LEIC – ALAMEDA

2017/18

Sistemas de Apoio à Decisão

Lab 3 – Clustering

Goals:

- K-means, EM
- Clustering validation
- Normalization
- PCA

1. Load the *iris* data, and apply the k-means algorithm, with 2 clusters and ignoring the *class* attribute.
    a. What are the centroids for each cluster?
    b. Knowing that there are 3 different types of plants, validate the results achieved visually. (In weka, right-click the result in the '*Result list'*, and select the '*Visualize cluster assignments'*. In this new context, you may save those assignments to an .arff file, by clicking in '*Save'*)
    c. How good is the clustering?

2. Run the algorithm for different number of clusters from 2 to 10.
    a. What is the best number of clusters? Validate it by visualizing the data.
    b. How good is the clustering and does it change with the number of clusters?

3. Load the *glass* data, and apply the k-means algorithm with 2 clusters, ignoring the *pep* attribute.
    a. What is the resulting SSE?
    b. Normalize numeric attributes and compare the results with the previous task.

4. Load the *cereals.csv* data, and apply the k-means algorithm with 2 clusters.
    a. Record the resulting SSE for the original dataset.
    b. Again but by ignoring the name attribute.

c. Apply PCA without centering the data. Record the resulting MSE.
d. Reload the original dataset, ignore the name attribute and apply PCA by centering the data. Record the resulting MSE.
e. How do the different results compare.

R packages

- stats
- caret
- clv
- klaR

Technique	**Weka**	**R**
kMeans	weka.clusterers.SimpleKMeans	stats.kmeans klaR.kmodes
Normalization	*weka.filters.unsupervised.instance.Normalize*	caret.preprocess
Dunn Index		clv.Dun
PCA	*weka.filters.unsupervised.attributes.*PrincipalComponents	stats.prcomp