

SOM vs MDS

Problemy SOM

SOM bardzo popularna metoda wizualizacji

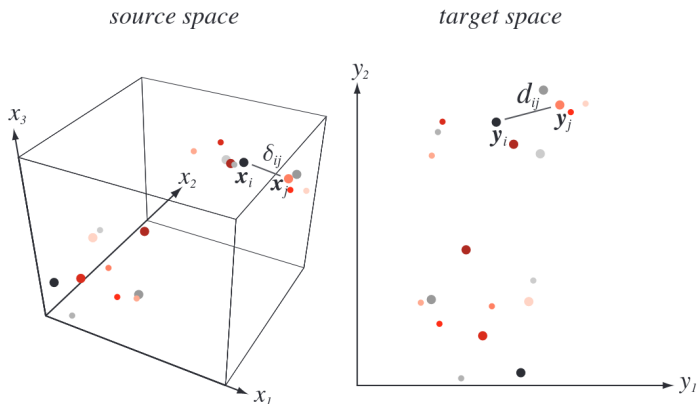
- nie gwarantuje wiarygodnej aproksymacji gęstości danych
- błąd kwantyzacji przydatny przy ocenie jakości odwzorowania ale nie mówi nic o jakości wizualizacji

MDS, Multi-Dimensional Scaling (Thorton 1954, Kruskal 1964), zwane także mapowaniem Sammona (Sammon 1964)

- redukcja wymiarowości z zachowaniem relacji odległości
- funkcja stresu: miara zniekształceń topograficznych
- zazwyczaj redukcja do 2D, 3D w celu wizualizacji danych

Idea MDS

Szukamy mapowania $\mathbf{x} \rightarrow \mathbf{y} = M(\mathbf{x})$ minimalizującego pewną funkcję kosztu opartą o odległości δ_{ij} oraz d_{ij}



Rys: Duda, Hart, Pattern Classification

Algorytm MDS

Przestrzeń cech $\mathbf{x}_i \in \mathbb{R}^d$, $i = 1, 2, \dots, n$

Odległość w przestrzeni cech

$$\delta_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$$

Przestrzeń docelowa $\mathbf{y}_i \in \mathbb{R}^K$, gdzie $K < N$

Odległość w przestrzeni docelowej

$$d_{ij} = \|\mathbf{y}_i - \mathbf{y}_j\|$$

Można użyć dowolnej metryki - najczęściej metryka Euklidesowa
W przypadku $K = 2$

$$E = \sum_{i < j}^n (\delta_{ij} - d_{ij})^2 = \sum_{i < j}^n \left(\delta_{ij} - \sqrt{(y_{i1} - y_{j1})^2 + (y_{i2} - y_{j2})^2} \right)^2$$

Funkcja E zależy od $2n$ parametrów $\mathbf{y}_i$

Jednak 3 współrzędne są redundantne: $\mathbf{y}_1 = (0, 0)^T$ wybór

środką układu wsp., $\mathbf{y}_2 = (0, y_{22})$ wybór orientacji układu wsp.

Miary zniekształceń

$$E(\mathbf{r}; \mathbf{w}) = \sum_{i < j}^n w_{ij} (\delta_{ij} - d_{ij})^2 \geq 0$$

wagi w_{ij} mogą zależeć od odległości, np. maleć wykładniczo ze wzrostem δ_{ij}

Czynnik normalizacyjny

$$E(\mathbf{r}) = \frac{\sum_{i < j}^n (\delta_{ij} - d_{ij})^2}{\sum_{i < j}^n \delta_{ij}^2 + \sum_{i < j}^n d_{ij}^2} \in [0, 1]$$

Stres wynosi 0, gdy idealne dopasowanie ($\delta_{ij} = d_{ij}$), wartość 1 np. gdy wszystkie $d_{ij} = 0$

