Redukce dimenzionality

Jan Šimbera

Přírodovědecká fakulta UK

10. listopadu 2016

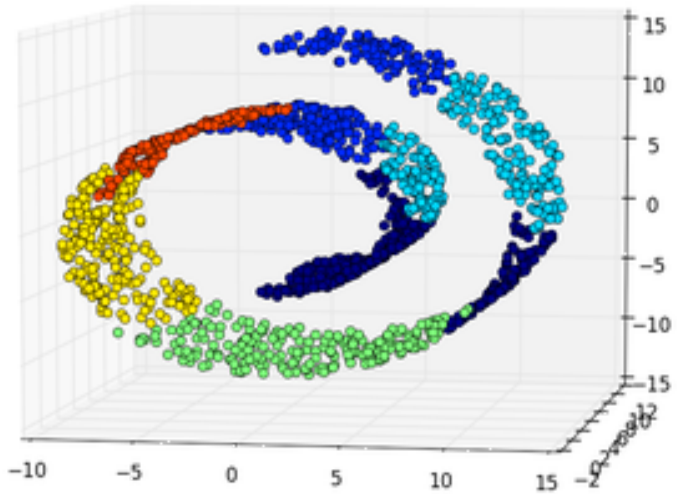
Obsah

1. Úvod
2. Konvexní metody
3. Generalizované konvexní metody
4. Nekonvexní metody
5. Porovnání
6. Experiment

Redukce dimenzionality

- zmenšit příznakovou matici $\mathbf{X}$ $n \times D$ na $\mathbf{Y}$ $n \times d$, $d < D$
- cíl:
  - zmenšení objemu zpracovávaných dat (D často více než stovky)
  - zachování relevantní informace a odstranění šumu
  - nalezení skrytých kauzalit a vysvětlujících proměnných

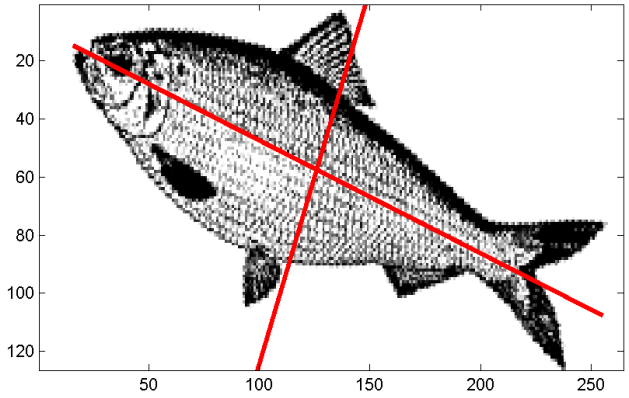
- předpoklad: data tvoří d-dimenzionální útvar *(manifold)* vyjádřený v D dimenzích
  - snažíme se jej „rozbalit“
  - problém se šumem!



Analýza hlavních komponent (PCA)

- lineární technika – dekorelace dle Pearsonova koeficientu
- rotace příznakového prostoru dle vlastních vektorů korelační matice

$$\mathrm{cov}(\mathbf{X})\mathbf{M} = \lambda\mathbf{M}$$



Výhody

- robustnost, rychlost pro malé D
- možnost volby d

Nevýhody

- nelineární vztahy
- zachovává globální, ne lokální geometrii

- $\mathcal{O}(D^3)$, lze reformulovat na $\mathcal{O}(n^3)$ – *classical scaling*

Isomap

- zachovává *geodetické vzdálenosti* mezi dvojicemi bodů
- graf k-sousedství – váhy hran euklidovské, vzdálenosti nejkratší cestou
  - k nutno zvolit
- minimalizace vzdáleností všech dvojic bodů:

$$\phi(\mathbf{Y}) = \sum_{ij} \left(d_{ij}^2 - |\mathbf{y_i} - \mathbf{y_j}|^2 \right)$$

