Home assgnment 1: Distance function, classification, clustering.

Urmas Pitsi, 192028IAPM. Data Mining ITI8730.

Exercise 1: Distance functions.

Implemented following functions: chebyshev, manhattan, canberra, euclidean, minkowski, mahalanobis and generic distance function that incorporates all previously mentioned distance functions. All functions work on n-dimensional (n-column) matrixes as inputs.
E.g: `euclidean <- function(x1, x2){ return(sqrt(rowSums((x1 - x2) ^ 2))) }`

Exercise 2: Clustering.

Implemented k-means algorithm, represented by the following pseudocode:

```
kmeans_clusters <- function(data, num_clusters, num_iterations, metric):
  Start iterating until solution.
  for (1 to num_iterations){
    1. Initialize random center points for clusters or store new centers.
    2. Assign each point to a cluster.
    3. Calculate center of each cluster.
    4. Check if solution. If found solution then exit function.
```

For cluster analysis implemented following functions:
intra_cluster_distances, inter_cluster_distances, intra_to_inter_ratio, silhouette_coefficient, silhouette_score, cluster_inertia.

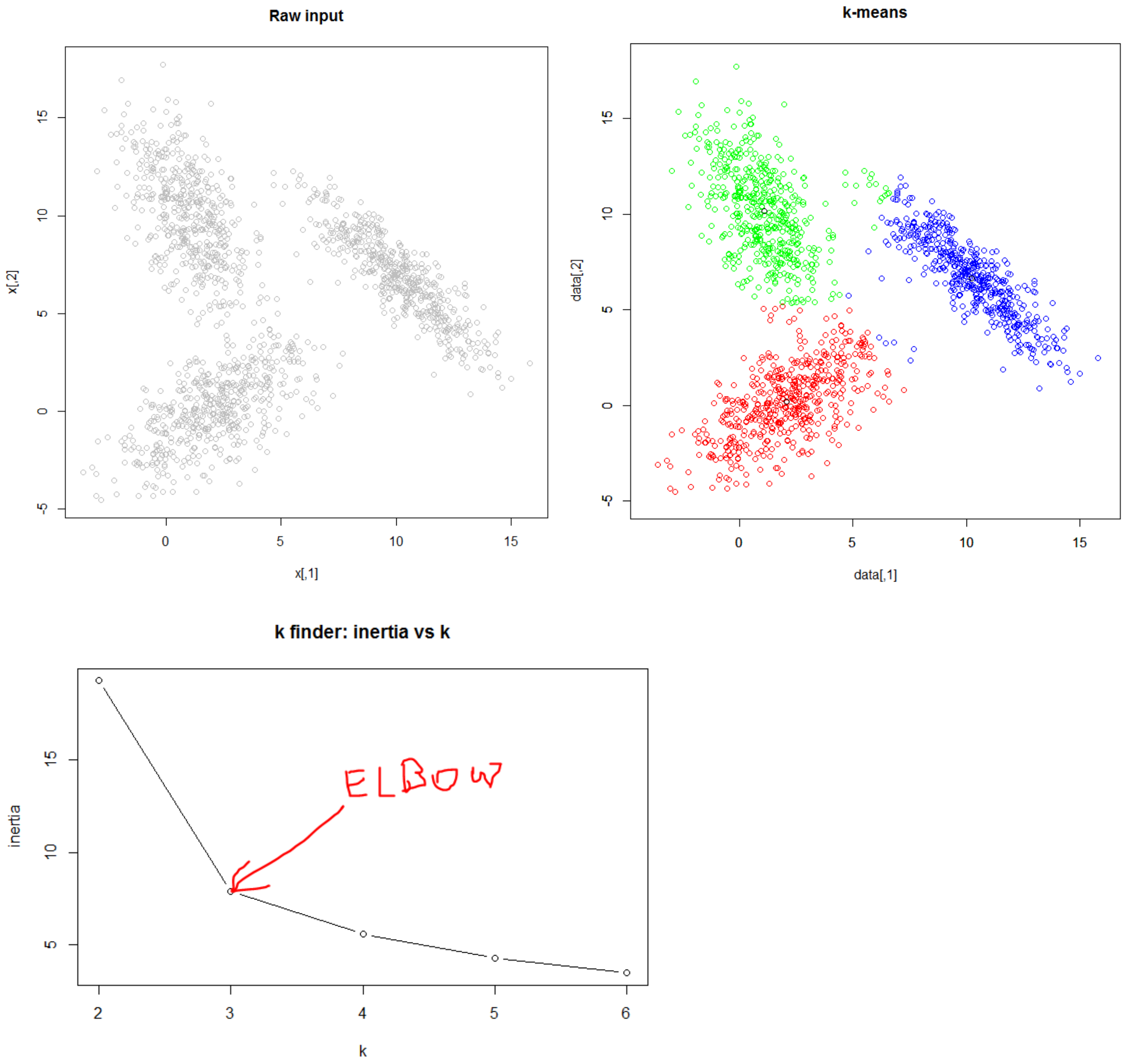
Exercise 3: Classification.

Implemented knneighbors algorithm. For each datapoint we calculate distances to all other datapoints and take most popular label among k closest datapoints.
For classification analysis I used same analytics as for cluster analysis. On top of that I implemented functions: accuracy_score, confusion_matrix function.

Exercise 4: Classification wrapper.

In order to find optimal k for knneighbors algorithm, I implemented an iterative function that finds "best" classification results for different k-values. Best classification result can be determined based on different analytics/scores: accuracy_score, cluster_inertia, silhouette_score etc.
I applied the similar analysis for clustering using same techniques in finding optimal k for kmeans algorithm ("elbow" rule, using inertia as clustering score).

Figure 1: k-means: from raw data to 3 clusters using “elbow” rule.



Raw input

k-means

k finder: inertia vs k

Appendix:

Source: https://gitlab.cs.ttu.ee/urpits/data-mining-iti8730/tree/master/assignment1

List of source files:

cluster_analysis.R
clusters_generator.R
demo_cluster_analysis.R
demo_mahalanobis.R
distances.R