Miary MDS

- Stres (Kruskal) błąd absolutny, mogą dominować duże odległości, zachowuje ogólną strukturę klastrów (skupisk)

$$E_1 = \sum_{i < j}^n (\delta_{ij} - d_{ij})^2 \geq 0$$

- Odwzorowanie Sammona, wpływ dużych odległości δ_{ij} jest zredukowany

$$E_2 = \frac{1}{c} \sum_{i < j}^n \frac{(\delta_{ij} - d_{ij})^2}{\delta_{ij}} \geq 0$$

gdzie

$$c = \frac{1}{\sum_{i < j} \delta_{ij}}$$

Miary MDS

- Błąd względny, wszystkie skale odległości traktowane są jednakowo

$$E_3 = \sum_{i=1}^n (1 - d_{ij}/\delta_{ij})^2 \geq 0$$

- Współczynnik alienacji (Guttman-Lingoes) – podobny do błędu względnego, trudniejszy do minimalizacji

$$E_4 = \sum_{i < j}^n (1 - \delta_{ij}/d_{ij})^2 \geq 0$$

Optymalizacja stresu

Sammon zaproponował optymalizację za pomocą zmodyfikowanej metody Newtona

$$y_{pq}(k+1) = y_{pq}(k) - \eta \Delta_{pq}(k)$$

gdzie

$$\Delta_{pq}(k) = \frac{\partial E}{\partial y_{pq}} \bigg/ \left| \frac{\partial^2 E}{\partial y_{pq}^2} \right|$$

składowa gradientu

$$\frac{\partial E}{\partial y_{pq}} = -\frac{2}{c} \sum_{j=1, j \neq p}^n \left(\frac{\delta_{pj} - d_{pj}}{d_{pj} \delta_{pj}} \right) (y_{pq} - y_{jq})$$

diagonalne składowe Hessianu

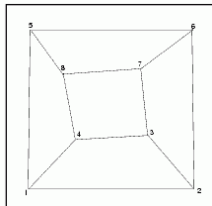
$$\frac{\partial^2 E}{\partial y_{pq}^2} = -\frac{2}{c} \sum_{j=1, j \neq p}^n \frac{1}{\delta_{pj} d_{pj}} \left[(\delta_{pj} - d_{pj}) - \frac{(y_{pq} - y_{jq})^2}{d_{pj}} \left(1 + \frac{\delta_{pj} - d_{pj}}{d_{pj}} \right) \right]$$

MDS i SOM

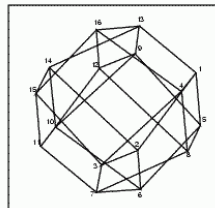
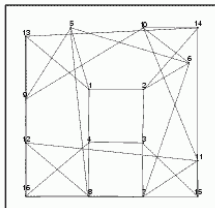
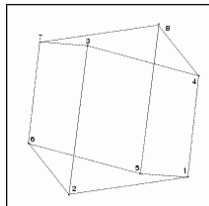
- w MDS brak funkcji mapującej $X \rightarrow Y$
- dla MDS dodanie nowego punktu wymaga ponownej optymalizacji
- SOM używany także w klasyfikacji
- MDS i SOM wrażliwe na szum w danych (nieistotne cechy)
- wyniki obu metod są uzależnione od punktu startowego, każda optymalizacja może generować odmienne wyniki

Wizualizacja: hiperkostka 2D i 4D

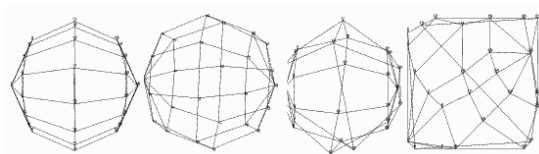
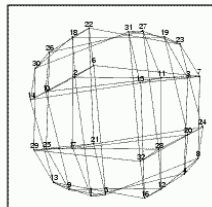
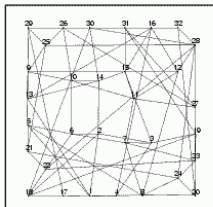
SOM warstwa 2D, 20x20



