Analiza głównych składowych, wykład 4

Joanna Jędrzejowicz

Instytut Informatyki

Konieczność redukcji wymiaru w eksploracji danych

- bazy danych spotykane w zadaniach eksploracji danych mają zwykle miliony rekordów i tysiące zmiennych; cześć zmiennych jest zwykle ze sobą powiązana,

Konieczność redukcji wymiaru w eksploracji danych

- bazy danych spotykane w zadaniach eksploracji danych mają zwykle miliony rekordów i tysiące zmiennych; cześć zmiennych jest zwykle ze sobą powiązana,
- użycie dużej liczby zmiennych może prowadzić do nadmiernego dopasowania (ang. overfitting),

Konieczność redukcji wymiaru w eksploracji danych

- bazy danych spotykane w zadaniach eksploracji danych mają zwykle miliony rekordów i tysiące zmiennych; cześć zmiennych jest zwykle ze sobą powiązana,
- użycie dużej liczby zmiennych może prowadzić do nadmiernego dopasowania (ang. overfitting),
- w niektórych zadaniach (np. analiza obrazów) utrzymanie pełnej wymiarowości utrudnia rozwiązanie

Cele metod redukcji wymiaru

- zmniejszenie liczby elementów opisujących,
- dążenie do niezależności elementów,
- stworzenie szkieletu do interpretacji wyników

Analiza głównych składowych - jedna z metod redukcji wymiaru

Analiza głównych składowych (ang. principal component analysis PCA) służy do wyszukiwania wzorców i polega na zastąpieniu struktury zmiennych przy użyciu mniejszego zbioru kombinacji liniowych tzw głównych składowych

- założmy, że oryginalne zmienne (inaczej: atrybuty) tworzą układ współrzędnych w m-wymiarowej przestrzeni; składowe główne reprezentują nowy układ współrzędnych, znaleziony poprzez rzutowanie oryginalnego układu wzdłuż kierunków maksymalnej zmienności,

Analiza głównych składowych - jedna z metod redukcji wymiaru

Analiza głównych składowych (ang. principal component analysis PCA) służy do wyszukiwania wzorców i polega na zastąpieniu struktury zmiennych przy użyciu mniejszego zbioru kombinacji liniowych tzw głównych składowych

- załóżmy, że oryginalne zmienne (inaczej: atrybuty) tworzą układ współrzędnych w m-wymiarowej przestrzeni; składowe główne reprezentują nowy układ współrzednych, znaleziony poprzez rzutowanie oryginalnego układu wzdłuż kierunków maksymalnej zmienności,
- oryginalne m zmiennych zostaje zamienione przez $k < m$ nowych zmiennych

Zastosowanie kowariancji

Kowariancja jest mierzona dla pary atrybutów (zmiennych) i mierzy zmiennosc danych wzgledem siebie.

$$cov(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \mu_X) \times (Y_i - \mu_Y)}{n-1}$$

Jesli uzyjemy tych samych zmiennych, to mamy wariancję.

PCA - kroki początkowe

- standaryzacja zmiennych tak, aby średnia każdej zmiennej była 0,

PCA - kroki początkowe

- standaryzacja zmiennych tak, aby średnia każdej zmiennej była 0,
- jeżeli zmienna X jest wektorem wymiaru n (liczba rekordów), to zmienna standaryzowana jest także wymiaru n oraz powstaje przez odjęcie wartości średniej μ (od każdego rekordu),

PCA - kroki początkowe

- standaryzacja zmiennych tak, aby średnia każdej zmiennej była 0,
- jeżeli zmienna X jest wektorem wymiaru n (liczba rekordów), to zmienna standaryzowana jest także wymiaru n oraz powstaje przez odjęcie wartości średniej μ (od każdego rekordu),
- utworzenie macierzy kowariancji C; macierz jest symetryczna, wymiaru mxm - na miejscu (i, j) stoi kowariancja między zmienną i-tą oraz j-tą (m jest liczbą zmiennych - atrybutów)

$$cov(X, Y) = \frac{\sum_{i=1}^{n} (X_i - \mu_X) \times (Y_i - \mu_Y)}{n - 1}$$

PCA - kroki początkowe

- standaryzacja zmiennych tak, aby średnia każdej zmiennej była 0,
- jeżeli zmienna X jest wektorem wymiaru n (liczba rekordów), to zmienna standaryzowana jest także wymiaru n oraz powstaje przez odjęcie wartości średniej μ (od każdego rekordu),
- utworzenie macierzy kowariancji C; macierz jest symetryczna, wymiaru mxm - na miejscu (i,j) stoi kowariancja między zmienną i-tą oraz j-tą (m jest liczbą zmiennych - atrybutów)

