SkaiWD – Laboratorium 4

PCA

Iwo Błądek

20 kwietnia 2020

1 Iloczyn skalarny jako projekcja

Iloczyn skalarny (ang. dot product, inner product) to funkcja $(\mathbb{R}^n, \mathbb{R}^n) \to \mathbb{R}$ przyjmująca jako argumenty dwa wektory i zwracająca skalar, zdefiniowana następująco:

$$\text{dot}(a,b) = \langle a, b \rangle = a^T \cdot b = \begin{bmatrix} a_1 & \dots & a_N \end{bmatrix} \cdot \begin{bmatrix} b_1 \\ \dots \\ b_N \end{bmatrix} = a_1 b_1 + \dots + a_N b_N$$

Iloczyn skalarny wektorów a i b, w alternatywnym zapisie $\langle a, b \rangle$, ma ważną interpretację: **projekcję wektora a na b**. Projekcja ta będzie dodatkowo przeskalowana przez długość wektora, na który projektujemy (dlatego w praktyce warto używać wektorów długości 1). W ogólności zachodzi wzór:

$$\langle a, b \rangle = |a| \cdot |b| \cdot \cos \alpha \tag{1}$$

gdzie α to kąt między oboma wektorami. Zarówno z tego wzoru jak i interpretacji projekcji wynika między innymi to, że iloczyn skalarny wektorów prostopadłych zawsze wynosić będzie 0.

Jeżeli mamy jakąś *ortonormalną* (tj. znormalizowaną i ortogonalną) bazę przestrzeni liniowej zawierającą wektory $v_1, \dots, v_n$, to każdy wektor x w tej przestrzeni liniowej możemy reprezentować jako:

$$x = \langle x, v_1 \rangle v_1 + \dots + \langle x, v_n \rangle v_n \tag{2}$$

Iloczyn skalarny $\langle x, v_i \rangle$ mówi nam, w jakim punkcie prostej wektora v_i znajdziemy się gdybyśmy prostopadle zrzutowali na niego wektor x.

Przykład: Mamy wektor $x = \begin{bmatrix} 3 \\ 5 \end{bmatrix}$ i bazę ortonormalną $\{v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, v_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}\}$.

$$\langle x, v_1 \rangle = \begin{bmatrix} 3 & 5 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 3 \cdot 1 + 5 \cdot 0 = 3$$

$$\langle x, v_2 \rangle = \begin{bmatrix} 3 & 5 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 3 \cdot 0 + 5 \cdot 1 = 5$$

Wynika z tego, że możemy zapisać:

$$x = \langle x, v_1 \rangle v_1 + \langle x, v_2 \rangle v_2 = 3 \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 5 \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 5 \end{bmatrix}$$

Mnożenie macierzy można alternatywnie przedstawić jako projekcje wierszy pierwszej macierzy na kolumny drugiej:

$$\begin{bmatrix} a & c \\ b & d \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} a & c \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ \begin{bmatrix} b & d \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} ax_1 + cx_2 \\ bx_1 + dx_2 \end{bmatrix} = \begin{bmatrix} \langle \overline{ac}, x \rangle \\ \langle \overline{bd}, x \rangle \end{bmatrix}$$

Innymi słowy, standardowa perspektywa na mnożenie macierzy to perspektywa „projekcyjna", dualna do perspektywy „transformacji liniowej" omówionej wcześniej. Dużo obliczeń na macierzach staje się łatwiejszymi do zrozumienia, kiedy ma się na uwadze właściwą perspektywę. Przy PCA będziemy mieć okazję zetknąć się z przypadkiem, gdy pomocna będzie perspektywa „projekcyjna".

2 Wprowadzenie do PCA

Analiza składowych głównych, *PCA* (ang. *Principal Component Analysis*), przekształca dane z układu współrzędnych z wymiarami odpowiadającymi oryginalnym zmiennym (atrybutom) do układu współrzędnych, w którym wymiary są określone przez nowe zmienne będące pewnymi *kombinacjami liniowymi* oryginalnych zmiennych. Nowe zmienne są tak konstruowane, by każdy kolejny wymiar miał jak największą możliwą wariancję.

