Human Activity Recognition

Report prepared by: Lizaveta Miasayedava (184361IVGM)

Problem description

The problem of human activity recognition is to classify accelerometer and gyroscope data recorded by special devices (e.g. smartphones) in order to recognize movements (activities). If a computer system is able to recognize the type of the subject activity, it can, for example, offer an assistance (to people with limited conditions). However, devices provide a large number of records (time series data), with lots of attributes, and it's not known for certain how sensor records relate to specific activities. Moreover, each subject may perform an activity differently, and this variations reflect in the recorded sensor data.

Data description

Data source: https://www.kaggle.com/uciml/human-activity-recognition-with-smartphones

The data was acquired from the smartphones of 30 experiment participants. The phone was configured to record two implemented sensors (accelerometer and gyroscope). The data was labeled manually with the following types of activities: walking, walking upstairs, walking downstairs, sitting, standing and laying. Also the data is partitioned into 2 sets (70% training data, 30% test data).

The data is preprocessed (raw data isn't available). The sensor signals (accelerometer and gyroscope) were preprocessed using noise filters. Then this data was splitted into the samples of fixed-width windows of 2.56 seconds and 50% overlap (128 data points/window), so the resulting points are equally spaced.

Task specification

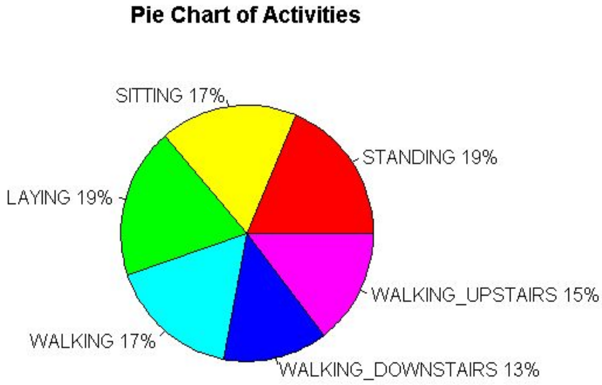
The goal of the project is to explore the data gathered in the experiment and create the model allowing for the classification of new unseen subjects from their sensor data.

Data exploration

Source code: har_exploration.R

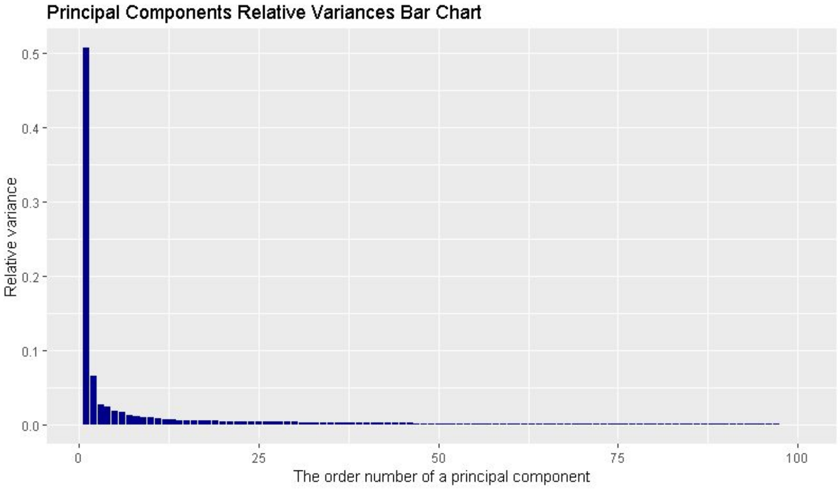
The available data is a dataset of 10299 rows of 563 features (7352 rows in a training subset and 2947 rows in a test subset). From 563 features, 561 features are sensor measurements and 2 features: subject and activity can be considered labels.

The proportions of data by activities:

