyczne skalowanie wielowymiarowe

- Jest także określane jako *Principal Coordinates Analysis (PCoA)*
- Jeśli odległości są odległościami euklidesowymi, klasyczny MDS daje łatwe rozwiązanie algebraiczne
- Na podstawie macierzy niepodobieństwa obliczana jest wartość funkcji kryterium nazywanej *strain*

$$Strain(x_1, \ldots, x_n) = \left(\frac{\sum_{ij} (b_{ij} - x_i' x_j)^2}{\sum_{ij} b_{ij}^2} \right)^{1/2}$$

- Elementy b_{ij} są wyliczane wg algorytmu

Algorytm

- Klasyczne skalowanie wielowymiarowe zakłada odległość Euklidesową L2.
- Klasyczne skalowanie wielowymiarowe wykorzystuje fakt, że macierz X można uzyskać poprzez przekształcenie macierzy $B = XX'$ na macierz wartości własnych
- Macierz B jest obliczana z macierzy podobieństwa poprzez podwójne centrowanie
  1. Wyznacz macierz odległości $D = [d_{ij}^2]$
  2. Zastosuj podwójne centrowanie $B = -\frac{1}{2}CDC$, gdzie $C = I - \frac{1}{n}Jn$, gdzie n to liczba obserwacji, J oznacza macierz 1.
  3. Wyznacz m jako największą wartość własną macierzy B
  4. Nowe $X = E_m \Lambda_m^{1/2}$. E_m to macierz m wektorów własnych, Λ_m to diagonalna macierz m wartości własnych macierzy B

Metryczne skalowanie wielowymiarowe

- Jest uogólnieniem procedury optymalizacji na różne funkcje strat i macierze wejściowe o znanych odległościach z wagami
- Wykorzystywana jest funkcją kryterium straty nazywana stres
- Zazwyczaj minimalizowana jest się z wykorzystaniem procedury *stress majorization*.
- Metryczny MDS minimalizuje funkcję kryterium stress, która jest resztową sumą kwadratów

$$Stress(x_1, \ldots, x_n) = \sqrt{\sum_{i \neq j = 1,\ldots,N} (d(i,j) - ||x_i - x_j||)^2}$$

Definicja

- Wielowymiarowa wartość odstająca to wartość odstająca jednocześnie w więcej niż jednym wymiarze
- Zatem jest to kombinacja wartości obserwacji obejmująca kilka zmiennych
- Może to być na przykład osoba o wzroście 2 metry (powyżej 95 percentyla rozkładu wzrostu) i wadze 50 kg (nie więcej niż 5 percentyl rozkładu) jednocześnie

Analiza graficzna

- Najprostszym sposobem wykrywania wielowymiarowych obserwacji odstających jest analiza graficzna
- Jej wykorzystanie jest ograniczone do analizy dwóch (czasami trzech) wymiarów jednocześnie
- Obserwacje odstające do takie, które znajdują się daleko od centrum wykresu
- Problem: przeanalizowanie wszystkich kombinacji zmiennych jest czasochłonne

Odległość Mahalanobisa

- Statystycznym narzędziem służącym m.in. do wykrywania wielowymiarowych obserwacji odstających jest odległość Mahalanobisa
- Jest zdefiniowana jako odległość punktu od rozkładu
- Można ją interpretować jako wielowymiarowe uogólnienie studentyzowanych reszt

Odległość Mahalanobisa

$$dist = \sqrt{(x - \mu_x)' C^{-1} (x - \mu_x)}$$

gdzie C oznacza empiryczną macierz kowariancji

- Jeżeli dane pochodzą z rozkładu normalnego to odległość Mahalanobisa jest zmienną losową z rozkładu χ-kwadrat, o liczbie stopni swobody równej liczbie wymiarów

- Sposób postępowania jest efektem subiektywnego wyboru badacza
- Z matematycznego punktu widzenia nie ma dobrej ani złej odpowiedzi na pytanie, jak traktować obserwacje odstające
- Ważniejsza jest wiedza odnosząca się do badanych zjawisk, czy w jaki sposób powstała wartość odstające
- Pierwszym krokiem jest decyzja czy są to obserwacje nietypowe czy błędne

Obserwacje błędne

- Błędna obserwacja odstająca to obserwacja odstająca, która wynika z niedokładnego pomiaru, błędnego wprowadzenia danych lub jest wynikiem manipulacji danymi.
- Takie obserwacje zwykle nie należą do populacji będącej przedmiotem zainteresowania
- Błędne obserwacje należy skorygować albo usunąć
- Usunięcie błędnych obserwacji zapobiega ewentualnym obciążeniem i nie niesie ryzyka związanego z utratą informacji

Winsoryzacja

- Winsoryzacja została zaproponowana przez Tukeya i McLaughlina w 1963 r.
- Polega na zastąpieniu każdej wartość zmiennej powyżej lub poniżej k-tego percentyla wartością samego k-tego percentyla
- W porównaniu z przycinaniem, winsoryzacja jest mniej ekstremalną opcją polegającą na przekodowaniu wartości odstających zamiast ich usuwania ze zbioru
- Winsoryzacji nie należy stosować, gdy celem jest poszukiwanie wartości odstających

- Szykując część dotyczącą obserwacji odstających korzystałem z materiałów udostępnionych przez Alicię Horsch na blogu *Towards Data Science* `https://towardsdatascience.com/detecting-and-treating-outliers-in-python`