PCA znajduje następujące zastosowania:

- kompresja danych,
- odkrywanie istotnych cech.

Schemat działania PCA jest zaskakująco prosty. Można go przedstawić w następujących krokach, gdzie X to nasz zbiór danych:

Oblicz macierz kowariancji/korelacji S_X

$\downarrow$

Znajdź wektory własne S_X

$\downarrow$

Utwórz macierz K zawierającą w kolumnach wektory własne S_X posortowane malejąco po wartościach własnych

$\downarrow$

Pomnóż oryginalne dane przez K: $Y = XK$

Nasze przetransformowane dane $Y = XK$ (uwaga: mnożenie przez K jest prawostronne; więcej o tym później) będą mieć tę własność, iż wariancja pierwszej zmiennej będzie maksymalna możliwa[1], wariancja drugiej zmiennej będzie maksymalna możliwa przy ograniczeniu, że pierwsza była maksymalna, itd. Kowariancja (i korelacja) między dowolnymi nowymi zmiennymi w Y będzie zawsze wynosić 0.

2.1 Macierze kowariancji i korelacji

Wariancja (estymacja z próbki) zmiennej losowej A wyrażona jest wzorem:

$$\mathrm{var}(A) = \frac{\sum_{i=1}^{n}(A_i - \bar{A})^2}{n-1}$$

[1] Przy ograniczeniu, że nowe zmienne tworzone są jako kombinacja liniowa oryginalnych zmiennych. Nie można wykluczyć, i zazwyczaj tak w praktyce będzie, iż istnieje nieliniowa zależność lepiej oddająca zmienność w danych.

gdzie A_i to i'ty zaobserwowany element, $\bar{A}$ to średnia, a n to liczba elementów w próbce. Kowariancja rozkładu łącznego dwóch zmiennych A i B zdefiniowana jest z kolei jako:

$$\mathrm{cov}(A,B) = \frac{\sum_{i=1}^{n}(A_i - \bar{A})(B_i - \bar{B})}{n-1}$$

Jeżeli mamy macierz X z danymi zawierającą w wierszach obserwacje (przypadki) a w kolumnach zmienne (atrybuty), to możemy obliczyć za pomocą operacji macierzowych kowariancje wszystkich par zmiennych naraz korzystając ze wzoru:

$$S_X = \frac{1}{n-1}\left((X_c)^T X_c\right)$$

gdzie X_c to macierz X z wycentrowanymi kolumnami powstała po odjęciu od każdego elementu danej kolumny jej średniej (tak więc średnia każdej kolumny w X_c to 0). Macierz S_X nazywana jest *macierzą kowariancji*. Od tego, po którym X_C damy transpozycję, zależeć będzie to, czy będzie to macierz kowariancji atrybutów (zazwyczaj pożądany efekt), czy też macierz kowariancji przypadków (bardzo rzadko jeżeli wcale pożądany efekt). Jako że $\mathrm{cov}(A,B) = \mathrm{cov}(B,A)$ to w każdym wypadku będzie to macierz symetryczna (co gwarantuje nam ortogonalność jej wektorów własnych).

Współczynnik korelacji liniowej rozkładu łącznego dwóch zmiennych losowych A i B liczony jest jako:

$$r_{AB} = \frac{\mathrm{cov}(A,B)}{\sigma_A \sigma_B}$$

gdzie σ_A i σ_B to odchylenia standardowe zmiennych.

Z perspektywy PCA użycie macierzy korelacji zamiast macierzy kowariancji wstępnie normalizuje wariancje zmiennych tak, by były jednolite dla wszystkich zmiennych. Nie zawsze jest to pożądany efekt, na przykład w przypadku gdy kolumny mają bardzo podobne interpretacje (np. zysk w kolejnych miesiącach) pewna informacja jest tracona. Z drugiej strony jeżeli atrybuty mają zupełnie różną charakterystykę i znaczenie, to atrybuty z dużymi wariancjami zdominują te ze słabszymi, co zazwyczaj nie będzie pożądanym wynikiem.