Výhody

- zachytí nelineární vztahy
- jednodušší, rozšířená

Nevýhody

- *short-circuiting*
- zachovává globální, ne lokální geometrii
- neumí díry, nekonvexní útvary

Kernel PCA

- analogie SVM pro redukci dimenzionality
- vlastní vektory se počítají na jádrové matici $\mathbf{K}k = (\kappa(\mathbf{x}_i, \mathbf{x}_j))$
- *jádrová funkce* $\kappa(\mathbf{x}_i, \mathbf{x}_j)$ – např. gaussovské jádro

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{|\mathbf{x}_i - \mathbf{x}_j|^2}{2\sigma^2}\right)$$

Výhody

- zachytí nelineární vztahy
- lepší odolnost vůči šumu

Nevýhody

- zachovává globální, ne lokální geometrii
- jádrová matice $\mathcal{O}(n^2)$

Difuzní mapy

- varianta Isomap zohledňující lépe lokální vzdálenosti
- vzdálenostní metrika založená na markovských procesech

$$w_{ij} = \exp\left(-\frac{|\mathbf{x}_i - \mathbf{x}_j|^2}{2\sigma^2}\right) \qquad \text{(váha hran)}$$

$$\mathbf{P} = \left(p_{ij}^{(1)}\right) = \left(\frac{w_{ij}}{\sum_k w_{ik}}\right) \qquad \text{(pravděpodobnosti přechodu)}$$

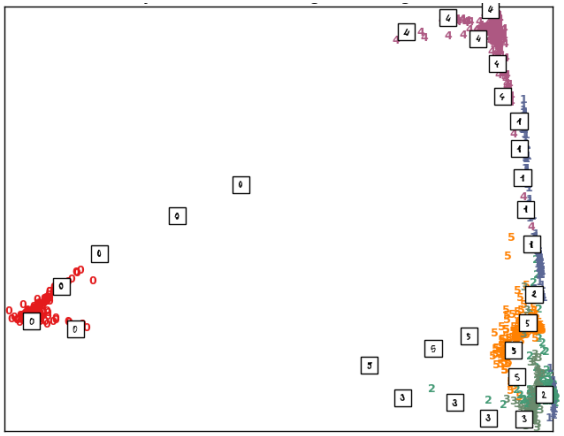
$$\psi_i = \frac{\sum_j p_{ij}^{(1)}}{\sum_j \sum_k p_{jk}^{(1)}} \qquad \text{(význam vyšším hustotám)}$$

$$D^{(t)}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_k \frac{\left(p_{ik}^{(t)} - p_{jk}^{(t)}\right)^2}{\psi_k}}$$

- průměrné

Local Linear Embedding

- vylepšení Isomapu na zachování lokální geometrie
- bod se vyjádří jako lineární kombinace svých sousedů
  - transformace se snaží zachovat kombinační váhy
  - předpokládá lokálně lineární chování



$$\phi(\mathbf{Y}) = \sum_i |\mathbf{y}_i - \sum_{j=1}^{k} w_{ij}\mathbf{y}_{ij}|^2$$

za podmínky $|\mathbf{y}^{(k)}|^2 = 1 \, \forall k$ – nutná pro vyloučení triviálního $\mathbf{Y} = \mathbf{0}$

- problémy – hvězdicové uspořádání, díry

Sammonovo zobrazení

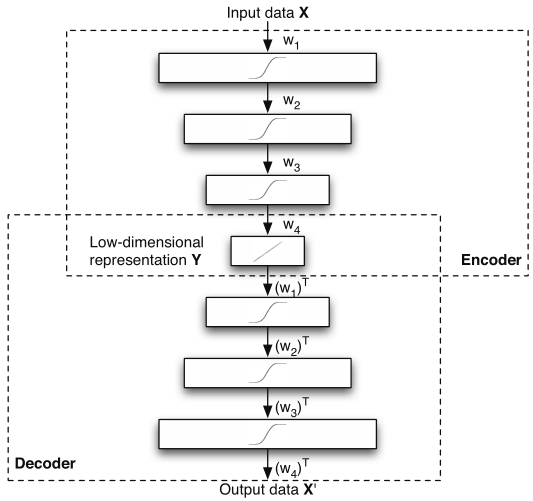
- snaha zachovat blízké vzdálenosti – vážení přímo v nákladové funkci:

