Data mining - first lab assignment report

Exercise 1. Distance function.

In this task I implemented the Minkowski distance function (and also the Manhattan, Euclidean and Chebyshev distance functions which are corresponding to Minkowski distance function with values p=1, p=2 and p=∞), the Canberra distance function and Mahalanobis distance function which can all be found in the “mydistfun” function in first_lab_assignment_Nikola_Tomažin.R.

Exercise 2. Feature selection.

In this task I implemented the function to compute entropy and Hopkins statistic.
The functions were tested on 2 datasets to have some comparison.

Entropy was calculated with formula:

$$E = -\sum_{i=1}^{m} \left[p_i \log(p_i) + (1 - p_i) \log(1 - p_i) \right]$$

where p_i is the proportion of the points in the region i, m - total number of regions.
Large values of entropy indicate poor clustering behaviour.

Let D be the data set to investigate and R is a representative sample of D, of power r. S is a synthetic data set of r data points randomly generated from the same domain. Let α1, . . . αr be the distances of each point of R to the nearest neighbour in D and β1, . . . βr are the distances of each point of S to the nearest neighbour in D. The Hopkins statistic is defined as follows:

$$H = \frac{\sum_{i=1}^{r} \beta_i}{\sum_{i=1}^{r} (\alpha_i + \beta_i)}.$$

While the values of the Hopkins statistic vary from 0-1, where values close to 1 tend to indicate the data is highly clustered, random data will tend to result in values around 0.5, and uniformly distributed data will tend to result in values close to 0.

Exercise 3. Clustering.

In the third task I implemented the k-medoids algorithm[1]. The dataset was 100 points split into 3 clusters so I used k=3 and got the best results for Euclidean and Manhattan distance functions, while for Canberra and Mahalanobis distance functions the result wasn’t that satisfying. The silhouette coefficient ranges from -1 to 1 where 1 means clusters are well apart from each other, 0 means clusters are indifferent, or we can say that the distance between clusters is not significant and -1 which means clusters are assigned in the wrong way. For the Euclidean and Manhattan we got values around 0.78 for all clusters and for Canberra and Mahalanobis there was one cluster with 0.7 but other 2 with 0.2 values.

Sources used:

https://www.geeksforgeeks.org/ml-k-medoids-clustering-with-example/
https://www.geeksforgeeks.org/ml-intercluster-and-intracluster-distance/
https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c
Data mining lectures at https://moodle.taltech.ee/course/view.php?id=31036
https://www.rdocumentation.org/

[1] Which i found at https://www.geeksforgeeks.org/ml-k-medoids-clustering-with-example/