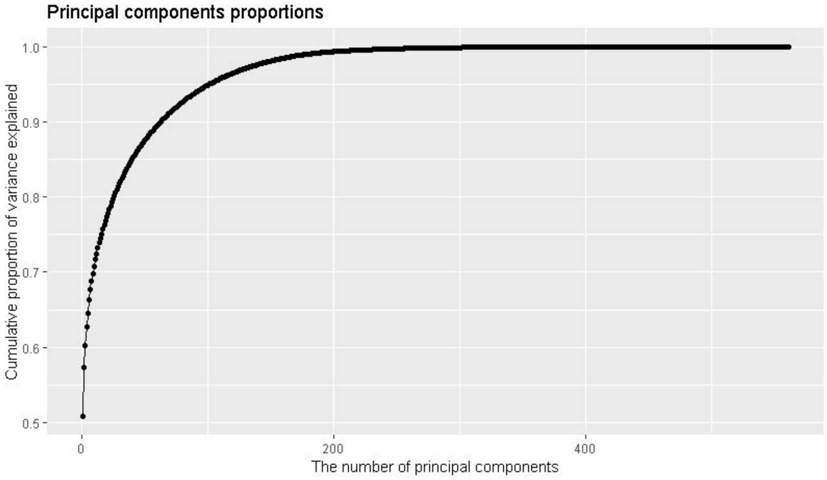


Since the dimensionality of the datasets is high, there were performed *Principal Component Analysis* that converts a set of correlated variables into a set of linearly uncorrelated variables (principal components). The principal components are orthogonal to each other and have descending variances.

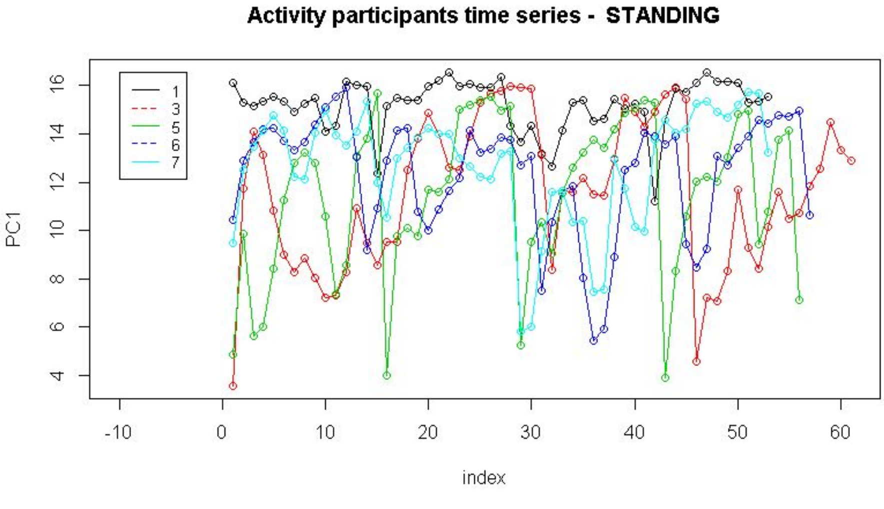
Resulted principal components provide the following information: the first principal component has the largest variance - slightly more than 50% (50,78%), other components provide less than 10% variance (6,58%; 2.81%; 2.5%; etc.)



95% of variation can be described with the first 100 principal components.



Addressing each activity separately, there can be noticed that each subject (experiment participant) performs an activity in a different way (shapes of time series plots differ).
For example, there are plotted time series of some standing subjects:



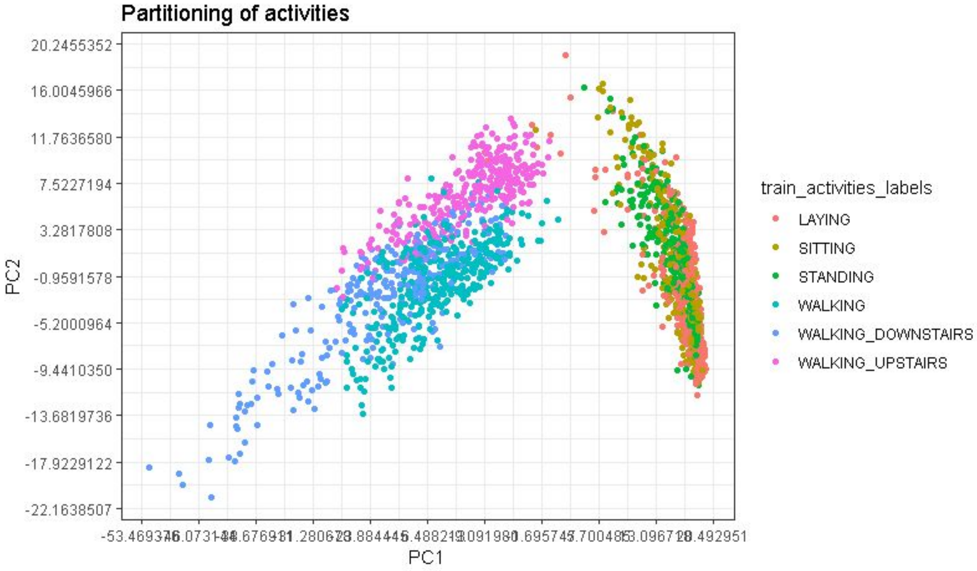
Thereby, the data can be viewed from 2 perspectives: point-wise and shape-wise.

