Eksploracja danych

Piotr Lipiński

Lista zadań nr 7 – redukcja wymiarowości - PCA

Zadanie 1. *(4 punkty + 4 punkty bonusowe za oddanie zadania przed 12 stycznia)*
a) Utwórz zestaw danych **X** składający się z 1000 wektorów dwuwymiarowych $\mathbf{x}_i = (x_{1i}, x_{2i})^T$, dla i = 1, 2, …, 1000, wygenerowanych losowo przy użyciu dwuwymiarowego rozkładu normalnego o średniej [3, 5] i macierzy kowariancji [12, 3; 3, 1].
b) Ustandaryzuj dane, tak aby średnia dla każdego z wymiarów wynosiła 0, a wariancja 1. Ustandaryzowane dane oznaczmy przez $\mathbf{X}^{(0)}$.
c) Policz macierz kowariancji **S** ustandaryzowanego zestawu danych $\mathbf{X}^{(0)}$.
d) Wyznacz wartości własne λ_1 i λ_2 oraz odpowiadające im wektory własne $\mathbf{v}_1$ i $\mathbf{v}_2$ macierzy kowariancji **S**. Dla ustalenia notacji, wartości własne porządkujemy malejąco, tzn. $\lambda_1 > \lambda_2$.
e) Wyznacz składowe główne zestawu danych, tzn. dla każdego punktu danych $\mathbf{x}_i^{(0)}$ wyznacz punkt $\mathbf{y}_i = (1/\mathrm{sqrt}(\lambda_1)\ \mathbf{v}_1^T\ \mathbf{x}_i^{(0)},\ 1/\mathrm{sqrt}(\lambda_2)\ \mathbf{v}_2^T\ \mathbf{x}_i^{(0)})^T$. Odpowiada to zrzutowaniu punktu $\mathbf{x}_i^{(0)}$ na osie nowego układu współrzędnych wyznaczonego przez wektory własne macierzy S i przeskalowaniu przez pierwiastki z wartości własnych.
f) Zrób rysunki pokazujące oryginalny zbiór danych (punkty $\mathbf{x}_i$), zbiór danych po standaryzacji (punkty $\mathbf{x}_i^{(0)}$) i zbioru danych po przekształceniu PCA (punkty $\mathbf{y}_i$). Na rysunkach z punktami $\mathbf{x}_i$ i $\mathbf{x}_i^{(0)}$ narysuj proste zawierające osie główne elipsy wyznaczanej przez punkty danych.
g) Sprawdź charakterystykę statystyczną (średnią, wariancję, macierz kowariancji i macierz korelacji) oryginalnego zbioru danych, zbioru danych po standaryzacji i zbioru danych po przekształceniu PCA.
h) Wyjaśnij dlaczego licząc punkty $\mathbf{y}_i$ dzielimy przez pierwiastki z wartości własnych. Jak wyglądałyby wyniki f) i g), gdybyśmy nie wykonywali tego dzielenia?

i) Powtórz wszystkie powyższe obliczenia dla zestawu danych **X** składającego się z 1000 wektorów dwuwymiarowych $\mathbf{x}_i = (x_{1i}, x_{2i})^T$, dla i = 1, 2, …, 1000, wygenerowanych losowo przy użyciu mieszaniny trzech rozkładów normalnych o średnich [-21, -2], [3, 5], [27, 12], macierzy kowariancji [12, 3; 3, 1] takiej samej dla wszystkich trzech rozkładów oraz wag równych dla wszystkich trzech rozkładów wynoszących 1/3.

j) Powtórz b), c), d) i e) dla zestawu danych IRIS. Zrób dwuwymiarowy rysunek ilustrujący zbiór danych IRIS wykreślając na osi x pierwszą składową główną y_{1i}, a na osi y_{2i} drugą składową główną każdego punktu $\mathbf{x}_i$.
k) Dla zestawu danych IRIS, spróbuj odtworzyć oryginalne punkty danych $\mathbf{x}_i$ z danych zredukowanych do dwóch składowych głównych, tzn. z dwuwymiarowych punktów $[y_{1i}, y_{2i}]^T$. Policz średniokwadratowy błąd odtworzenia, tzn. sumę kwadratów odległości między oryginalnym punktem danych a odtworzonym punktem danych.
l) Jakie znaczenie ma suma wartości własnych macierzy kowariancji **S**, które zostały użyte do konstrukcji danych zredukowanych?

Wskazówka: Dołączony do listy zadań skrypt Matlaba, omawiany na wykładzie, pokazuje jak przykładowo wykonać część zadania w Matlabie.

