Eksploracja danych

Piotr Lipiński

Lista zadań nr 3 – Grupowanie danych

Zadanie 0. *(2 punkty)*
Zaimplementuj algorytm K-Means. Postaraj się, żeby implementacja była efektywna, bo będzie on w przyszłości używany do przetwarzania dużych zbiorów danych. Sprawdź działanie zaimplementowanego algorytmu na danych IRIS.

Zadanie 1. *(1 punkt)*
Napisz program, który generuje zbiór N = 5000 losowych wektorów danych z mieszaniny rozkładów gaussowskich o niżej podanych parametrach i sprawdź działanie algorytmu K-Means na tym zbiorze (rozkład Choleskiego macierzy kowariancji można w Matlabie uzyskać poleceniem chol, a w Pythonie poleceniem cholesky).
a) d = 2, K = 5, $\mathbf{p} = 1/K * \mathbf{1}$, $\mu_k = 3k * \mathbf{1}$, $\Sigma_k = \mathbf{I}$ (dla k = 1, 2, …, K),
b) jak w a), oprócz $\Sigma_3 = [3\ 0; 0\ 1]$,
c) jak w b), oprócz $\Sigma_1 = [3\ 1; 1\ 1]$,
d) jak w c), oprócz $\mathbf{p} = [0.2; 0.1; 0.3; 0.1; 0.3]$,
e) jak w a), oprócz d = 3 i $\Sigma_3 = [3\ 1\ 0; 1\ 1\ 0; 0\ 0\ 1]$,
f) jak w a), oprócz d = 100 i K = 10.
(**1** to wektor odpowiedniej długości złożony z samych jedynek, **I** to macierz identycznościowa).
Jak pogrupowałbyś te dane, jeśli nie znałbyś wartości K użytej w generatorze?

Zadanie 2. *(1 punkt)*
Wygeneruj zbiór N = 1000000 losowych danych z d = 1000 wymiarowej mieszaniny K = 1000 rozkładów gaussowskich takiej, że odległość między środkami μ_k każdych dwóch rozkładów tej mieszaniny jest nie mniejsza niż q = 10. Prawdopodobieństwa wyboru rozkładów mogą być równe, $\mathbf{p} = 1/K * \mathbf{1}$, a same rozkłady mogą mieć nieskorelowane zmienne, $\Sigma_k = \mathbf{I}$. Sprawdź działanie algorytmu K-Means na tym zbiorze. Poeksperymentuj z innymi wartościami parametru q.

Zadanie 3. *(2 punkty)*
W UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/) można znaleźć przykładowe dane do testowania działania algorytmów eksploracji danych. Pobierz 5 zestawów danych (Iris, Wine oraz 3 inne wybrane przez siebie). Zapoznaj się z opisem tych danych. Spróbuj opisać charakterystykę tych danych. Spróbuj przedstawić te dane na wykresach. Sprawdź jak dobrze można je pogrupować za pomocą znanych Ci algorytmów grupowania (sprawdź co najmniej K-Means).

Zadanie 4. *(2 punkty)*
Zdjęcie fotograficzne o rozdzielczości A x B w formacie RGB można traktować jako zbiór N = A * B trójwymiarowych punktów danych. Po przeprowadzeniu grupowania takich danych, z liczbą grup K, można obniżyć liczbę kolorów zdjęcia do K (tworząc paletę K kolorów określonych przez środki uzyskanych grup, a następnie zastępując kolor każdego piksela najbliższym mu kolorem ze stworzonej palety). Napisz program wczytujący zdjęcie z pliku JPG (polecenie imread w Matlabie lub w Pythonie), redukujący liczbę kolorów w powyższy sposób i wyświetlający zredukowane zdjęcie na ekranie. Sprawdź działanie programu na 5 wybranych przez siebie zdjęciach.

Zadanie 5. *(2 punkty)*
Zbiór danych Kosarak (stworzony przez Ferenca Bodona, dostępny m.in. na http://fimi.ua.ac.be/data/kosarak.dat.gz) zawiera informacje o transakcjach zawartych w sieci

supermarketów. Każdy wiersz tego pliku reprezentuje jedną transakcję i zawiera identyfikatory produktów kupionych w tej transakcji.

a) Wyznacz T = 1000 najczęściej kupowanych produktów (tzn. produktów, które pojawiły się w największej liczbie transakcji). Oznaczmy je kolejno Z_1, Z_2, …, Z_T.

b) Dla każdego produktu, policz ile razy był on kupowany razem z produktem Z_i (dla i =1, 2, …, T) tworząc w ten sposób wektor liczbowy $\mathbf{p} = (p_1, p_2, \ldots, p_T)$ opisujący dany produkt.

c) Pogrupuj produkty używając ich powyższej reprezentacji oraz algorytmu K-Means z różną liczbą grup (poeksperymentuj).

Przedstaw wyniki swoich eksperymentów. Możesz także poeksperymentować z parametrem T. Które grupy produktów można uznać za szczególnie istotne, a które za raczej przypadkowe?