Clustering

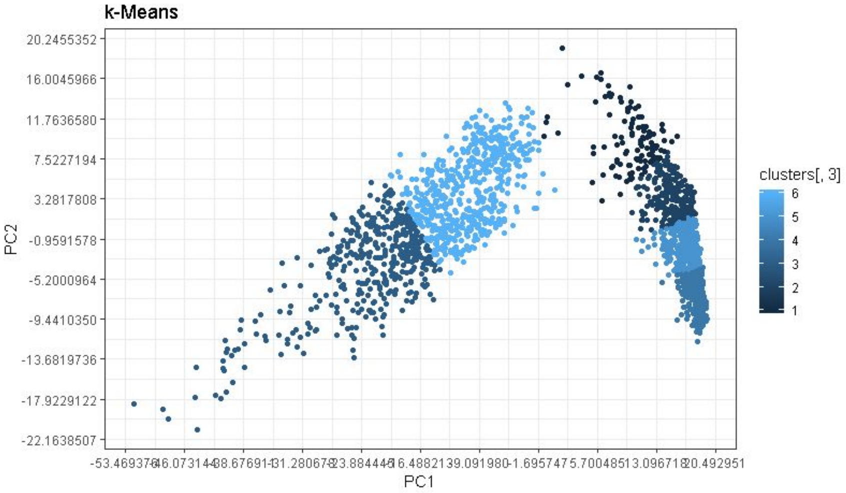
Source code: har_clustering.R, kmeans.R, dist_func.R, graph_clustering.R, my_graph_helper_functions.R

Clustering can be performed in the data exploration or preprocessing stage in order to investigate the specificities of data points and find possible outliers which removal may result in more accurate classification.

For the clustering of the human activity data represented by its principal components the k-Means algorithms with Euclidean distance function was used. Clusters have elliptic form, but since activities can be viewed in the context of smooth transitions one into another, they are overlapping by nature:



According to the provided data, there can be seen 2 separable clusters (moving and staying at the same place). The clustering into 6 partitions with k-Means algorithm looks in the following way:



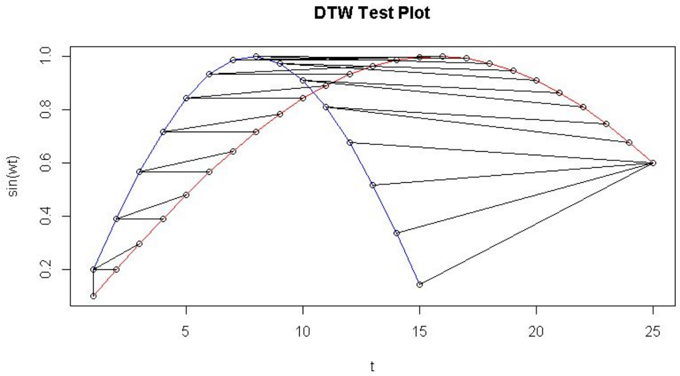
In order to evaluate the separability of clusters silhouette coefficients were calculated:

k	2	3	4	5	6	7	8
Average silhouette coefficient	0.4153 2	0.045240 2	-0.094404	-0.172815	-0.240613	-0.245986	-0.284772

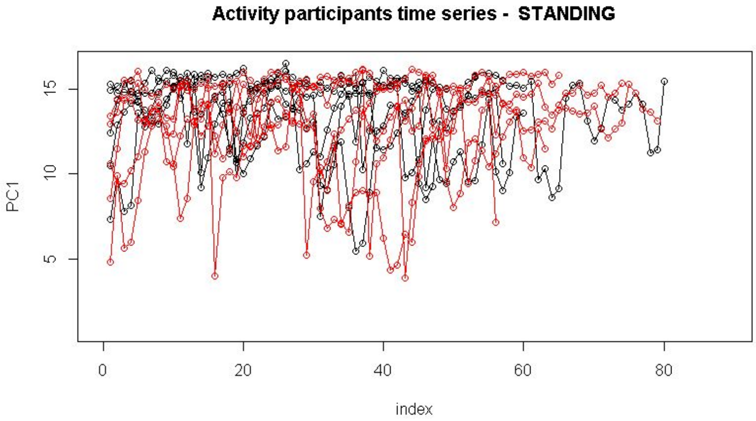
Since larger positive values indicate better separation of clusters, the most separable clustering result is provided by the k-Means algorithm in case of k=2.