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \mu_X) \times (Y_i - \mu_Y)}{n-1}$$

μ_X, μ_Y oznaczają srednie (u nas 0)

PCA - kroki początkowe, kowarianacja

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \mu_X) \times (Y_i - \mu_Y)}{n-1}$$

Kowariancja jest miarą tego, jak dwie zmienne różnią się od siebie. Dodatnia wartość kowariancji oznacza, że jeśli jedna zmienna rośnie, to druga ma również tendencję do wzrostu. Wartość ujemna oznacza, że jeśli jedna zmienna rośnie, to druga maleje. por. dane w arkuszu wyklad3i5.xls

Przyklad

Dane				po standaryzacji	
	X_1	X_2		X_1	X_2
	2,5	2,4		0,69	0,49
	0,5	0,7		-1,31	-1,21
	2,2	2,9		0,39	0,99
	1,9	2,2		0,09	0,29
	3,1	3		1,29	1,09
	2,3	2,7		0,49	0,79
	2	1,6		0,19	-0,31
	1	1,1		-0,81	-0,81
	1,5	1,6		-0,31	-0,31
	1,1	0,9		-0,71	-1,01
srednia	**1,81**	**1,91**		**0,00**	**0,00**

Przyklad, cd

liczba zmiennych $m = 2$, macierz kowariancji jest wymiaru $2x2$

macierz kowariancji		
0,6166	0,6154	
0,6154	0,7166	

PCA - kolejne kroki

- znalezienie wartości własnych macierzy kowariancji C oraz odpowiadających im wektorów własnych:
  λ jest wartością własną macierzy C odpowiadającą wektorowi własnemu x jeżeli

$$C \cdot x = \lambda \cdot x$$

- dla danych z przykładu

$$wartosciWlasne = \begin{pmatrix} 0,049083 \\ 1,284027 \end{pmatrix}$$

$$wektoryWlasne = \begin{pmatrix} -0,735170 & -0,677873 \\ 0,677873 & -0,735178 \end{pmatrix}$$

pierwsza kolumna odpowiada pierwszej wartosci własnej

PCA - kolejne kroki, sprawdzenie

$$wartosciWlasne = \begin{pmatrix} 0,049083 \\ 1,284027 \end{pmatrix}$$

$$wektoryWlasne = \begin{pmatrix} -0,735170 & -0,677873 \\ 0,677873 & -0,735178 \end{pmatrix}$$

sprawdzenie dla drugiej wartości:

$$\begin{pmatrix} 0,6165 & 0,6154 \\ 0,6154 & 0,7166 \end{pmatrix} \cdot \begin{pmatrix} -0,677873 \\ -0,735178 \end{pmatrix} = 1,28 \cdot \begin{pmatrix} -0,677873 \\ -0,735178 \end{pmatrix}$$

wartości własne sortuje się w porządku nierosnącym (i odpowiadające im wektory własne)

PCA - kolejne kroki, sprawdzenie

$$wartosciWlasne = \begin{pmatrix} 0,049083 \\ 1,284027 \end{pmatrix}$$

$$wektoryWlasne = \begin{pmatrix} -0,735170 & -0,677873 \\ 0,677873 & -0,735178 \end{pmatrix}$$

sprawdzenie dla drugiej wartości:

$$\begin{pmatrix} 0,6165 & 0,6154 \\ 0,6154 & 0,7166 \end{pmatrix} \cdot \begin{pmatrix} -0,677873 \\ -0,735178 \end{pmatrix} = 1,28 \cdot \begin{pmatrix} -0,677873 \\ -0,735178 \end{pmatrix}$$

wartości własne sortuje się w porządku nierosnącym (i odpowiadające im wektory własne)

