MEIC 2018/19 Ciência de Dados

Lab 8b: Dimensionality Reduction

1. **Load** and analyze high-dimensional **data**
   a. Load real-valued datasets *prostate* (and transpose it) and *colon*
   b. Load the *decathlon2* data with active individuals (rows 1 to 23), supplementary individuals (rows 24 to 27), active variables (columns 1 to 10) and class variables (columns 11 to 13). E.g. in R:

      data(decathlon2) from library("factoextra")
      decathlon2.active <- decathlon2[1:23, 1:10]
      head(decathlon2.active)

2. **Principal component analysis**
   a. obtain the eigenvalues
   b. analyze how much data variance is explained by each components
   c. retrieve the feature composition of each component
   d. analyze the number of selected components when varying the allowed amount of noise
   e. fix a number of components and apply the inverse transformation to reconstruct the original data and analyze its properties

3. **Visualize** reduced data spaces
   a. plot the new data space resulting data from previous PCA analyzes
      i. use 2-dimensional (or 3-dimensional) plots by projecting data into 2 (or 3) dimensions
      ii. for *colon* and *decathlon2* data: color data points according to their nominal output or apply scales based on numeric output
   b. plot the graph of explanatory components
   c. biplot (a) and (b) information

4. **PCA variants**

   For each dataset apply:
   a. kernel PCA: an extension of PCA to achieve non-linear dimensionality reduction through the use of kernels
   b. sparse PCA: a variant of PCA to extract the set of non-sparse components that best reconstruct the data
   c. apply the inverse transformation to reconstruct the original data and analyze its properties

5. Others. Considering the *colon* dataset:
    a. compare PCA with **supervised** forms of dimensionality reduction; Suggestion: apply Linear Discriminant Analysis (LDA) on the given datasets to identify attributes that account for the most variance between classes
    b. apply **random projections** to efficiently reduce the dimensionality

6. **Evaluate** dimensionality reduction procedures
   Select some of the previous dimensionality reduction settings:
    a. apply and analyze clustering performance before and after using:
        i. unsupervised clustering metrics
        ii. supervised clustering metrics
    b. analyze the visual plots
    c. analyze the improvements from applying classifiers and regression methods before and after the dimensionality reduction

Notes

Bonus: show the results from **exercise 2.d** and **2.f** to have your mark

Resources (packages):

- R: stats, kernlab, ggbiplot, factoextra, dplyr
- Python: sklearn.feature_selection