For each activity type it’s possible to perform clustering of experiment participants to find out what participants move in a similar way. Shape similarities of time series may be measured with dynamic time warping.

Source code: test_dtw_dist_func.R



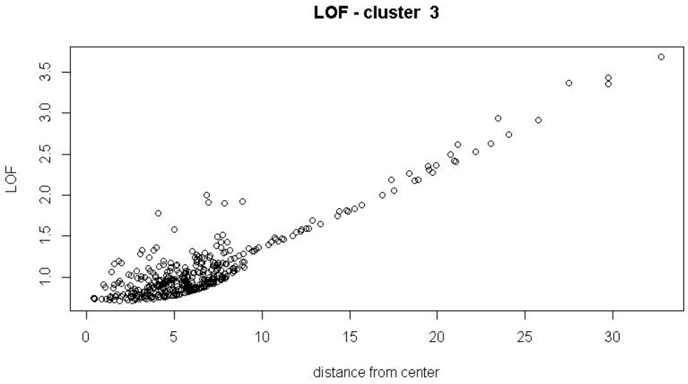
In this case time series of different participants can be considered graph nodes and similarities in time series - the weights of edges. Considering dynamic time warping distance matrix as adjacency matrix, it’s possible to use Kernighan-Lin algorithm (often used for community detection) to split participants time series into 2 balanced groups.



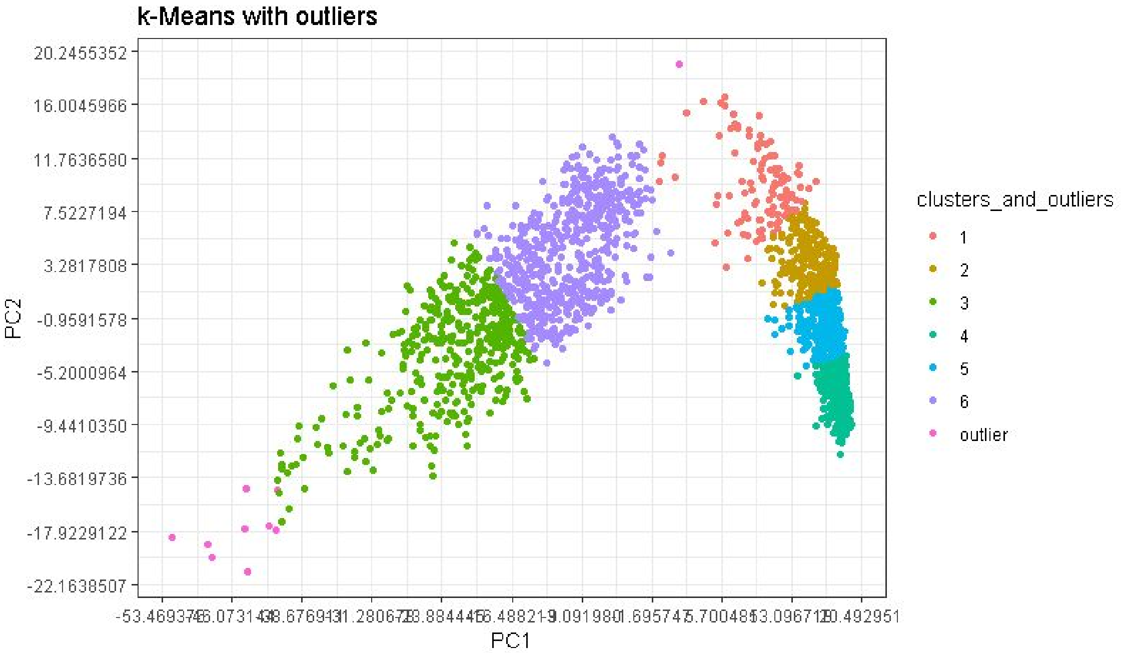
Outlier detection

Source code: har_outlier_detection.R, my_LOF.R

Outlier detection in this work is based on the local outlier factor algorithm that measures local deviations of points with respect to its neighbours. Outlier detection in activities is performed for each cluster obtained with k-Means algorithm.