- wektor własny odpowiadający największej wartości nazywany jest składową główną

PCA - kolejne kroki

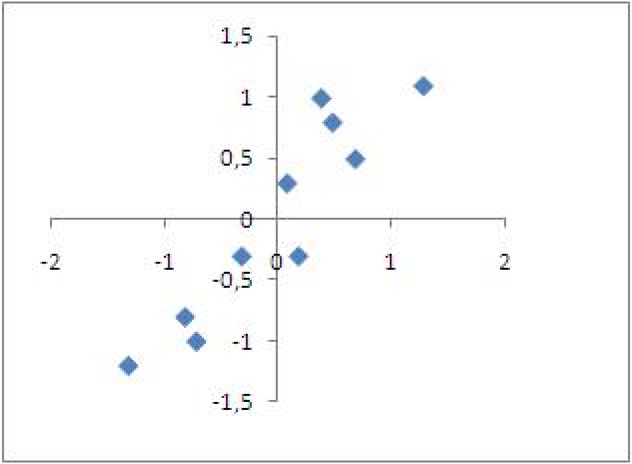
Po uporządkowaniu wartości własnych (i odpowiadających im wektorów własnych) otrzymuje się uporządkowanie głównych składowych w kolejności od najbardziej istotnych.

PCA - kolejne kroki

Po uporządkowaniu wartości własnych (i odpowiadających im wektorów własnych) otrzymuje się uporządkowanie głównych składowych w kolejności od najbardziej istotnych.
Redukcja polega na zmniejszeniu wymiaru i odrzuceniu najmniejszych wartości własnych i odpowiadajacych im wektorów własnych.

Przyklad, cd

na wykresie poniżej umieszczono zmienne zestandaryzowane: skupienie danych wokół wektora własnego $(-0,68; -0,73)$ odpowiadajacego większej wartości własnej



PCA - nowe współrzędne

Niech D_{nxm} oznacza macierz danych początkowych, n - liczba rekordów, m - liczba atrybutów,
X_{mxm} - macierz zawierająca w **wierszach** wektory własne,

- nowe współrzędne:

$$\hat{D} = X \times D^T$$

D^T - oznacza macierz transponowaną

PCA - nowe współrzędne

Niech D_{nxm} oznacza macierz danych początkowych, n - liczba rekordów, m - liczba atrybutów,
X_{mxm} - macierz zawierająca w **wierszach** wektory własne,

- nowe współrzędne:

$$\hat{D} = X \times D^T$$

  D^T - oznacza macierz transponowaną
- korzystając z tego, że macierz odwrotna X^{-1} jest równa X^T (wynika z postaci X) możemy wyrazic początkowe dane w terminach nowego układu

$$D^T = X^T \times \hat{D}$$

Przyklad, cd

Korzystając ze wzorów z poprzedniego przeźrocza i ograniczając się do jednej wartości własnej

$$\hat{D} = \begin{pmatrix} -0.6787 & -0.73517 \end{pmatrix} \times$$

$$\begin{pmatrix} 0.69 & -1.31 & 0.39 & 0.09 & 1.29 & 0.49 & 0.19 & -0.81 & -0.31 & -0.71 \\ 0.49 & -1.21 & 0.99 & 0.29 & 1.09 & 0.79 & -0.31 & -0.81 & -0.31 & -1.01 \end{pmatrix}$$

$$=$$

$$\begin{pmatrix} -0.83 & 1.78 & -0.99 & -0.27 & -1.68 & -0.91 & 0.10 & 1.14 & 0.44 & 1.22 \end{pmatrix}$$

np. -0.83=-0.6787*0.69+(-0.73517)*0.49, itd

Przyklad, cd

stosujac wzór $D^T = X^T \times \hat{D}$ otrzymujemy poczatkowy zbiór punktów skupionych wokół głównej składowej

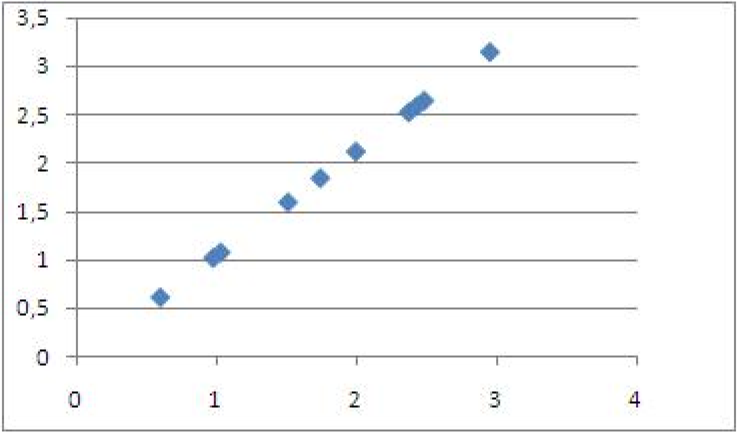
$$X^T \times \hat{D} = \begin{pmatrix} -0.6787 \\ -0.73517 \end{pmatrix} \times$$

$$\begin{pmatrix} -0.83 & 1.78 & -0.99 & -0.27 & -1.68 & -0.91 & 0.10 & 1.14 & 0.44 & 1.22 \end{pmatrix}$$

$$=$$

$$\begin{pmatrix} 2.37 & 0.60 & \cdots & 0.98 \\ 2.51 & 0.60 & \cdots & 1.01 \end{pmatrix}$$

Przyklad, cd



PCA - ile składowych należy wybrać?

