Data mining ITI8730

Jaroslav Gorškov

Homework assignment 1

Exercise 1.

minkowski_distance(v1,v2,p)
canberra_distance(v1,v2)
mahalanobis_distance(v1,v2,dt)

Please note that in the case of mahalanobis distance it is assumed that both v1 and v2 are taken from data frame dt. Latter is used to calculate the inverted covariance matrix.

Exercise 2.

I have chosen to implement k-means, and compare the clustering results to k-medoids, for which I've used in-built solution (function *pam*).
K-means operates on the notion of readjusting centroids. The first centroids are picked randomly from the given data frame. Subsequent ones will be picked in a way to minimize the distance between them and other points of the cluster. In case of k-means, the coordinates of a newly generated centroid(*k*) are the means of the data points of the cluster *k*. The process stops when the newly generated centroids do not differ from the previously generated ones.

Running k_means function with different hyperparameter *p* (responsible for the metric function choice), we can observe, that in the case of manhattan and euclidean distances, the results are almost identical, varying by a very small degree. Running the k_medoids algorithm with the same metric function, we verify the correctness of our own k_means implementation, since the results are extremely similar.

To estimate the quality of clustering, we use two different approaches (considered different, but there are many common factors): intra-inter cluster ratio and silhouette coefficient. First one mentioned is self-descriptive - we find the ratio of sum of distances between points in the same cluster to sum of distances between points in different clusters. In case of silhouette coefficient, to use the formula depicted on the right, we need the intra cluster sum of means, and sum of distances of points from one cluster to points in other clusters.

$$s(i) = \frac{D^{out}_{min_i} - D^{in}_{avg_i}}{max\{D^{out}_{min_i}, D^{in}_{avg_i}\}}$$

Please note, that my implementation is very time consuming, yet works correctly. The slowness of algorithm could be due to the fact that for-loops are not optimal in R. Since we have them nested, the big O notation reminds us of that O(n^2). The other reason (a very probable one!) is that we didn't take smaller samples and considered all points of the data set.

The intra-inter cluster ratio I got is 0.2417539. The closer to zero, the better is clustering performance. Silhouette coefficient was calculated to be 236.3081. Higher positive numbers signify of a good clustering.

Exercise 3.

We partition our data in two, such that first part consists of 70% of initial records (used for training purpose), and other of 30% (for testing purpose).

For each point in the testing data, we find the sum of pairwise distance to points in training data. Then we sort ascending order and get first k points (hence the name, k-nearest-neighbours). Having k points and their labels (classes they belong to), we find the class that the majority of the said points are member of. The class received is the predicted class of the corresponding point in test data.

To evaluate the accuracy of the classifier, we just need to divide amount of correct guesses by the total number of points in testing data. For the data frame *kdata_with_labels.RData* , the accuracy is very high, ranging from 98% to 100% , depending on the choice of hyperparameter *k*.

Exercise 4.

In order to find the optimal parameter *k* for knn algorithm, we check the accuracy of classification for different *k* values provided as an interval in the function parameters (*k_min, k_max*). We choose the *k* for which the classification has the highest accuracy.

References:

https://stats.stackexchange.com/questions/65705/pairwise-mahalanobis-distances

https://www.datanovia.com/en/lessons/cluster-validation-statistics-must-know-methods/#silhouette-coefficient

http://enhancedatascience.com/2017/10/24/machine-learning-explained-kmeans/

https://www.machinelearningplus.com/statistics/mahalanobis-distance/

https://dataaspirant.com/2017/01/02/k-nearest-neighbor-classifier-implementation-r-scratch/

https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorith-and-practical-examples/

Aggarwal, Charu. *Data mining: The Textbook*. Switzerland, 2015

Lecture slides