

Metody redukcji wymiarowości danych

Piotr Lipiński

wykład z eksploracji danych

Metoda składowych głównych

- ▶ Metoda składowych głównych (ang. Principal Component Analysis, PCA) służy do badania rzeczywistej wymiarowości danych.
- ▶ Niech $D = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\} \subset \mathbb{R}^d$ będzie analizowanym zbiorem danych złożonym z N wektorów d -wymiarowych.
- ▶ Celem PCA jest przekształcenie d -wymiarowych wektorów $\mathbf{x}_i \in \mathbb{R}^d$ w M -wymiarowe wektory $\mathbf{z}_i \in \mathbb{R}^M$, dla pewnego $M < d$.

Metoda składowych głównych

- ▶ Niech $\mathcal{U} = \{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_d\}$ będzie pewnym zbiorem d ortonormalnych wektorów w \mathbb{R}^d . Wówczas każdy wektor $\mathbf{x} \in \mathbb{R}^d$ może być przedstawiony jako

$$\mathbf{x} = \sum_{k=1}^d z_k \mathbf{u}_k, \quad \text{gdzie } z_k = \mathbf{u}_k^T \mathbf{x}.$$

- ▶ Odpowiada to zmianie układu współrzędnych: każdy wektor \mathbf{x} reprezentowany przez współrzędne x_1, x_2, \dots, x_d w standardowym układzie współrzędnych, może być reprezentowany przez współrzędne z_1, z_2, \dots, z_d w układzie współrzędnych wyznaczonym przez bazę ortonormalną \mathcal{U} .
- ▶ Przypuśćmy, że zamiast wszystkich współrzędnych używalibyśmy tylko wybranych M z nich, a pozostałe zastąpilibyśmy stałymi, otrzymując pewne przybliżenie oryginalnego wektora \mathbf{x} przez wektor

$$\tilde{\mathbf{x}} = \sum_{k=1}^M z_k \mathbf{u}_k + \sum_{k=M+1}^d b_k \mathbf{u}_k, \quad \text{gdzie } b_k \text{ to pewne stałe.}$$

Metoda składowych głównych

- ▶ Błąd przybliżenia to

$$\mathbf{x} - \tilde{\mathbf{x}} = \sum_{k=M+1}^d (z_k - b_k) \mathbf{u}_k.$$

- ▶ Łączny błąd przybliżenia dla całego zbioru danych D można zatem określić jako

$$F_{err} = \frac{1}{2} \sum_{i=1}^N \|\mathbf{x}_i - \tilde{\mathbf{x}}_i\|^2 = \frac{1}{2} \sum_{i=1}^N \sum_{k=M+1}^d (z_{ik} - b_k)^2.$$

Metoda składowych głównych

- ▶ F_{err} można traktować jako funkcję b_k i wyznaczyć b_k dla których F_{err} przyjmuje wartość najniższą. Różniczkując F_{err} po b_k i przyrównując do zera, otrzymujemy

$$b_k = \frac{1}{N} \sum_{i=1}^N z_{ik} = \frac{1}{N} \sum_{i=1}^N \mathbf{u}_k^T \mathbf{x}_i = \mathbf{u}_k^T \bar{\mathbf{x}},$$

gdzie

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i.$$

- ▶ Wracając do funkcji błędu, można więc ją zapisać jako

$$F_{err} = \frac{1}{2} \sum_{k=M+1}^d \sum_{i=1}^N (\mathbf{u}_k^T (\mathbf{x}_i - \bar{\mathbf{x}}))^2 = \frac{1}{2} \sum_{k=M+1}^d \mathbf{u}_k^T \mathbf{\Sigma} \mathbf{u}_k,$$

gdzie $\mathbf{\Sigma}$ to macierz kowariancji zbioru danych D , tzn.

$$\mathbf{\Sigma} = \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T.$$

Metoda składowych głównych

- ▶ Można pokazać, że funkcja F_{err} osiąga minimum dla wektorów \mathbf{u}_k spełniających

$$\mathbf{\Sigma} \mathbf{u}_k = \lambda \mathbf{u}_k,$$

czyli dla wektorów własnych macierzy kowariancji $\mathbf{\Sigma}$.

- ▶ Funkcję błędu można wówczas zapisać jako

$$F_{err} = \frac{1}{2} \sum_{k=M+1}^d \lambda_k,$$

zatem należy wybrać M wektorów własnych odpowiadających największym wartościom własnym macierzy kowariancji.

- ▶ Wektory własne \mathbf{u}_k zwane są składowymi głównymi zbioru danych D .

