10. Redukcja wymiaru - metoda PCA

dr inż. Urszula Libal

Politechnika Wrocławska

2015

1. PCA

Analiza składowych głównych:

— w skrócie nazywana PCA (od ang. *Principle Component Analysis*)

— znana także transformacją Karhunena-Loeve'go (KLT).

— Polega na wybraniu k ortogonalnych n-wymiarowych wektorów, które najlepiej reprezentują dane, $k \leq n$.

— Oryginalne dane są rzutowane na przestrzeń rozpiętą przez k wybranych wektorów (**składowe główne**), co prowadzi do **redukcji wymiaru** wektorów cech (z n do k).

2. PCA w kilku krokach

1. Unormowanie cech
2. Obliczenie składowych głównych
3. Sortowanie składowych głównych od najmocniejszych do najsłabszych
4. Wybranie k znaczących składowych głównych i usunięcie pozostałych

2.1. Unormowanie

— Dane wejściowe (wektory cech) są **unormowane**, aby każda cecha wpadała do tego samego przedziału.

— Krok ten pomaga w zapewnieniu, że cechy szerzej rozłożone nie zdominują cech mocniej skoncentrowanych.

2.2. Obliczenie składowych głównych

— Następnie wylicza się k ortonormalnych wektorów, które tworzą bazę dla unormowanych danych wejściowych. Wektory te są to wektory jednostkowe, wskazujące w kierunku prostopadłym do pozostałych wektorów z utworzonej bazy.

Procedura PCA polega na wyliczeniu wartości własnych $\lambda_1^S, \lambda_2^S, \ldots, \lambda_n^S$ macierzy rozproszenia danych, np. macierzy kowariancji S. Dane są reprezentowane przez zestaw N wektorów cech $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ o n wymiarach, tj.

$$\left\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}\right\}.$$

Macierz rozproszenia S wyliczamy ze wzoru

$$S = \sum_{i=1}^{N} \left(\mathbf{x}^{(i)} - \overline{\mathbf{x}} \right) \left(\mathbf{x}^{(i)} - \overline{\mathbf{x}} \right)^T, \tag{1}$$

gdzie $\mathbf{x}^{(i)}$ to wektory cech, $i = 1, 2, \ldots, N$, a $\overline{\mathbf{x}}$ to ich empiryczna średnia.

Następnie wyznacza się wektory własne oraz wartości własne macierzy S, np. przy pomocy dekompozycji macierzy do postaci (tzw. rozkład spektralny macierzy S, [2])

$$S = A \Lambda A^T, \tag{2}$$

gdzie A to macierz wektorów własnych, a Λ to macierz diagonalna, na przekątnej której znajdują się **wartości własne macierzy** S: $\lambda_d^S,\ d = 1, 2, \ldots, n$.

2.3. Sortowanie składowych głównych

— Uporządkowujemy wartości własne macierzy kowariancji S w kolejności malejącej

$$\lambda_1^S \geqslant \lambda_2^S \geqslant \ldots \geqslant \lambda_n^S \geqslant 0. \tag{3}$$

— Redukcja cech opiera się na wyznaczeniu podzbioru cech w nowej przestrzeni, rozpiętej przez ortonormalne składowe główne. Nowy zestaw cech po transformacji jest wyznaczony według zasady **maksymalizującej zmienność danych** wraz z jednoczesną minimalizacją ubytku informacji spowodowanej ich redukcją.

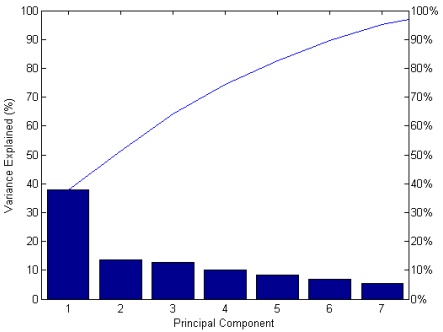
2.4. Selekcja k składowych głównych

Analiza skumulowanej wariancji k składowych głównych $(k \leqslant n)$ opiera się na procentowej mierze *var* wyjaśniania zmienności danych przez pierwszych k składowych głównych, która jest zdefiniowana następująco

$$var = \left(\frac{\sum_{d=1}^{k} \lambda_d^S}{\sum_{d=1}^{n} \lambda_d^S} \right) \times 100\%. \tag{4}$$

Technika PCA zakłada, że jeżeli wartości danej cechy w_d $(d = 1, 2, \ldots, n)$ charakteryzują się dużą wariancją, a odpowiadająca jej wartość własna λ_d^S przyjmuje dużą wartość, to cecha ta posiada dużą wartość informacyjną, np. dobrze dyskryminuje obiekty z różnych klas.

Redukcja wymiaru polega na wybraniu tylko tych składowych głównych (k składowych z n), dla których zmienność danych *var* (4) jest nie mniejsza od założonego procentu wyjaśnianej zmienności danych (np. 90%).