MDS

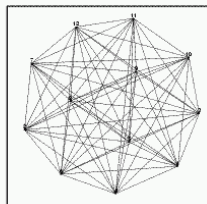
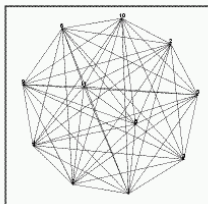
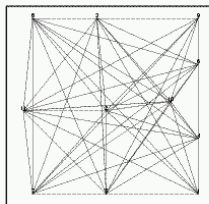
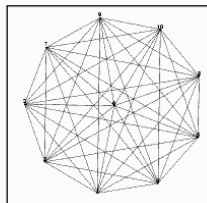
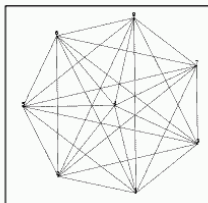
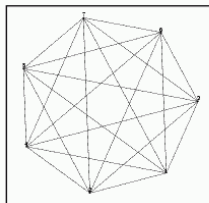


Wizualizacja: hiperkostka 5D + sfera w 3D

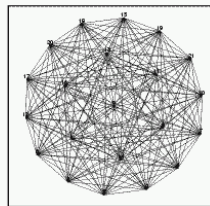
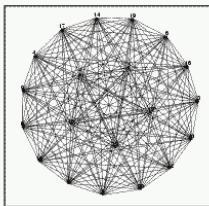
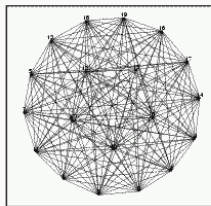
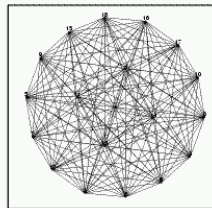
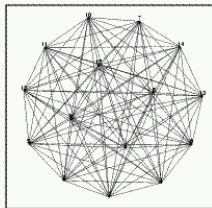
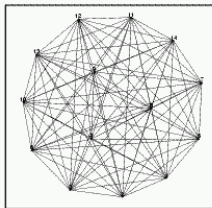


The two-dimensional representations of the 26 points on the sphere obtained by minimization of S , E , A , and by **SOFM** (left to right) with a 20×20 neurons map.

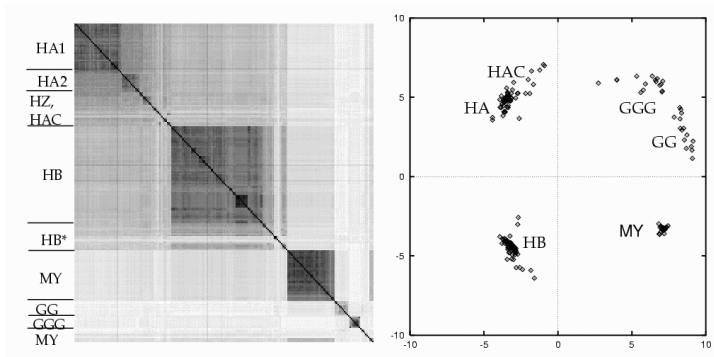
Wizualizacja: sympleks 6-11



Wizualizacja: sympleks 15-20

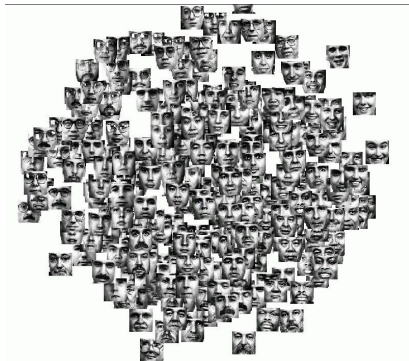


Sekwencje rodziny białek



sekwencja białek rodziny Globin, macierz podobieństwa, MDS odkrywa strukturę danych (Klock & Buhmann 1997).

Podobieństwo twarzy



300 twarzy (from Klock & Buhmann 1997).

Mapy semantyczne

- próba uchwycenia sensu słów i pojęć poprzez własności oraz relacji semantycznych
- Przykład: 8 ptaków i 8 ssaków
dove, hen, duck, goose, owl, hawk, eagle,
fox, dog, wolf, cat, tiger, lion, horse, zebra, cow.
- każde pojęcie opisane 13 cechami binarnymi
size is: small, medium large;
has 2 legs, 4 legs,
has hair, hoofs, mane, feathers;
likes to: hunt, run, fly, swim.
- utwórz zdania opisujące zwierzęta za pomocą wybranych cech
Horse is big, has 4 legs, mane, hair, hoofs, likes to run.