Metoda mnożników Lagrange'a

- ▶ Niech $F : \mathbb{R}^d \rightarrow \mathbf{R}$ (funkcja celu).
- ▶ Niech $h : \mathbb{R}^d \rightarrow \mathbf{R}$ (funkcja ograniczeń).
- ▶ Jeśli $\mathbf{x}_0 \in \mathbb{R}^d$ jest minimum lub maksimum funkcji F na zbiorze $\{\mathbf{x} \in \mathbb{R}^d : h(\mathbf{x}) = c\} \subset \mathbb{R}^d$, gdzie c jest pewną stałą, to istnieje $\boldsymbol{\lambda} \in \mathbb{R}^d$ taka że

$$\nabla F(\mathbf{x}_0) = \boldsymbol{\lambda} \nabla h(\mathbf{x}_0), \quad (1)$$

lub równoważnie:

$$\frac{\partial F}{\partial x_i}(\mathbf{x}_0) = \lambda_i \frac{\partial h}{\partial x_i}(\mathbf{x}_0), \quad \text{dla każdego } i = 1, 2, \dots, d, \quad (2)$$

gdzie $\mathbf{x} = (x_1, x_2, \dots, x_d)^T$ oraz $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_d)^T$.

- ▶ Jeśli dla $\mathbf{x}_0 \in \mathbb{R}^d$ spełniony jest warunek (1) lub (2), to \mathbf{x}_0 może, ale nie musi, być minimum lub maksimum funkcji F na rozważanym zbiorze (powyższy warunek jest konieczny, ale nie jest dostateczny).

Dowód

- ▶ Będziemy wykorzystywać metodę mnożników Lagrange'a.
- ▶ Funkcja błędu to

$$F_{err} = \frac{1}{2} \sum_{k=M+1}^d \mathbf{u}_k^T \mathbf{\Sigma} \mathbf{u}_k,$$

gdzie $\mathbf{\Sigma}$ to macierz kowariancji zbioru danych D .

- ▶ Wektory \mathbf{u}_i są ortonormalne. Mamy więc ograniczenia $\mathbf{u}_i^T \mathbf{u}_j = \delta_{ij}$.

Dowód

- ▶ Wykorzystamy metodę mnożników Lagrange'a w trochę innej notacji. Niech

$$L = \frac{1}{2} \sum_{k=M+1}^d \mathbf{u}_k^T \mathbf{\Sigma} \mathbf{u}_k - \frac{1}{2} \sum_{i=M+1}^d \sum_{j=M+1}^d \mu_{ij} (\mathbf{u}_i^T \mathbf{u}_j - \delta_{ij}).$$

- ▶ Używając notacji macierzowej

$$L = \frac{1}{2} \text{Tr}(\mathbf{U}^T \mathbf{\Sigma} \mathbf{U}) - \frac{1}{2} \text{Tr}(\mathbf{M}(\mathbf{U}^T \mathbf{U} - \mathbf{I})).$$

- ▶ Przyprowadzimy ∇L względem \mathbf{U} do zera otrzymując

$$0 = (\mathbf{\Sigma} + \mathbf{\Sigma}^T) \mathbf{U} - \mathbf{U}(\mathbf{M} + \mathbf{M}^T).$$

Dowód

- ▶ Macierz Σ jest symetryczna. Można przyjąć, że macierz \mathbf{M} jest także symetryczna, bo macierz \mathbf{U} jest symetryczna.
- ▶ Zatem poprzednie równanie sprowadza się do

$$\Sigma \mathbf{U} = \mathbf{U} \mathbf{M}.$$

- ▶ Z założenia (i ograniczeń), macierz \mathbf{U} ma ortonormalne kolumny, więc

$$\mathbf{U}^T \Sigma \mathbf{U} = \mathbf{M}.$$

- ▶ Zatem jednym z możliwych rozwiązań jest macierz wektorów i wartości własnych. Nie musi to jednak być jedyne rozwiązanie.

Dowód

- ▶ Rozpatrzmy równanie dla wektorów i wartości własnych macierzy \mathbf{M} :

$$\mathbf{M}\mathbf{V} = \mathbf{V}\mathbf{D}$$

- ▶ Ponieważ macierz \mathbf{M} jest symetryczna, więc macierz wektorów własnych \mathbf{V} może być tak wybrana, że jej kolumny są ortonormalne, stąd

$$\mathbf{D} = \mathbf{V}^T \mathbf{M} \mathbf{V}$$

- ▶ Rozpisując dalej

$$\mathbf{D} = \mathbf{V}^T \mathbf{U}^T \Sigma \mathbf{U} \mathbf{V} = (\mathbf{U} \mathbf{V})^T \Sigma (\mathbf{U} \mathbf{V}) = \tilde{\mathbf{U}}^T \Sigma \tilde{\mathbf{U}},$$

gdzie $\tilde{\mathbf{U}} = \mathbf{U} \mathbf{V}$.

Dowód

- ▶ Ponieważ $\mathbf{V}^T \mathbf{V} = \mathbf{I}$, bo \mathbf{V} jest ortonormalne, mamy

$$\mathbf{U} = \tilde{\mathbf{U}} \mathbf{V}^T,$$

więc inne rozwiązania niż macierz wektorów własnych macierzy kowariancji to tylko jej przekształcenia ortonormalne.