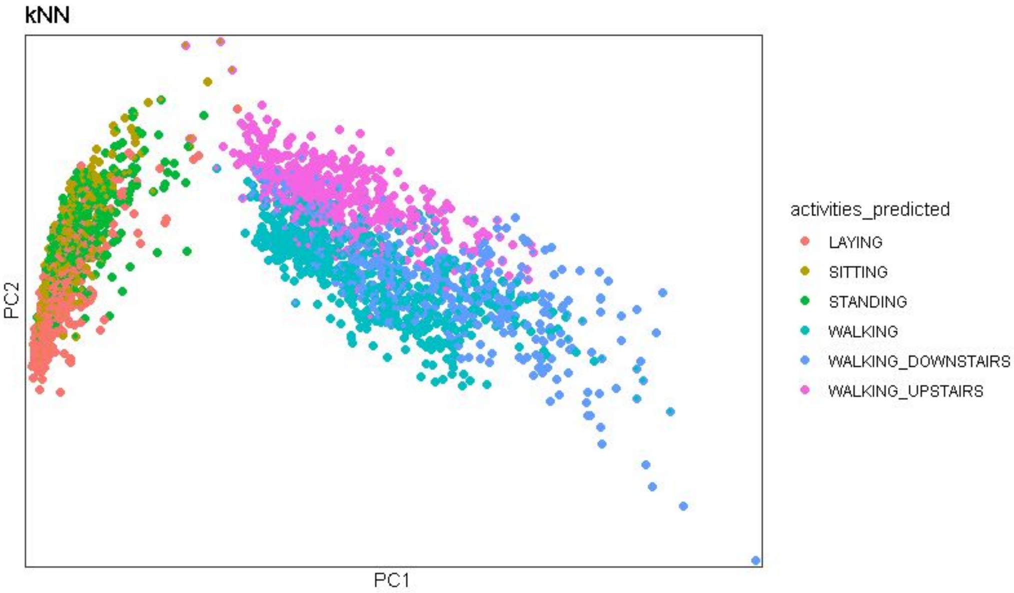
10 data points with the highest local outlier factors are represented below:



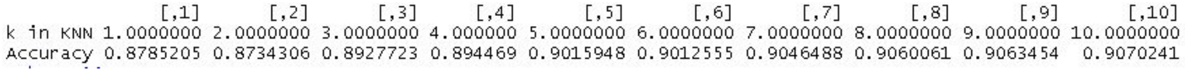
Classification

Source code: har_classification.R, my_knn_dtw.R, my_knn.R, har_classification_helper.R

Point-wise classification of activities: kNN + Euclidean distance function.

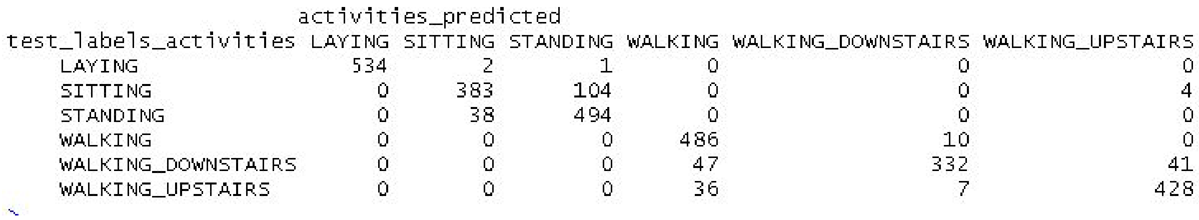


Accuracies of classification for different values of k hyperparameter:



```
              [,1]      [,2]      [,3]     [,4]      [,5]      [,6]      [,7]      [,8]      [,9]      [,10]
k in KNN 1.0000000 2.0000000 3.0000000 4.000000 5.0000000 6.0000000 7.0000000 8.0000000 9.0000000 10.0000000
Accuracy 0.8785205 0.8734306 0.8927723 0.894469 0.9015948 0.9012555 0.9046488 0.9060061 0.9063454  0.9070241
```