$$\phi(\mathbf{Y}) = \frac{1}{\sum_{ij} d_{ij}} \sum_{i \neq j} \frac{(|\mathbf{x}_i - \mathbf{x}_j| - |\mathbf{y}_i - \mathbf{y}_j|)^2}{|\mathbf{x}_i - \mathbf{x}_j|}$$

- optimalizace pseudonewtonovsky
- vážení až příliš důsledné

Autoenkodér

- neuronová síť s „hrdlem“
  - MLP se sigmoidami mimo hrdlo
- učí se identita
  - náročné (mnoho stupňů volnosti, pomalá konvergence)
  - předtrénování po vrstvách RBM
- pro DR se vezme část po hrdlo



- vhodné pro zašuměná data
- parametrické vyjádření – lze dosazovat další body *(out-of-sample extension)*
- časově i výpočetně náročné

Samoorganizující mapy (SOM)

- RD spojená s klasifikací – omezený počet cílových bodů
- neurony $\hat{\mathbf{x}}_i(\mathbf{y}_i)$ uspořádané v pravidelné mřížce Γ s dimenzí $d-1$ vrstva
- iterativní učení hodnot $\hat{\mathbf{x}}_i$ v neuronech z dat $\mathbf{x}_k$:

$$\hat{\mathbf{x}}_W(t) = \arg\min_{\mathbf{y}_i \in \Gamma} |\hat{\mathbf{x}}_i(t) - \mathbf{x}(t)| \qquad \text{(výběr vítěze)}$$

$$\hat{\mathbf{x}}_i(t+1) = \hat{\mathbf{x}}_i(t) + \alpha(t) \mathrm{e}^{-\frac{|\mathbf{y}_i - \mathbf{y}_W|^2}{2\sigma^2(t)}} [\hat{\mathbf{x}}_i(t) - \mathbf{x}(t)] \qquad \text{(úprava hodnot)}$$

- zejména pro klasifikační úlohy, výkon velmi závislý na mřížce a vzdálenostní metrice

t-SNE

- t-distributed Stochastic Neighbor Embedding
- převod shodnosti bodů na společné pravděpodobnosti
- minimalizace Kullback-Leibler divergence mezi původními (N) a redukovanými (t) daty
- vhodné pro vizualizaci

