Димензиона редукција

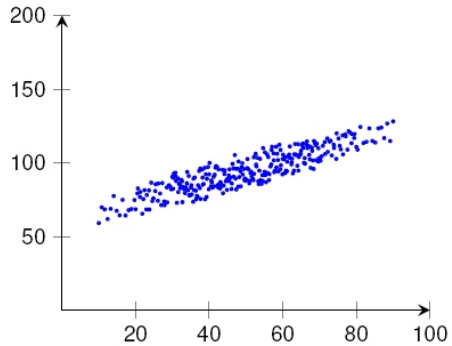
22. oktobar 2022.

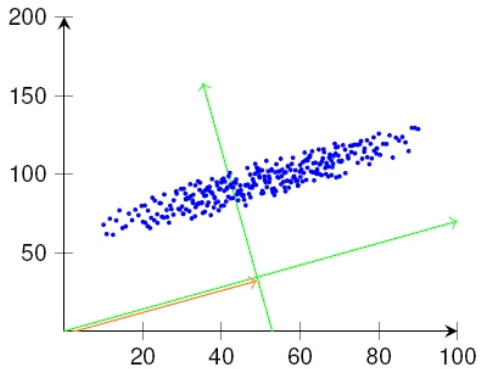
Избор карактеристика

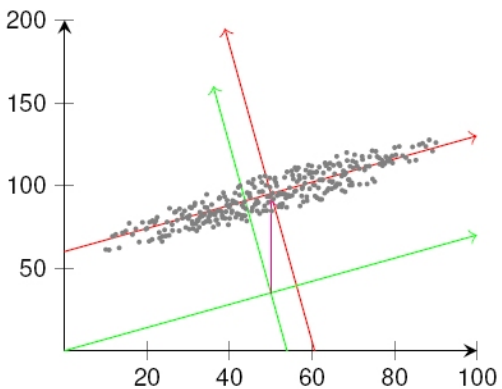
1. Један од начина за смањење димензионалности
2. Елиминација редундантних карактеристика
3. Елиминација ирелевантних карактеристика
4. Развијен је велики број техника, поготову за класификацију
5. Формирање нових атрибута
6. Пресликавање у нови простор (нпр. Фуријеова анализа, таласићи)
7. Уводна литература: Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar: Introduction to Data Mining, 2nd ed, стр. 76-83, Додатак Б у првом издању књиге

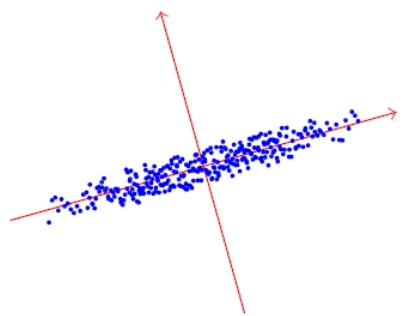
Редукција помоћу ротације оса

Основна идеја









Редукција помоћу ротације оса

- Аутоматско уклањање координатних оса помоћу ротације?
- *PCA* (Principal Component Analysis)
- *SVD* (Singular Value Decomposition)

Principal Component Analysis

- Смањење броја димензија података
- Налажење образаца у подацима велике димензионалности
- Визуелизација података велике димензионалности

Principal Component Analysis (наставак)

- Основна идеја: ротација података у систем са осама где је највећи број варијанси покривен најмањим бројем димензија
- Нови систем са осама зависи од корелације између атрибута
- *PCA* се (најчешће) примењује после одузимања средње вредности од сваке тачке

Principal Component Analysis (наставак)

- За матрицу података D реда $m \times n$ може да се формира матрица коваријанси C са елементима $c_{ij} = cov(d_{*i}, d_{*j})$
  (c_{ij} је коваријанса i-те и j-те колоне података)
- Коваријанса је мера како се атрибути мењају у пару. Ако је $i = j$ тада је коваријанса једнака варијанси атрибута
- Ако се матрица D претходно припреми тако да је средња вредност сваког од атрибута једнака 0, тада је $C = D^T D$

Principal Component Analysis (наставак)

Циљ *PCA* је налажење трансформације података за коју важи

1. Сваки пар новодобијених атрибута има коваријансу 0
2. Атрибути су уређени (у опадајућем редоследу) у односу на величину варијансе која је покривена од стране атрибута
3. Захтева се ортогоналност између атрибута, тако да сваки наредни атрибут покрива што је могуће већи број преосталих варијанси

Principal Component Analysis (наставак)

Трансформација се врши употребом сопствених вредности матрице коваријанси. Нека су

