Report prepared by: Lizaveta_Miasayedava (184361IVGM)

Exercise 1. Distance functions.

Implementation: *my_matlib.R* (different auxiliary calculations), *my_distfunc.R*

Description: there were implemented several distance functions, including Minkowski (of orders 1,2,3,4,5), Canberra, Mahalanobis, Chebyshev and cosine distances for any finite number of dimensions.

These functions are hereinafter used for solving clustering and classification problems.

Exercise 2. Clustering.

Implementation: *my_kmeans.R, my_eval_criteria.R, test_kmeans.R*

Description: k-means algorithm implementation allowing different distance functions (implemented in exercise 1). The clustering solution for one of the datasets can be seen in the Figure 1 (Mahalanobis distance function, k=3).

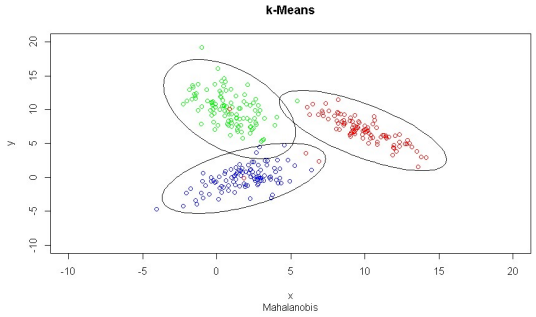

Figure 1

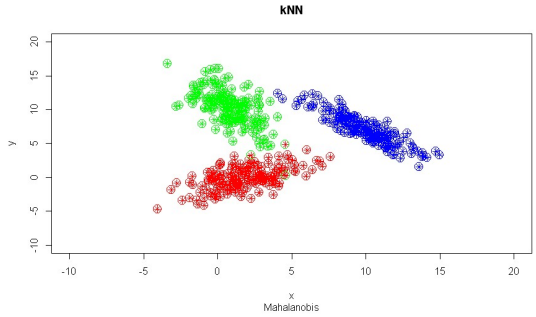

Figure 2

Then the internal cluster validation for different distance functions was carried out. Used criteria: silhouette coefficients and ratios of intra- to inter- cluster distances:

Ratio of intra- to inter- cluster distances:

	Cluster 1	Cluster 2	Cluster 3	Average
Minkowski1	0.29626090	0.3374498	0.3056188	*0.3131098*
Minkowski2	0.31079107	0.3404993	0.2844094	**0.3118999**
Minkowski3	0.31079107	0.3404993	0.2844094	**0.3118999**
Minkowski4	0.31079107	0.3404993	0.2844094	**0.3118999**
Minkowski5	0.31079107	0.3404993	0.2844094	**0.3118999**
Canberra	0.65560584	0.7453502	0.2147378	0.5385646
Mahalanobis	0.31079107	0.3404993	0.2844094	**0.3118999**
Chebyshev	0.31690543	0.3373982	0.2844094	*0.3129044*
Cosine	0.07701449	0.2358751	0.7146070	0.3424988

Silhouette coefficients:

	Number of SC >= 0	Number of SC < 0	Average value
Minkowski1	244	56	*0.26426109*
Minkowski2	252	48	**0.31032607**
Minkowski3	252	48	**0.31032607**
Minkowski4	252	48	**0.31032607**
Minkowski5	252	48	**0.31032607**
Canberra	97	203	-0.35550963
Mahalanobis	252	48	**0.31032607**
Chebyshev	252	48	*0.30563878*
Cosine	200	100	-0.01691794

Conclusion: since smaller ratio values of intra- to inter- cluster distances and larger positive values of silhouette coefficients indicate better separation of clusters**,** for given dataset and implemented k-means algorithm Minkowski (order 2-5) and Mahalanobis distance functions demonstrate better performance.

Exercises 3. Classification.

Implementation: *my_knn.R, test_knn.R*

Description: k nearest neighbors algorithm implementation allowing different distance functions and different values of hyper-parameter k. With Mahalanobis distance function and k=3 there was obtained the following classification of points (Figure 2).

Conclusion: the evaluation of accuracy achieved with different distance functions has shown the same high algorithm performance almost for all of the implemented distance functions:

	Accuracy	Fischer score x	Fischer score y
Minkowski1	0.994	4.71828767	4.2607901
Minkowski2	0.994	4.71828767	4.2607901
Minkowski3	0.994	4.71828767	4.2607901
Minkowski4	0.994	4.71828767	4.2607901
Minkowski5	0.994	4.71828767	4.2607901
Canberra	0.988	4.74190567	3.8381997
Mahalanobis	0.994	4.71828767	4.2607901
Chebyshev	0.994	4.71828767	4.2607901

Exercise 4. Classification Wrapper.

Implementation: *my_classwrap.R, test_class_wrapper.R*

Description: there was implemented a wrapper method that calls implemented previously k nearest neighbours algorithm with different values of k and calculates corresponding Fischer score and accuracy of the results.

	Accuracy	Fischer score x	Fischer score y
k=2	0.994	4.718288	4.260790
k=3	0.994	4.718288	4.260790
k=4	0.994	4.718288	4.260790
k=5	0.994	4.718288	4.260790
k=6	0.992	4.797954	4.317369
k=7	0.992	4.797954	4.317369
k=8	0.992	4.797954	4.317369
k=9	0.992	4.797954	4.317369
k=10	0.990	4.860293	4.361906

Conclusion: the highest accuracy was achieved with k=2..5. According to the obtained Fischer scores, both features (x and y) have almost the same discriminatory power.

Exercise 5. Local outlier factor.

Implementation: *my_LOF.R, test_LOF.R*

Description: for given dataset of points there were found k clusters with k-means algorithm and calculated local outlier factors for points of one of the clusters. According to obtained results, there was build a plot where along x-axis are distances of different cluster's points to the cluster center and along y-axis – their LOFs (Figure 3).

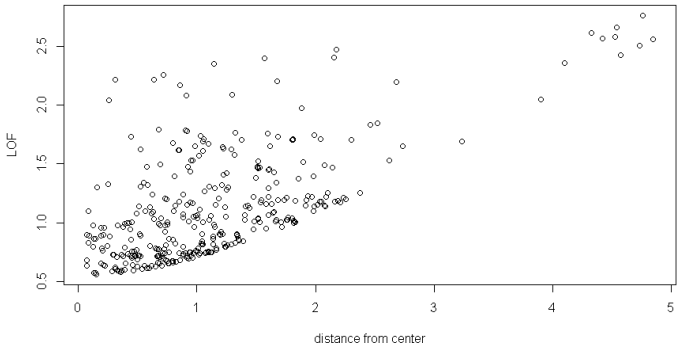

Figure 3

Conclusion: According to this plot, there is a general trend of increasing values of local outlier factors with the increase of the distance from the elliptic cluster's center.