2.2 Redukcja wymiarowości

PCA domyślnie nie redukuje wymiarowości danych, więc jako wynik uzyskamy macierze o następujących wymiarach:

$$\underset{(p,a)}{Y} = \underset{(p,a)}{X} \cdot \underset{(a,a)}{K}$$

gdzie p to liczba przypadków, a to liczba atrybutów, a notacja (i,j) oznacza wymiary macierzy o i wierszach i j kolumnach.

Można jednak zauważyć, że przez proces maksymalizacji wariancji (reprezentowanej przez wartości własne S_x) kolejnych nowych zmiennych, ostatnie z nich często będą mieć bardzo mały udział w zmienności danych. Można więc bez istotnej straty informacji usunąć z macierzy K kolumny odpowiadające najmniejszym wartościom własnym. Zakładając, że wzięliśmy po uwagę r kolumn, otrzymujemy:

$$\underset{(p,r)}{Y} = \underset{(p,a)}{X} \cdot \underset{(a,r)}{K}$$

2.3 Interpretacja $Y = XK$

Nasze dane X mają w wierszach przypadki a w kolumnach atrybuty. Z kolei macierz K ma w kolumnach wektory własne macierzy S_X, i są one zarówno ortogonalne jak i długości 1 – tworzą więc bazę ortonormalną. Najłatwiej wyjaśnić kolejność mnożenia XK korzystając z perspektywy „projekcyjnej", czyli że każdy przypadek jest rzutowany na kolejne nowe zmienne. Jako że kolumny K stanowią bazę ortonormalną, to możemy zapisać przypadek p_1 jako:

$$p_1 = \langle p_1, k_1\rangle k_1 + \ldots + \langle p_1, k_a\rangle k_a$$

Zadanie 2.1: Zrób Zadanie 3.1, które jest również zadaniem domowym dla nieobecnych (i tych, którzy nie zdążą go zrobić w trakcie zajęć).

Zadanie 2.2: Mamy dane X o następującej macierzy kowariancji:

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 0.7 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Czy warto użyć dla tych danych PCA? Jak by się zmieniła macierz kowariancji po użyciu PCA (bez redukcji wymiarowości)?

Zadanie 2.3: Pokaż, że macierz kowariancji S_y dla $Y = XK$ jest macierzą diagonalną z wartościami własnymi na przekątnej (czyli $S_y = L$).

Zadanie 2.4: *„Prawo zachowania wariancji"* – pokaż, że suma wariancji zmiennych nie zmienia się po zastosowaniu PCA, czyli $tr(S_y) = tr(S_x)$, gdzie tr to ślad macierzy. Konieczne może być wykorzystanie własności cykliczności śladu iloczynu macierzy: $tr(ABC) = tr(CAB) = tr(BCA)$.

Zadanie 2.5: W metodzie PCA nowe dane tworzone są jako $Y = XK$. Wyjaśnij tę kolejność X i K w kontekście projekcji.