- kryterium wartości własnej,
- kryterium części wariancji wyjaśnionej przez składowe główne,

Kryterium wartości własnej

Każda składowa powinna wyjasniać przynajmniej taką zmienność, jak zmienna pierwotna i dlatego kryterium wartości własnej stwierdza, że tylko składowe o wartościach własnych większych od 1 należy pozostawić do analizy

Kryterium części wariancji wyjaśnionej przez składowe główne

W tym kryterium analityk najpierw określa jaka część zmienności ma zostać wyjaśniona przez składowe główne. Wybiera się po kolei składowe, aż do momentu gdy zostanie osiągnieta oczekiwana wartość zmienności.

Kryterium części wariancji wyjaśnionej przez składowe główne

W tym kryterium analityk najpierw określa jaka część zmienności ma zostać wyjaśniona przez składowe główne. Wybiera się po kolei składowe, aż do momentu gdy zostanie osiągnieta oczekiwana wartość zmienności.
Każda wartość własna zwraca wariancję względem odpowiedniej składowej, w przykładzie wartość własna 1.28 odpowiada wariancji
$1.28/(1.28+0.049)=96{,}31\%$.

Kryterium części wariancji wyjaśnionej przez składowe główne

W tym kryterium analityk najpierw określa jaka część zmienności ma zostać wyjaśniona przez składowe główne. Wybiera się po kolei składowe, aż do momentu gdy zostanie osiągnieta oczekiwana wartość zmienności.
Każda wartość własna zwraca wariancję względem odpowiedniej składowej, w przykładzie wartość własna 1.28 odpowiada wariancji
$1.28/(1.28+0.049)$=96,31%.
Wartość może zależeć od dziedziny, w której prowadzone sa badania, np nauki socjologiczne versus przyrodnicze- socjologowie mogą być zadowoleni, jeżeli składowe wyjasniaja 60 % zmienności, bo czynnik ludzki jest bardzo nieprzewidywalny, w naukach przyrodniczych oczekuje się wyższego procentu, bo miary są mniej zmienne

Przykłady zastosowań PCA, przykład 1

- Załóżmy, że danych jest 20 obrazków. Każdy obrazek jest prostokątem złożonym z $N \times N$ pikseli. Zatem każdy obrazek może być zapamiętany jako wektor

$$X = (x_1, \cdots, x_{N^2})$$

zaś 20 obrazów jako macierz

$$macierzObrazow = \begin{pmatrix} X_1 \\ X_2 \\ \cdots \\ X_{20} \end{pmatrix}$$

Przykłady zastosowań PCA, przykład 1

- Załóżmy, że danych jest 20 obrazków. Każdy obrazek jest prostokątem złożonym z $N \times N$ pikseli. Zatem każdy obrazek może być zapamiętany jako wektor

$$X = (x_1, \cdots, x_{N^2})$$

zaś 20 obrazów jako macierz

$$macierzObrazow = \begin{pmatrix} X_1 \\ X_2 \\ \cdots \\ X_{20} \end{pmatrix}$$

- Niech obrazki przedstawiają twarze i dla danego obrazka chcemy znaleźć jeden spośród danych 20, który jest najbardziej podobny. Po wykonaniu przekształcenia PCA możemy szukać szukać różnic nie wzgledem oryginalnych współrzędnych, tylko przekształconych i ograniczyć się do głównych składowych.

Przykład2 - kompresja obrazu

Jeżeli danych jest 20 obrazków, każdy złożony z $N \times N$ pikseli, to można także rozpatrywać N^2 wektorów- każdy 20 wymiarowy.
Czyli każdy wektor składa się z wartości *dla tego samego piksela*. Wykonując przekształcenie PCA otrzymujemy 20 wektorów własnych. Aby skompresować dane możemy w transformacji skorzystać np. z 15 wektorów - mamy do czynienia z kompresją *stratną*.