CS-E4650 Methods of Data mining
Home assignments 1

Deadline Sun 27.9.2020 24:00
Each task has maximum 20 points.
Note: Clarified task 1 and added hints to tasks 4c and 4d to make them easier!

1. Load data nba2013.csv from MyCourses home page under Home assignments. **Prune out non-numerical features, you will need only numerical features in the following tasks.**

   a) Perform K-means and evaluate the goodness of clustering using Silhouette Coefficient, Calinski Harabasz and Davies-Bouldin indices. Try the following values of K: a) $K = 10$, b) $K = 5$, c) $K = 2$.

   b) What is an optimal K and why?

   Hint: You can find the equations of goodness indices in the lecture 4 slides or use some library function to calculate them.

2. Use the same data as in task 1.

   a) Perform hierarchical agglomerative clustering using the following linkage metrics:

      – single-linkage metric
      – complete-linkage metric
      – average-linkage metric
      – distance of centroids metric

   b) What is the optimal metric for this data and why?

3. Use the same data as in task 2. Implement a program that shuffles the data set randomly. Shuffle the data and repeat task 2. Repeat shuffling and clustering at least 5 times to see if the clustering results (and goodness measures) change. What do you observe? Can you make any conclusions how robust different linkage metrics are to data order?

4. Look at cow data in Table 1. The task is to calculate distances and similarities between cows and present them as nearest neigbour/similarity graphs. It is enough to connect each cow to two nearest neighbours (unless there are more equally similar neighbours). You can present the graph as a matrix or draw it.

a) Scale the numerical features as needed and calculate pairwise distances using only numerical features. You can use the common Euclidean distance, but extra points for the Mahalanobis distance! Present the results as a similarity graph.

b) Calculate pairwise similarities using only categorical features. Use the Goodall measure that was presented in lecture 3 (see slides or text book section 3.3.2, since there are many Goodall measures). Present the result as a similarity graph.

c) Create a distance or similarity measure that combines the numerical and categorical measures so that neither group dominates (see lecture 3 and text book 3.2.3). Create now a similarity or nearest neighbour graph using the combined measure. **Change: You can use the overlap similarity for categorical features, since it is an easier measure to modify into a distance measure.**

d) If your combined measure was a similarity measure, present it as a distance measure. Is the distance measure a metric? Prove your answer. **Change: this is easier if you used the overlap measure in c).**

Table 1: Cow data: name, race, age (years), daily milk yield (litres), character and music taste.

name	race	age	milk/d	character	music
Clover	Holstein	2	20	lively	rock
Sunny	Ayrshire	2	10	kind	rock
Rose	Holstein	5	15	calm	country
Daisy	Ayrshire	4	25	calm	classical
Strawberry	Finncattle	7	35	calm	classical
Molly	Ayrshire	8	45	kind	country

5. **Note**: you can use code to do the computations, but you must describe all steps and report intermediate results.

   Consider the following two-variable data set, where each row corresponds to a point in 2-dimensional space:

$$\begin{pmatrix} 0 & 1 \\ -1/2 & 3/2 \\ 3/2 & 5/2 \\ 1 & 3 \end{pmatrix}.$$

(a) (5 points) Carry out the principal component analysis of these data, that is, compute the eigenvalue decomposition of the corresponding sample covariance matrix.

(b) (5 points) Consider the resulting decomposition.

  i. Use it to transform the original 2-dimensional data set into a 1-dimensional representation (a 4 × 1 matrix) such that the variance of the resulting data is equal to the largest eigenvalue.

  ii. Next, use it to transform the original data set into a 2-dimensional representation, such that the variance of one of the columns is equal to the smallest eigenvalue.

(c) (5 points) Given two points in d-dimensional Euclidean space,

$$x = (x_1, x_2, \ldots, x_d)^T,$$
$$y = (y_1, y_2, \ldots, y_d)^T,$$

the *Euclidean distance* between them is computed as follows:

$$d(x, y) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}.$$

  i. Compute the Euclidean distance between all pairs of points in the original data set.

  ii. Compute the Euclidean distance between all pairs of points in the 1-dimensional representation obtained in exercise 5b.

  iii. Compute the Euclidean distance between all pairs of points in the 2-dimensional representation obtained in exercise 5b.

  iv. What is the effect of the previous transformations on these distances?

(d) (5 points) Now consider the following data set:

$$\begin{pmatrix} \sqrt{1/2} & \sqrt{1/2} \\ \sqrt{1/2} & 2\sqrt{1/2} \\ 4\sqrt{1/2} & \sqrt{1/2} \\ 4\sqrt{1/2} & 2\sqrt{1/2} \end{pmatrix}.$$

Repeat exercises 5a and 5b on these data. What are the similarities and differences between the results on this data set and the first one? Can you give a geometric explanation for the similarities? Hint: plot the two data sets.