Mapy semantyczne

animal		d	o	h	u	s	w	a	g	f	d	w	c	i	l	h	z
		e	n	k	e	l	k	e	x	g	f	l	a	g	e	r	e
is	small	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	0
	medium	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
	big	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
has	2 legs	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
	4 legs	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
	hair	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
	hooves	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
	mane	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0
	feathers	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
likes to	hunt	0	0	0	0	1	1	1	1	0	1	1	1	1	1	0	0
	run	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	0
	fly	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0
	swim	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0

Mapy semantyczne

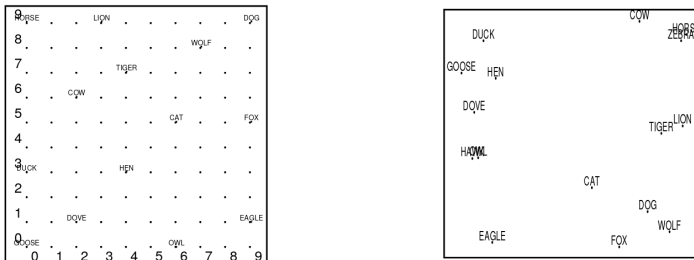


FIG. 1. The two-dimensional representations of the 13-dimensional semantic data obtained by SOM (left) with a 10 x 10 neurons map, a training of 10000 cycles, with final stress of 0.25, and the MDS (right) with final stress of 0.20 after 10 iterations.

SOM, Ritter and Kohonen 1989,
MDS, Naud & Duch (1996).

- Naturalna klasyfikacja zwierząt zachowująca ich podobieństwo
 - ssaki odseparowane od ptaków
 - małe zwierzęta od dużych
 - drapieżników od roślinożerców
- podobieństwo zwierząt jest powiązane z odległością na mapie

Samoorganizacja

Uczenie hebbowskie

Sieci samoorganizujące się

Uczenie konkurencyjne

- reguła Kohonena
- konkurencja WTA, WTM
- kwantyzacja wektorowa, SOM

Uczenie hebbowskie

- reguła Hebba, reguła Oja
- asocjacje i korelacje między sygnałami
- dekompozycja składowych głównych PCA (Principal Component Analysis)
- dekompozycja na składowe niezależne ICA (Independent Component Analysis) , separacja źródeł BSS (Bling Source Separation), nielinowe PCA, ...

Reguła Hebba (1949)

„Neurons, that fire together, wire together.”

Jeżeli neuron A jest cyklicznie pobudzany przez neuron B, to staje się on jeszcze bardziej czuły na pobudzenie tego neuronu

- synchroniczne pobudzanie neuronów A i B wzmacnia połączenie synaptyczne między nimi
- jeżeli neurony nie są pobudzane jednocześnie (asynchronicznie) to połączenie jest osłabiane

Zmiana wagi połączenia pomiędzy neuronami A i B

$$w_{AB}(k+1) = w_{AB}(k) + \eta y_A(k) y_B(k)$$

Reguła Hebba

Dla neuronu liniowego

$$y = \mathbf{w}^T \mathbf{x}$$

reguła Hebba przybiera postać

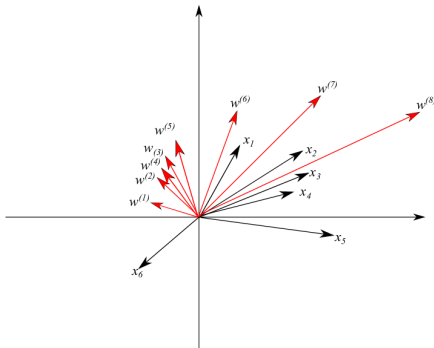
$$\Delta w_i = \eta y x_i$$

Kierunek $\mathbf{w}$ podążą w kierunku $\mathbf{x}$, decydujący wpływ na wartość końcową będą miały sygnały zgrupowane (rozciągające się) wzdłuż znalezionej kierunku.