Zadanie 2. *(4 punkty)*
Wybierz 5 zestawów danych z UCI Machine Learning Repository (proponuję wybrać te same zestawy co w zadaniu 3 z listy 3). Dla każdego zestawu danych, zrób dwuwymiarowy rysunek pierwszej i drugiej składowej głównej. Czy pokazuje on coś więcej niż wykresy par atrybutów danych oryginalnych? Następnie zredukuj wymiarowość danych (zdecyduj do ilu wymiarów) i sprawdź jak dobrze można je pogrupować za pomocą znanych Ci algorytmów grupowania (sprawdź co najmniej KMeans). Wyniki porównaj do grupowania oryginalnych, niezredukowanych, danych. Jeżeli wybrane dane dotyczą problemu klasyfikacji (mają etykietę klasy), to zrób podobne porównanie dotyczące klasyfikacji danych (klasyfikacji oryginalnych danych i klasyfikacji danych zredukowanych).

Zadanie 3. *(4 punkty)*
Zapoznaj się z dołączonymi do listy zadań danymi (pochodzącymi z bazy danych AR Face Database[1] stworzonej przez prof. Aleixa Martineza z Ohio State University i udostępnionej przez niego na potrzeby naszego wykładu). Przygotowany przeze mnie zbiór danych zawiera odpowiednio przeskalowane zdjęcia twarzy o rozdzielczości 82 x 60 pikseli w 256 odcieniach szarości każdy (zatem zdjęcie twarzy można utożsamić z punktem przestrzeni 82 x 60 = 4920 wymiarowej). Wczytaj pierwszy zestaw danych (zawierający 250 zdjęć – po 5 zdjęć każdej z 50 osób), spróbuj zredukować ich wymiarowość. Dokładnie przeanalizuj wyniki (jak można interpretować uzyskane wektory własne?).

Zadanie 4. *(2 punkty)*
Zaimplementuj prostą metodę rozpoznawania twarzy:

- wczytane są zdjęcia wzorcowe (5 zdjęć każdej z 50 osób), czyli 250 wektorów 4920-wymiarowych $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{250} \in \mathbf{R}^{4920}$,
- wiadomo które zdjęcia wzorcowe odpowiadają którym osobom (w zbiorze danych zdjęcia osób są zapisane kolejno: 5 zdjęć osoby nr 1, 5 zdjęć osoby nr 2, ..., 5 zdjęć osoby nr 50), co można określić funkcją osoba($\mathbf{x}_k$) = (k-1) div 5 + 1,
- na wejściu podawane jest zdjęcie nieznanej osoby, czyli wektor 4920 wymiarowy $\mathbf{y} \in \mathbf{R}^{4920}$,
- dla wektora **y** należy wyznaczyć najbliższy mu wektor $\mathbf{x}_k$, dla k = 1, 2, …, 250, (najbliższy w sensie odległości euklidesowej w $\mathbf{R}^{4920}$),
- jeżeli odległość między **y** a $\mathbf{x}_k$ nie przekracza pewnego ustalonego progu, to można uznać, że **y** jest zdjęciem twarzy osoby osoba($\mathbf{x}_k$).

Wczytaj drugi zestaw danych (zawierający 100 zdjęć – po 2 zdjęcia każdej z 50 osób). Przetestuj na tym zestawie danych efektywność zaimplementowanej metody (zdjęcia wzorcowe pochodzą z pierwszego zestawu danych, drugi zestaw danych jest używany tylko do testów). Dokładnie przeanalizuj wyniki.

Zadanie 5. *(2 punkty)*
Zmień metodę rozpoznawania twarzy z poprzedniego zadania w taki sposób, że zamiast pracować w przestrzeni $\mathbf{R}^{4920}$ będziemy pracować w przestrzeni mniej wymiarowej (wyznaczonej przez redukcję wymiarowości pierwszego zestawu danych metodą PCA). Porównaj tę metodę z metodą oryginalną (używając do testów drugiego zestawu danych). Dokładnie przeanalizuj wyniki.

Uwaga: Przekształcenie redukujące wymiarowość powinno być określone wyłącznie w oparciu o pierwszy zestaw danych (do drugiego zestawu danych należy zastosować to samo przekształcenie bez wprowadzania żadnych modyfikacji).

[1] Martinez, A., M., Benavente, R., The AR Face Database, CVC Technical Report #24, 1998.

Zadanie 6. *(2 punkty)*
Zamiast proponowanej prostej metody rozpoznawania twarzy użyj klasyfikatora KNN. Porównaj skuteczność klasyfikacji dla różnych k, dla danych oryginalnych i danych zredukowanych (dla różnej liczby wymiarów). Porównanie zrób metodą cross-validation dla połączonych obu zestawów danych.

Zadanie 7. *(nieobowiązkowe - 4 punkty bonusowe)*
Przekształć opracowany system rozpoznawania twarzy do wersji pracującej na zdjęciach o rozdzielczości 165 x 120 pikseli w kolorze (kodowanie RGB) lub w 256 odcieniach szarości. Porównaj działanie tego systemu z uproszczoną poprzednią wersją. Dokładnie przeanalizuj wyniki.

Uwaga: Ze względu na prawa autorskie do używanych zestawów danych, proszę o indywidualny kontakt emailem w celu uzyskania dostępu do bazy danych.