Rysunek 1. Procent wyjaśnienia zmienności danych przez kolejne składowe główne.

Źródło: opracowanie własne

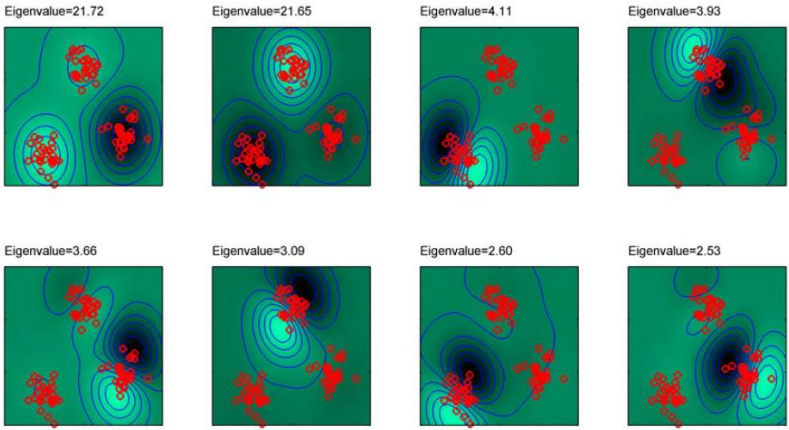
W ostatnim kroku dokonujemy **transformacji (rzutowania)** wektorów cech $\mathbf{x}^{(i)}$, $i = 1, 2, \ldots, N$, do nowego układu współrzędnych rozpiętego przez wektory własne macierzy S. Przekształcenie to jest zwane rozwinięciem lub transformacją Karhunena-Loevego (patrz np. [5]).

Wektory cech $\mathbf{x}^{(i)}$ zostają przemnożone przez macierz $A_{n\times k}$ zawierającą tylko k kolumn macierzy wektorów własnych A, odpowiadających k największym wartościom własnym,

$$\mathbf{x}' = A_{n\times k}^T \mathbf{x}. \tag{5}$$

W wyniku obrotu układu współrzędnych n-wymiarowe wektory cech $\mathbf{x}^{(i)}$ zostają w ten sposób przekształcone w wektory cech $\mathbf{x}' = (x'_1, x'_2, \ldots, x'_k)^T$ o jedynie k składowych.

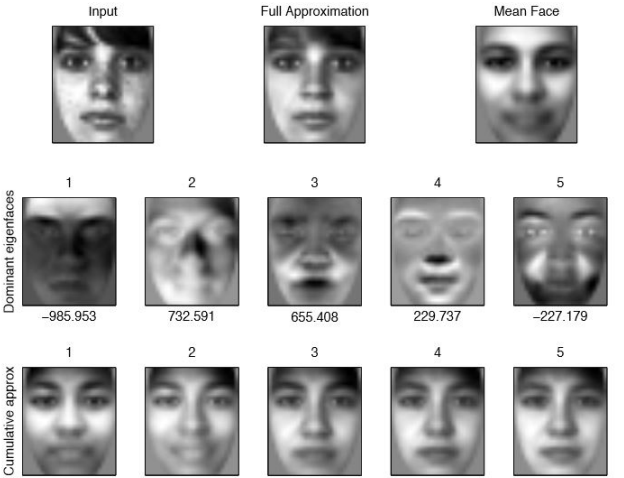
Transformacja Karhunena-Loevego posiada tę własność, że dowolna para współrzędnych nowego układu (tj. cech x'_{n_1} i x'_{n_2}, $n_1, n_2 \in \{1, 2, \ldots, n\}$) jest wzajemnie **nieskorelowana** (patrz np. [5]). Dlatego transformacja Karhunena-Loevego może zostać użyta w celu usunięcia korelacji cech. Z kolei wektory cech poddane selekcji przy użyciu metody PCA mogą następnie zostać użyte przy klasyfikacji obiektów.



Rysunek 2. Przykład jądrowego PCA z gaussowskim jądrem $\kappa(\mathbf{x}, \mathbf{x}') = \exp(-||\mathbf{x} - \mathbf{x}'||^2/0.1)$.

Źródło: [6]

4. Eigenfaces



Rysunek 3. Eigenfaces (5 pierwszych z 50).

Źródło: [3]

Literatura

[1] J. Han, M. Kamber, J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Elsevier, (2012)

[2] J. Koronacki, J. Ćwik, *Statystyczne systemy uczące się*, WNT, Warszawa (2005)

[3] M. Turk and A. Pentland, *Face recognition using eigenfaces*, Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–591, (1991)

[4] A.R. Webb, K.D. Copsey, *Statistical Pattern Recognition*, 3rd ed., Wiley (2011)

[5] W. Sobczak, W. Malina, *Metody selekcji i redukcji informacji*, WNT, Warszawa (1985)

[6] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer Series: Information Science and Statistics (2006)