Uczenie hebbowskie prowadzi do wyznaczenia miary podobieństwa między sygnałem wejściowym x_i a rozkładem gęstości prawdopodobieństwa $p(x)$ sygnału.

Interpretacja geometryczna reguły Hebba

Wektor wag przemieszcza się w kierunku wyznaczonym przez środek masy wektorów treningowych (maleje kąt pomiędzy kierunkiem największej wariancji danych)



Trening prowadzi do nieograniczonego wzrostu wag (norma $\|\mathbf{w}\|$ rośnie w każdym kroku)

Modyfikacje reguły Hebba

Zapobieganie nieograniczonemu wzrostowi wag (2 strategie):

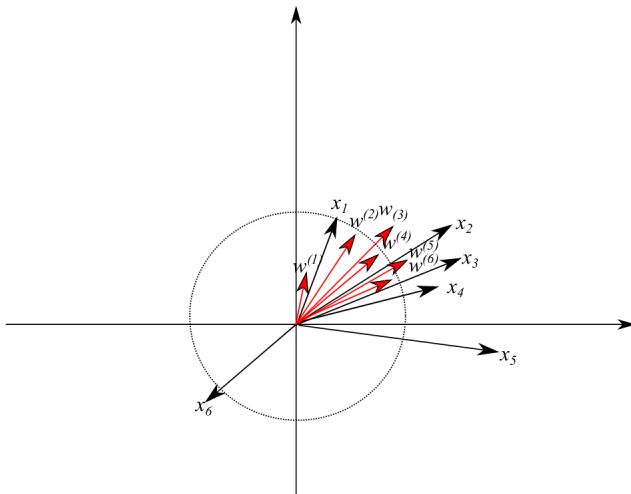
- normalizacja wag po każdym kroku uczenia
- dodanie czynnika normalizującego wagi, Oja (1982)

Reguła Oji

$$\Delta w_i = \eta y (x_i - y w_i)$$

- dodanie wyrazu zanikania wagi proporcjonalnego do y^2
- adaptacja prowadzi do unormowania wag $\|\mathbf{w}\| = 1$
- wektor wag zbliża się do wektora własnego macierzy kowariancji $E[\mathbf{X}\mathbf{X}^T]$ o największej wartości własnej
- wektor wag leży wzdłuż kierunku maksymalizującego wartość oczekiwaną y^2 (wariancja wartości wyjściowej neuronu)

Interpretacja geometryczna reguły Oji



Własności reguły Oji

Dla wyuczonego wektora wag wartość oczekiwana

$$E[\Delta \mathbf{w}] = 0$$

stąd

$$\begin{aligned} E[\Delta \mathbf{w}] &= \eta E[y(\mathbf{x} - y\mathbf{x})] = \eta E[(\mathbf{w}^T \mathbf{x})\mathbf{x} - y^2 \mathbf{w}] \\ &= \eta (E[\mathbf{x}\mathbf{x}^T] \mathbf{w} - E[y^2] \mathbf{w}) = 0 \end{aligned}$$

Macierz kowariancji $E[\mathbf{x}\mathbf{x}^T] = \mathbf{R}_x$

Reguła Oji prowadzi do wyznaczenia wektora własnego $\mathbf{R}_x$

$$\mathbf{R}_x \mathbf{w} = \lambda \mathbf{w}$$

Analiza składowych głównych PCA

Transformacja liniowa

$$\mathbf{y} = \mathbf{W}\mathbf{x}$$

gdzie

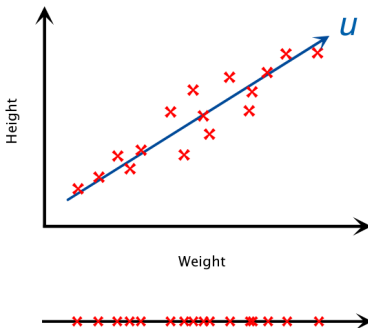
$$\mathbf{x} \in \mathbb{R}^N \quad \mathbf{y} \in \mathbb{R}^K \quad \mathbf{W} \in \mathbb{R}^{K \times N} \quad K < N$$

Redukcja wymiarowości z zachowaniem maksymalnej informacji zawartych w danych.