Výhody

- dobrý výkon na reálných datech
- funguje na nespojité útvary

Nevýhody

- výpočetní náročnost $\mathcal{O}(n^2)$ (předřazuje se PCA)
- $d \leq 3$

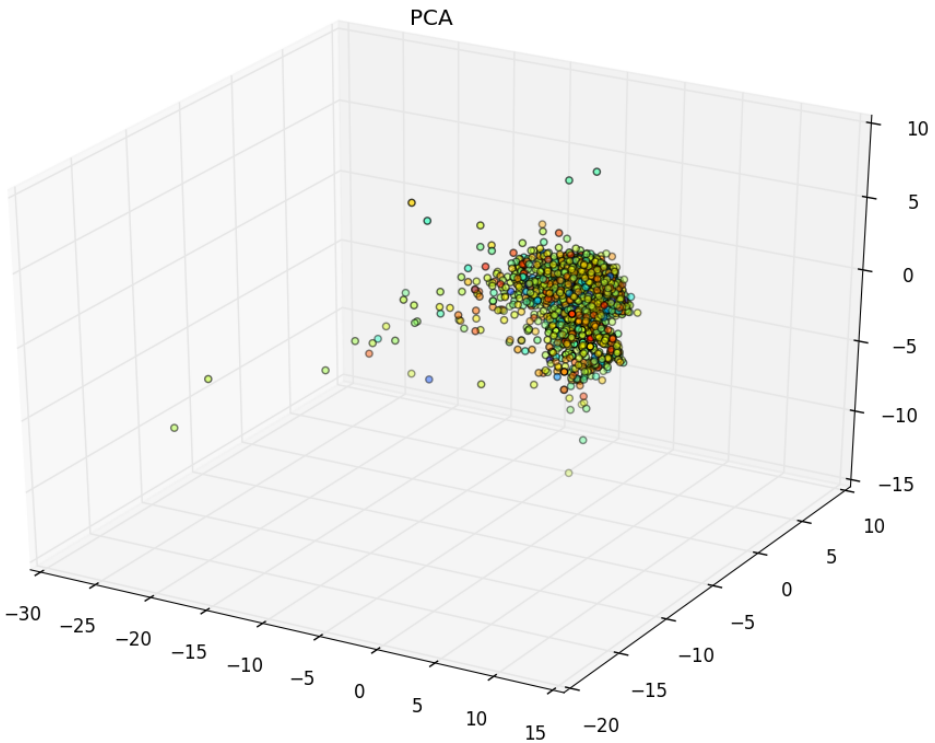
Porovnání

- van der Maaten et al. (2009)
- konvexní metody umí dobře umělé datasety (swiss roll)
- nekonvexní metody jsou lepší na reálných vysokodimenzionálních datech s vyšším množstvím šumu