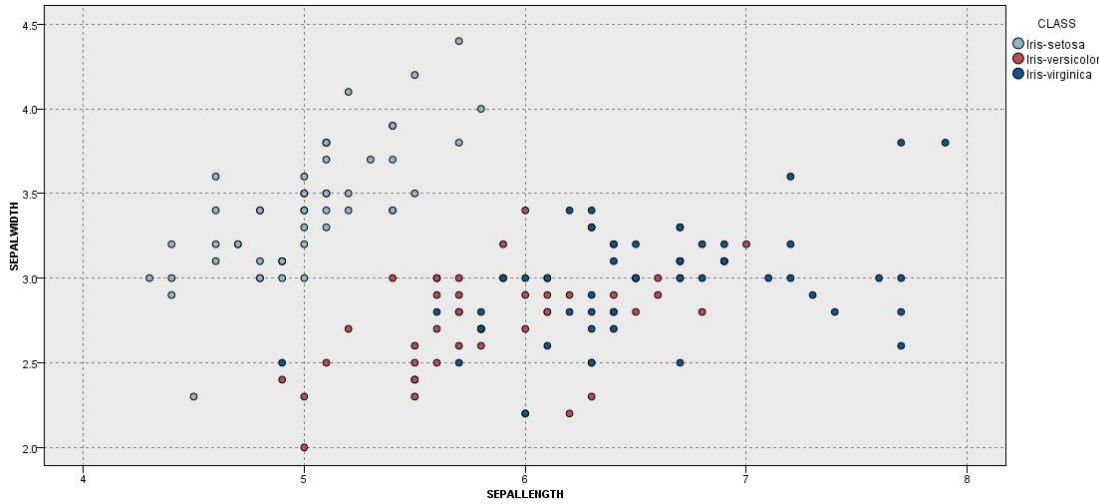
1. λ_i (ненегативне) сопствене вредности од C уређене у редоследу $\lambda_1 \geq \lambda_2 \geq ...\lambda_{m-1} \geq \lambda_m$
2. $U = [u_1, ..., u_n]$ матрица сопствених вектора од C уређена тако да i-ти вектор одговара i-тој највећој сопственој вредности
3. Нека је матрица D претходно припремљена тако да је средња вредност сваког од атрибута једнака 0

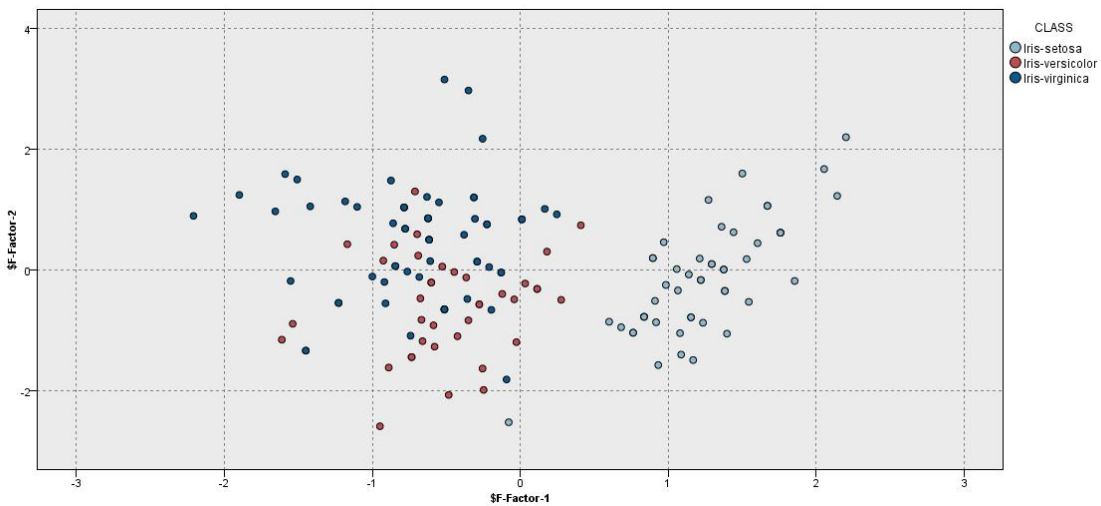
Principal Component Analysis (наставак)

Тада важи:

1. матрица $D' = DU$ је тражена трансформација матрице података
2. нови атрибут је линеарна комбинација старих атрибута; тежина линеарне комбинације i-тог атрибута су компоненте i-тог сопственог вектора
3. варијанса новог i-тог атрибута је λ_i; збир варијанси оригиналних је једнак збиру варијанси нових атрибута
4. нови атрибути се називају *главне компоненте* ; први нови атрибут је прва главна компонента, итд.

Principal Component Analysis (наставак)





Димензиона редукција - литература

Литература:

1. Abdi H., Williams L. - Principal Component Analysis
2. Shlens J. - A tutorial on Principal Component Analysis (2005)
3. B.K. Tripathy, Anveshrithaa Sundareswaran etc. - Unsupervised Learning Approaches for Dimensionality Reduction and Data Visualization
4. Yang, Xin-She - Introduction to algorithms for data mining and machine learning (стр. 98-115)
5. M. Garzon et al. - Dimensionality reduction in data science
6. V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos - Feature selection for high-dimensional data

Softver

- Python - sklearn
- R
- Matlab
- SPSS Modeler
- SETDIR (https://setdir.engineering.iastate.edu/doku.php?id=download), *A_software_framework_for_data_dimensionality_reduction*
  Test:
  .\ Windows \ setdr -num_objects 200 -higher_dim 3 -inputfile d:/ip2/SETDIR_1.7.04_user/workspace/Arc200.dat
- Matlab Toolbox for Dimensionality Reduction (https://lvdmaaten.github.io/drtoolbox/)
- dimRed: A Framework for Dimensionality Reduction (https://cran.r-project.org/web/packages/dimRed/vignettes/dimensionality-reduction.pdf)