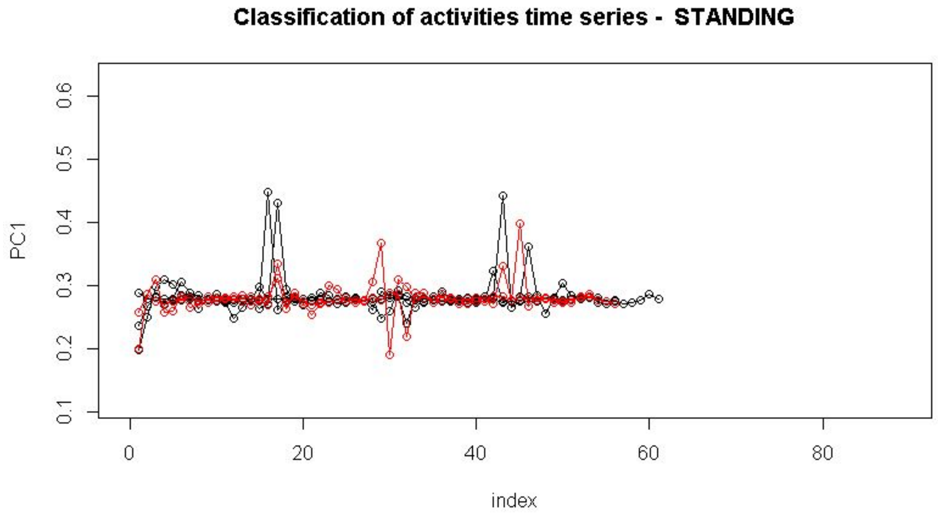
Confusion matrix in case of k=10:



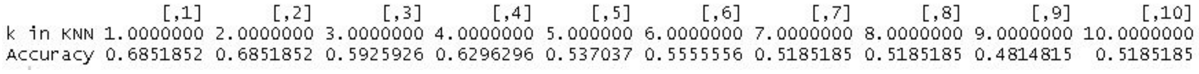
```
                          activities_predicted
test_labels_activities LAYING SITTING STANDING WALKING WALKING_DOWNSTAIRS WALKING_UPSTAIRS
    LAYING                534       2        1       0                  0                0
    SITTING                 0     383      104       0                  0                4
    STANDING                0      38      494       0                  0                0
    WALKING                 0       0        0     486                 10                0
    WALKING_DOWNSTAIRS      0       0        0      47                332               41
    WALKING_UPSTAIRS        0       0        0      36                  7              428
```

Shape-wise classification of activities: kNN + DTW

Black plots - train time series, red plots - test time series.

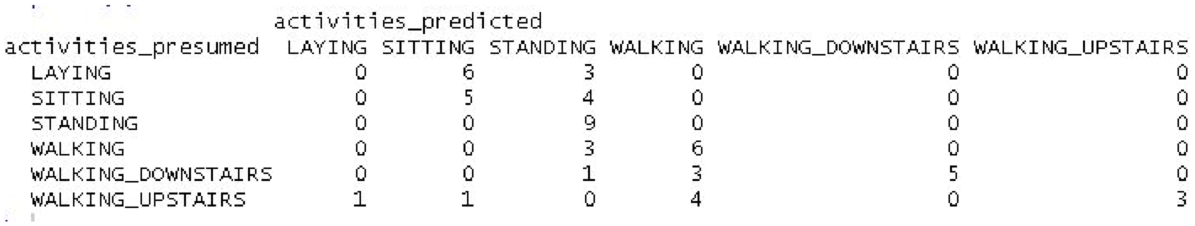


Accuracies of classification for different values of k hyperparameter:



```
              [,1]      [,2]      [,3]      [,4]     [,5]      [,6]      [,7]      [,8]      [,9]      [,10]
k in KNN 1.0000000 2.0000000 3.0000000 4.0000000 5.000000 6.0000000 7.0000000 8.0000000 9.0000000 10.0000000
Accuracy 0.6851852 0.6851852 0.5925926 0.6296296 0.537037 0.5555556 0.5185185 0.5185185 0.4814815  0.5185185
```

Confusion matrix in case of k=1:



```
                    activities_predicted
activities_presumed  LAYING SITTING STANDING WALKING WALKING_DOWNSTAIRS WALKING_UPSTAIRS
  LAYING                  0       6        3       0                  0                0
  SITTING                 0       5        4       0                  0                0
  STANDING                0       0        9       0                  0                0
  WALKING                 0       0        3       6                  0                0
  WALKING_DOWNSTAIRS      0       0        1       3                  5                0
  WALKING_UPSTAIRS        1       1        0       4                  0                3
```

Conclusion: as a result of the work, the gathered human activity data was explored, analyzed for clusters and outliers, and used to create the model allowing for the classification of activities of unseen subjects from their sensor data.