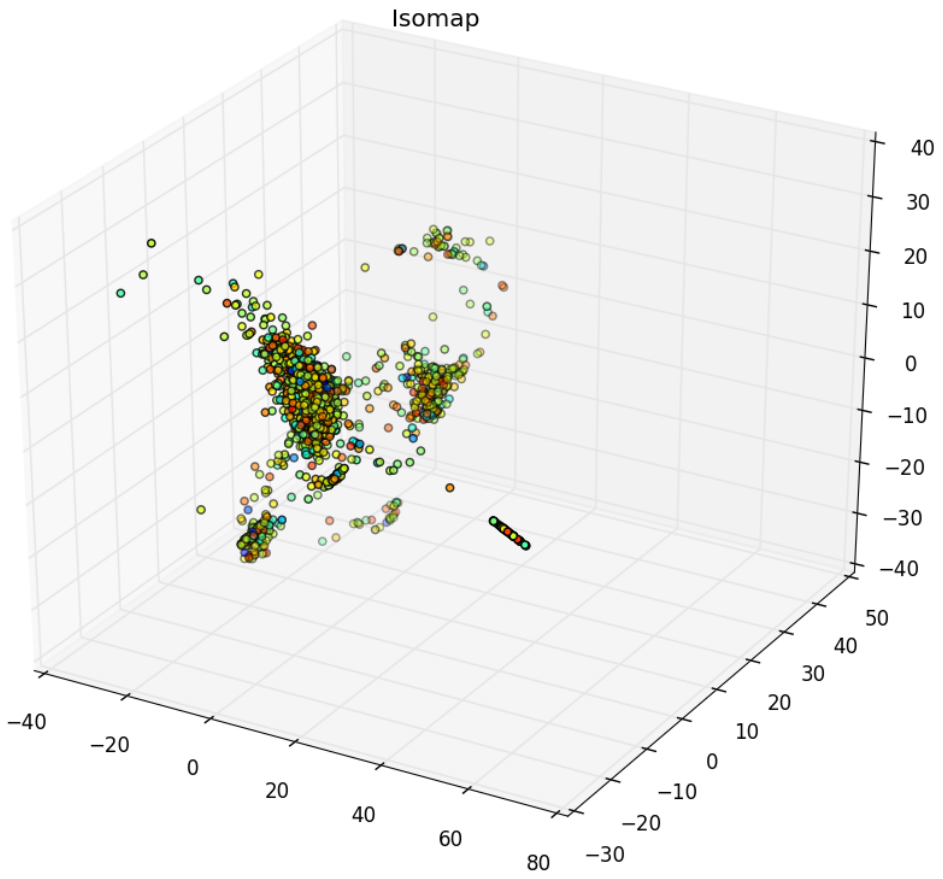
Experiment

- implementace ze `scikit-learn`
- PCA, Isomap, LLE a t-SNE
- charakteristiky prostředí pro 3000 území v Praze (73D) – land cover, zastavěnost, množství bodů zájmu, hustota silniční sítě
- redukce do 3D
- odhalí se korelace s hustotou zalidnění?
- časy:
  - PCA 26 ms
  - Isomap 7,8 s
  - LLE 4,3 s
  - t-SNE 79 s

