TÉCNICO LISBOA

MEIC 2018/19 Ciência de Dados

Lab 7: Clustering

Part I

1. **Load, generate, preprocess** and **plot data**
    a. Generate datasets
        i. analyze and run generateCData.py file to produce and save datasets with two attributes and varying properties
        ii. if you are using R, save and open the produced data or use MixSim package generate similar datasets
    b. Load the following numeric datasets
        i. *iris* (remove the class attribute for cluster analysis)
        ii. *glass* (remove the class attribute for cluster analysis)
        iii. *colon* (remove the class attribute for cluster analysis)
    c. Center and scale data (normalization)
    d. Plot heatmaps to explore each dataset
        i. reorder and color elements
        ii. infer dendograms
        iii. customize heatmaps

2. **Distance matrices**
    Compute pairwise similarity matrices for the observations in the *iris*, *glass* and *colon* datasets using:
    a. Euclidean distance
    b. Pearson correlation

3. **Clustering algorithms**
    Apply clustering algorithms to the generated datasets (with k=2 and k=3) and to the *iris* dataset, including:
    a. **Hierarchical** and agglomerative clustering
        i. parameterize the linkage criterion
            1. min and max
            2. complete
            3. average
            4. Ward
        ii. perform bootstrap analysis (pvclust package in R) to assess the uncertainty by calculating cluster p-values via multiscale bootstrap resampling

b. ***k*-Means** with parameterized distance metrics:
    i. Euclidean
    ii. others
c. **DBSCAN** (Python)
d. **fuzzy clustering** (R)
    i. analyze the memberships of each observation to each cluster
e. advanced
    i. **Gaussian mixtures** in Python
    ii. Q-clustering (maximum diameter) and Clara in R
    iii. affinity propagation and spectral clustering in Python

f. repeat the previous steps for the *colon* dataset and discuss the challenges of clustering high-dimensional data, including:
    i. inter- and intra-similarity
    ii. efficiency

4. Clustering **mixed data**
a. Load the *college* data – composed of both continuous attributes (including acceptance rate, out of school tuition, number of new students enrolled) and categorical attributes (whether a college is elite or public/private) – and analyze its properties
b. Apply k-medoids clustering algorithm with k=3 and adequate numeric and categoric distance measures
c. Compare the produced clusters in the presence and absence of categorical data

Part II

5. Visualize clustering solutions
Considering some of aforementioned clustering settings:
a. plot clusters and centroids
b. visualize clusters in the 2D or 3D space (e.g. scatterplot3d in R)
Note: if you are using data with more than 3 attributes, you can use manifold learning techniques to retrieve 2D/3D embeddings.
c. plot observations distinctly classified by clustering algorithms

6. Number of clusters
Select one of the aforementioned clustering settings:
a. Run a *k*-dependent clustering algorithm with a varying number of clusters
b. Analyze the error of the produced clustering solutions using the silhouette coefficient
c. Hypothesize what is the true number of clusters based on the produced curve (number of clusters x error)

7. **Evaluating clustering**
   Select the clustering solutions produced for the *iris* dataset and evaluate their quality:
   a. In the absence of class information using the silhouette coefficient
   b. In the presence of class information using:
      i. adjusted Rand index
      ii. sum of squared errors
      iii. mutual Information based scores (in Python)
      iv. homogeneity, completeness and V-measure (in Python)

Part III

8. **Pencil-and-paper** exercise. Consider the following dataset:

	y1	y2	y3	y4	y5
g1	0.365	0.912	-0.463	A	0
g2	0.971	-1.571	0.750	D	1
g3	-0.730	1.334	-0.986	A	1
g4	-0.182	0.268	-1.303	C	1
g5	0.080	0.980	0.676	B	0
g6	-0.244	0.117	1.652	B	0
g7	-1.357	1.070	0.850	C	0
g8	1.278	0.135	0.437	C	1
g9	0.411	-1.032	1.383	D	0
g10	-0.687	-0.088	-1.177	B	0

   a. Apply the 2 iterations of the k-medoids algorithms on centered-scaled data with k=2 when considering:
      i. Euclidean distance on numeric attributes and binary matches on categoric attributes
      ii. centroid given by means and modes
   b. Apply agglomerative clustering using average distances
   c. Compare the produced clustering solutions with regards to their:
      i. cohesion and separation
      ii. silhouette coefficient

Bonus: show the results from **exercise 7.a** and **7.b.i** to have your mark

Resources (packages):

R packages for clustering: cluster, pvclust, kmeds
R packages with visual facilities: gplots, pheatmap, scatterplot3d
Other relevant R packages: stats, caret, clv, klaR
Python: sklearn.cluster