W PCA kierunek zawierający najwięcej informacji to kierunek o największej wariancji

Projekcja danych na wektor wag

Projekcja danych 2D na kierunek o największej wariancji



Rys: Boedecker, Machine Learning Summer 2015

PCA

Niech $\mathbf{X} \in \mathbb{R}^N$ wektor losowy o zerowej wartości oczekiwanej, wówczas wartość oczekiwana macierzy kowariancji

$$R_x = E[\mathbf{X}\mathbf{X}^T]$$

dla skończonej próbki o liczebności p

$$R_x \approx \frac{1}{p} \sum_{i=1}^p \mathbf{x}_i \mathbf{x}_i^T$$

PCA

Wartości własne λ_i oraz wektory własne $\mathbf{w}_i$ macierzy kowariancji

$$\mathbf{R}_x \mathbf{w}_i = \lambda_i \mathbf{w}_i \quad i = 1, \dots, N$$

R jest symetryczna i nieujemna $\Rightarrow$ wartości własne rzeczywiste i nieujemne

Niech

$$\lambda_1 > \lambda_2 > \dots > \lambda_N \geq 0$$

Ograniczając do K pierwszych składowych

$$\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_k]^T$$

otrzymujemy transformację PCA

$$\mathbf{y} = \mathbf{W}\mathbf{x}$$

PCA

Wariancja wzdłuż i -tego kierunku

$$\text{Var}(\mathbf{w}_i^T \mathbf{x}) = E[\|\mathbf{w}_i^T \mathbf{x}\|^2] = E[(\mathbf{w}_i^T \mathbf{R}_x \mathbf{w}_i)] = \lambda_i$$

Celem PCA jest znalezienie wektorów $\mathbf{w}_1, \dots, \mathbf{w}_K$, które maksymalizują wariancję przy zachowaniu ortogonalności

$$\mathbf{w}_i^T \mathbf{w}_j = 0 \quad \text{dla} \quad j \geq i$$

$$\mathbf{w}_i^T \mathbf{w}_i = 1$$

Algorytm PCA

1. Wyśrodkowanie danych dla każdej cechy j

$$x_{ij} \leftarrow x_{ij} - \frac{1}{n} \sum_{j=1}^n x_{ij}$$

2. Wyznaczenie macierzy kowariancji

$$\mathbf{R}_x = \frac{1}{n} \mathbf{X} \mathbf{X}^T$$

3. Diagonalizacja za pomocą rozkładu SVD

$$\mathbf{S} = \mathbf{W}^{-1} \mathbf{R}_x \mathbf{W}$$

$\mathbf{W}$ zawiera wektory własne, $\mathbf{S}$ wartości własne na diagonalu

4. Projektcja danych na kierunki wektorów własnych odpowiadającym K największym wartościom własnym

$$\mathbf{y}_i = \mathbf{W}_K \mathbf{x}_i$$

Rekonstrukcja

Rekonstrukcja obrazu wejściowego

$$\hat{\mathbf{x}} = \mathbf{W}^T \mathbf{y}$$

PCA minimalizuje błąd rekonstrukcji

$$E_r = E[\|\mathbf{x} - \hat{\mathbf{x}}\|^2]$$

dla k składowych PCA

$$E_r = \sum_{i=k+1}^N \lambda_i$$

Minimalizacja E_r odpowiada maksymalizacji $\sum_{i=1}^K \lambda_i$

PCA

- Gdy zmienne są skorelowane to znajomość tylko części zmiennych pozwala określić pozostałe
- Względny wkład poszczególnych składowych

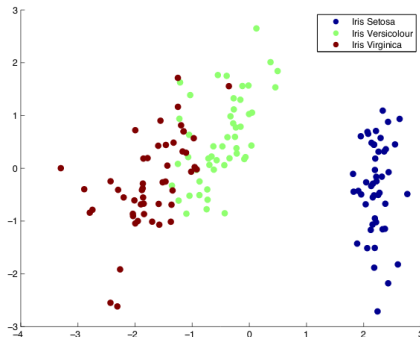
$$m_i = \frac{\lambda_i}{\sum \lambda_j}$$