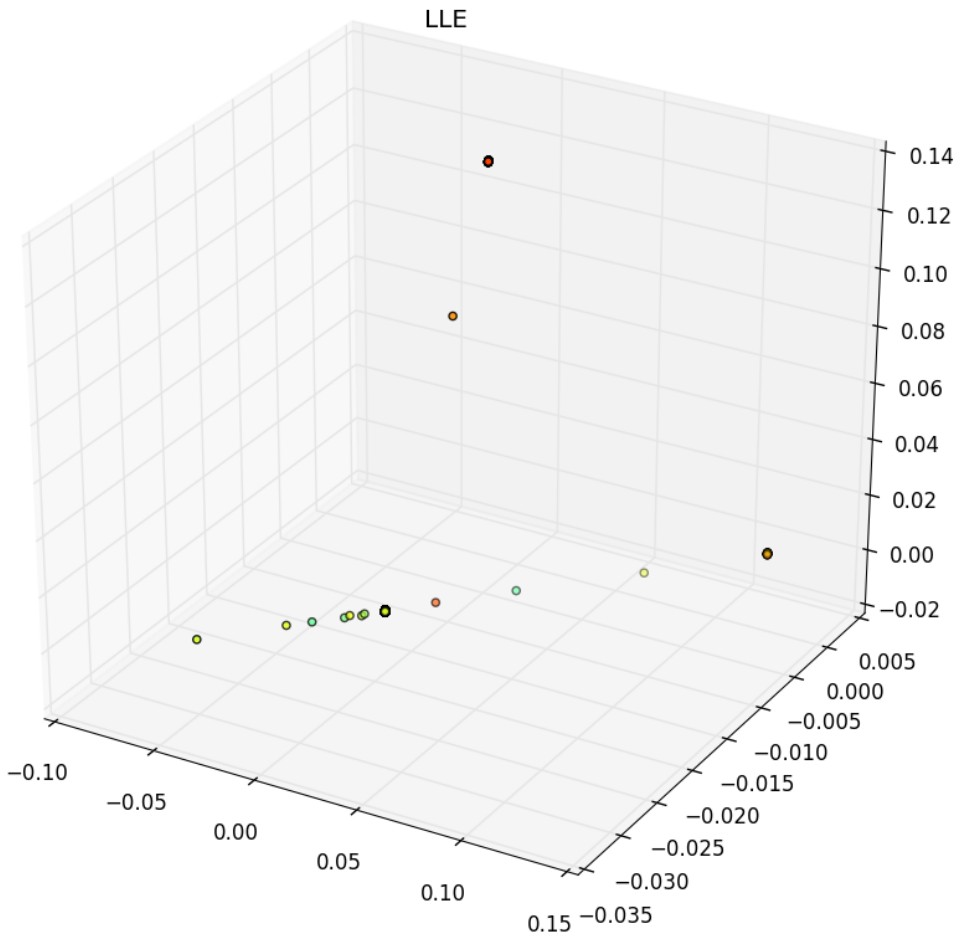
PCA



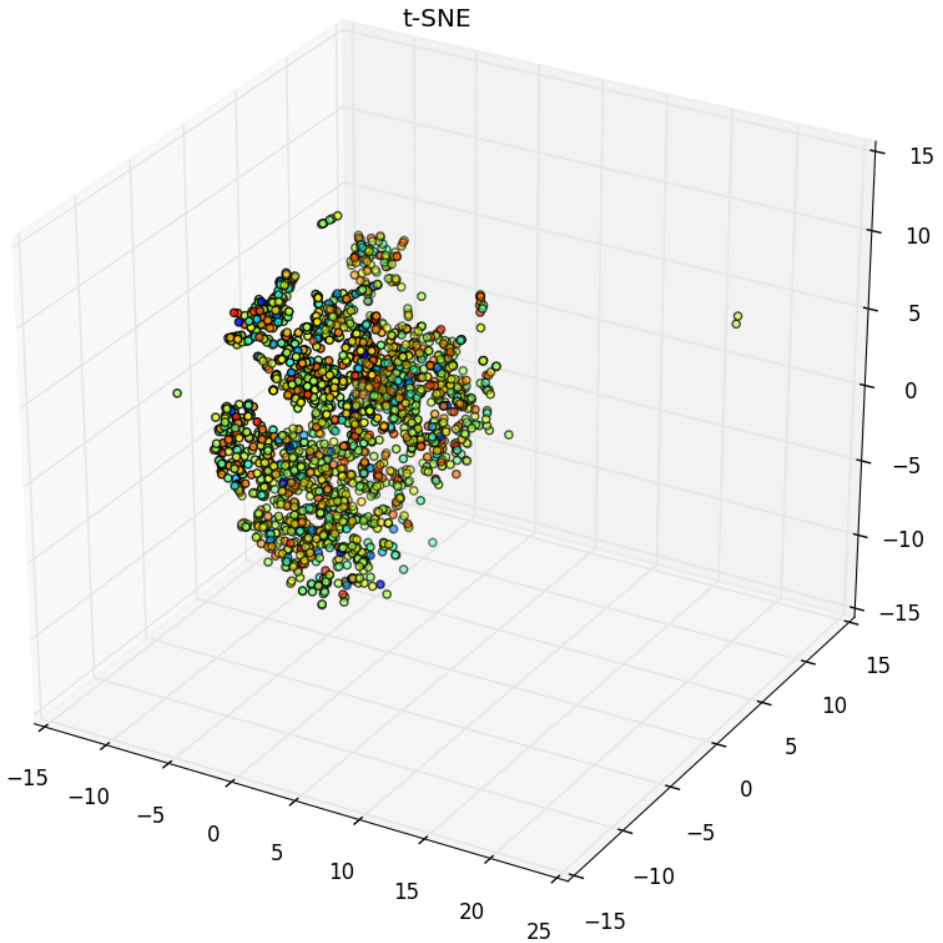
Isomap


Isomap
40
30
20
10
0
−10
−20
−30
−40
50
40
30
20
10
0
−10
−20
−30
−40
−20
0
20
40
60
80

LLE


LLE
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
−0.02
0.005
0.000
−0.005
−0.010
−0.015
−0.020
−0.025
−0.030
−0.035
−0.10
−0.05
0.00
0.05
0.10
0.15

t-SNE



Literatura

- LESKOVEC, J., RAJARAMAN, A., ULLMAN, J. (2014): *Mining of Massive Datasets, Chapter 11: Dimensionality Reduction.* Cambridge University Press, s. 405–426.
- VAN DER MAATEN, L. J. P., POSTMA, E., VAN DEN HERIK, J. (2009): *Dimensionality Reduction: A Comparative Review.* Technická zpráva, Tilburg Centre for Creative Computing, Tilburg.
- VAN DER MAATEN, L. J. P., HINTON, G. E. (2008): *Visualizing High-Dimensional Data Using t-SNE.* Journal of Machine Learning Research, 9, s. 2579–2605.
- YIN, H. (2007): *Nonlinear Dimensionality Reduction and Data Visualization: A Review.* International Journal of Automation and Computing, 4, č. 3, s. 294–303.