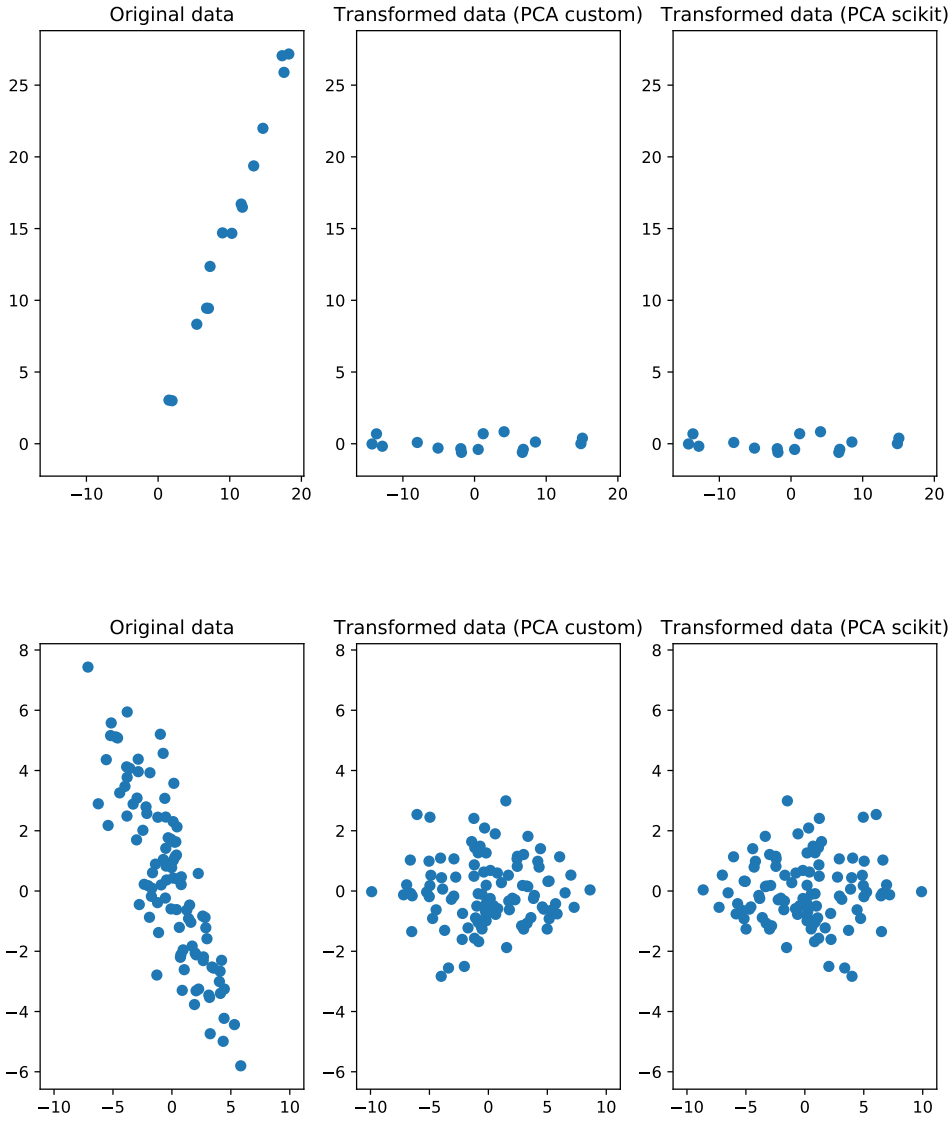
3 Zadanie domowe (5 punktów)

Zadanie 3.1: (2 pkt) Wypełnij **ręcznie** arkusz zadaniowy dostępny na stronie przedmiotu. Zadanie można oddać pokazując na zajęciach wydrukowaną i wypełnioną kartkę lub przesyłając jej skan/zdjęcie.

Zadanie 3.2: (2 pkt) Zaimplementuj algorytm PCA korzystając z szablonu na stronie i wypełniając w nim ciało funkcji `pca_manual`. Uzupełnij w niej również instrukcje print tak by wypisywały odpowiednie informacje.

Zadanie 3.3: (1 pkt) Uzupełnij w szablonie funkcję `pca_sklearn` obliczającą PCA przy użyciu biblioteki scikit-learn (`http://scikit-learn.org/stable/`). Klasa odpowiedzialna w tej bibliotece za PCA to `sklearn.decomposition.PCA`. Zapoznaj się z jej dokumentacją w internecie.

Poniżej przedstawione są przykładowe wyniki, jakie można uzyskać. Dla drugiego zbioru danych można zauważyć, że wykresy dla implementacji PCA są różne – wynika to z wybrania innych wektorów własnych w macierzy K (w szczególności: wektorów przeciwnych).



Uwaga: kolumny macierzy K to wektory własne o długości równej 1. Oznacza to, że można przemnożyć w K dowolną kolumnę (wektor własny) przez -1 i nie zmienić powyższych własności. Praktyczna konsekwencja jest taka, że różne implementacje PCA mogą dać różne wyniki zależnie od wybranych wektorów własnych (jednak wariancje zmiennych zawsze będą takie same). Widać to na rysunku powyżej, gdzie strzałka jest odwrócona – wynikło to właśnie z tego, że wystąpiła różnica w znakach w kolumnach K.

Zadanie 3.4: (0 pkt) Uczyń zajęcia lepszymi poprzez wypełnienie ankiety z zajęć.