- Standardowe metody wyznaczania wektorów własnych macierzy $\mathbf{R}_x$ dla dużych wymiarów są zbyt złożone

Metody adaptacyjne wyznaczania składowych głównych (np. reguła Oji)

- nie wymagają obliczania macierzy $\mathbf{R}_{xx}$
- mogą być stosowane w metodach on-line, gdy nie mamy dostępu do wszystkich danych jednocześnie

Przykład: PCA na danych Iris

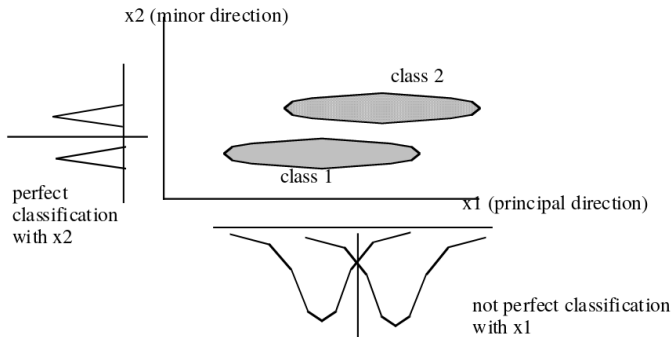


$$\lambda_1 = 2.8914 \quad \lambda_2 = 0.9151 \quad \lambda_3 = 0.1464 \quad \lambda_4 = 0.0205$$

PCA często używane do przetworzenia danych przed treningiem algorytmów uczenia maszynowego

Rys: Boedecker, Machine Learning Summer 2015

PCA i klasyfikacja



Projekcja PCA nie gwarantuje uzyskania najlepszej dyskryminacji

MDS i PCA

- MDS wynik zależny od punktu startowego, stosowany wielokrotny start
- PCA może zostać użyte jako punkt startowy dla MDS. Nieistotne składowe są usuwane i nie mają wpływu na wizualizację MDS.
- PCA jest linowe, nie zachowuje odległości pomiędzy obiektami
- PCA wymaga utworzenia macierzy korelacji $d \times d$, koszt $O(nd^2)$, diagonalizacja tej macierzy $O(d^3)$
- MDS wymaga obliczenia macierzy odległości $O(n^2d)$ i minimalizacji $2n - 3$ parametrów

Trening pierwszego kierunku

Estymator pierwszego kierunku

$$y_1 = \mathbf{w}^T \mathbf{x} = \sum_{j=0}^N w_{1j} x_j$$

za pomocą reguły Oja

$$\mathbf{w}_1(k+1) = \mathbf{w}_1(k) + \eta(k) y_1(k) (\mathbf{x}(k) - \mathbf{w}_1(k) y_1(k))$$

Dobór $\eta(k)$ jest istotny - powinien maleć z czasem, np.:

$$\eta(k) = \frac{\eta(0)}{k^\gamma} \quad \text{gdzie} \quad 0.5 \leq \gamma \leq 1$$

Uogólniony algorytm hebbowski (GHA)

Jednowarstwowa sieć liniowa (Sanger, Oja, 1989)

$$y_i = \mathbf{w}_i^T \mathbf{x}$$

Reguła uczenia (Sanger)

$$\Delta \mathbf{w}_i = \eta y_i \left(\mathbf{x}_i - \sum_{k=1}^i y_k \mathbf{w}_k \right)$$

oznaczając

$$\hat{\mathbf{x}}_i(k) = \mathbf{x}_i - \sum_{k=1}^{i-1} y_k \mathbf{w}_k$$

otrzymujemy regułę Oji dla przeskalowanych wejść $\hat{\mathbf{x}}_i$

$$\Delta \mathbf{w}_i = \eta y_i (\hat{\mathbf{x}}_i - \mathbf{w}_i y_i)$$

Wartości wag $\mathbf{w}_i$ kolejnych neuronów uzyskują wartości kolejnych wartości własnych macierzy kowariancji $\mathbf{C}$

Uczenie anty-hebbowskie

Reguła anty-Hebbowska

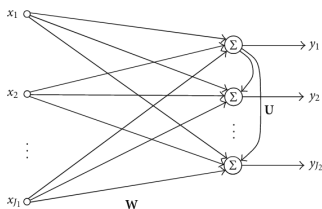
$$\Delta w_{ij} = -\eta x_i y_j$$

- znak minus sprawia, że trening prowadzi do znalezienia kierunku **minimalizującego** wariancję wyjścia
- uczenie jednostki liniowej prowadzi do zerowania wyjścia - trening dąży do znalezienia projekcji danych do punktu
- jeżeli nie ma kierunku, wzdłuż którego wariancja wynosi 0 to wagi zbiegają do 0

APEX

Adaptive Principal Component Extraction (Diamantaras 1990)

Dla danych wektorów $\mathbf{w}_1, \dots, \mathbf{w}_{i-1}$ realizujących kierunki PCA sieć iteracyjnie wyznacza i -tą składową PCA



$$\mathbf{y} = \mathbf{W}^T \mathbf{x}$$

$$y_i = \mathbf{w}_i^T \mathbf{x} + \mathbf{u}_i^T \mathbf{y}$$

$\mathbf{y} = (y_1, \dots, y_{i-1})^T$ wyjścia pierwszych $i - 1$ neuronów

$\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_{i-1}]^T$ macierz wag pierwszych $i - 1$ neuronów

$\mathbf{u}_i = [u_{1i}, u_{2i}, \dots, u_{(i-1)i}]^T$ wagi połączeń wewnątrz warstwy

Rys: Qiu, et.al., *Neural Network Implementations for PCA and Its Extensions*

Aktualizacja wag APEX

Aktualizacja wag $\mathbf{w}_i$ za pomocą reguły Oji

$$\Delta \mathbf{w}_i = \eta(n) y_i(n) [\mathbf{x}(n) - y_i(n) \mathbf{w}_i]$$

Połączenia między-warstwowe u_{ij} uczone reguła anty-hebbowską odpowiedzialną za ortogonalizację y_i względem poprzednich składowych

$$\Delta \mathbf{u}_i = -\eta(n) y_i(n) [\mathbf{y}(n) + y_i(n) \mathbf{u}_i]$$

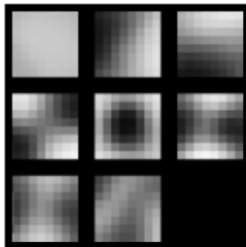
Zastosowanie PCA

- Redukcja wymiarowości
- Wizualizacja danych $nD \rightarrow 2D, 3D$
- Usuwanie szumu
- Usuwanie korelacji z danych
- Kompresja sygnału, np. kompresja obrazów
- Dane po transformacji PCA często używane do treningu algorytmów uczenia maszynowego

Kompresja danych za pomocą PCA

Wektory treningowe to ramki obrazu ($8 \times 8 = 64$ współrzędne)

Przykład: obraz 512×512 z 8 bitami na piksel



ordering
of the neurons:

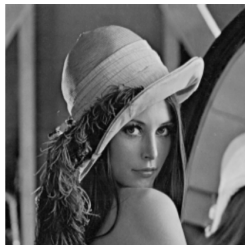
neuron 0	neuron 1	neuron 2
neuron 3	neuron 4	neuron 5
neuron 6	neuron 7	(unused)

Neurony w kolejności zawierającej najwięcej informacji potrzebnej do rekonstrukcji

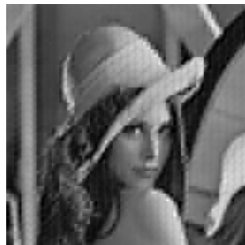
- neuron 0: średni poziom szarości
- neuron 1 i 2: gradient obrazu
- neurony 3-5: pochodna drugiego rzędu obrazu

Przykład: Kompresja obrazu za pomocą APEX

- a) oryginał
- b) 4 PC
- c) 8 PC
- d) 12 PC



(a)



(b)



(c)



(d)

Rys: Krose, Smagat, An Introduction to Neural Networks, 1996