Course 4 – Unsupervised Learning

The project uses the Human Activity Recognition Using Smartphone Dataset from [UCI Machine Learning Repository: Human Activity Recognition Using Smartphones Data Set] which contains 562 columns, of which 561 are various sensor readings from the mobile phone, and the final column is the activity that was being performed by the user (walking, walking upstairs, walking downstairs, sitting, standing, and laying). There are 10299 rows of data. Some of the columns are shown below:

	tBodyAcc-mean()-X	tBodyAcc-mean()-Y	tBodyAcc-mean()-Z	tBodyAcc-std()-X	tBodyAcc-std()-Y	tBodyAcc-std()-Z
0	0.29	-0.02	-0.13	-1.00	-0.98	-0.91
1	0.28	-0.02	-0.12	-1.00	-0.98	-0.96
2	0.28	-0.02	-0.11	-1.00	-0.97	-0.98
3	0.28	-0.03	-0.12	-1.00	-0.98	-0.99
4	0.28	-0.02	-0.12	-1.00	-0.98	-0.99

angle(X,gravityMean)	angle(Y,gravityMean)	angle(Z,gravityMean)	Activity
-0.84	0.18	-0.06	STANDING
-0.84	0.18	-0.05	STANDING
-0.85	0.18	-0.05	STANDING
-0.85	0.18	-0.05	STANDING
-0.85	0.19	-0.04	STANDING

The data set was clean and no pre-processing was required:

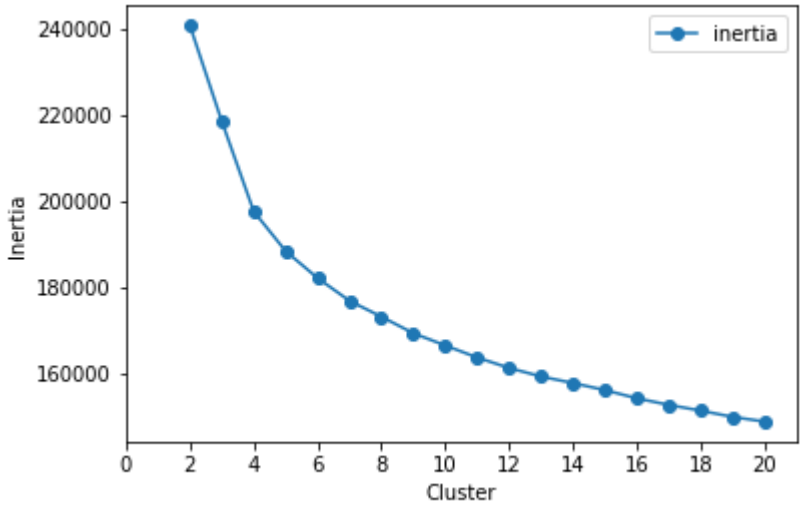
- No missing values
- All values in the same range
- No categorical data (except Activity column)

The target column (Activity) was removed and unsuvervised learning techniques were used to cluster the data. The following steps were performed:

1. K-Means Clustering
2. HAC with
    a. Average Linkage
    b. Ward Linkage
3. PCA for dimensionality reduction
    a. K-Means with PCA data

K-Means

K-Means was done for cluster size ranging from 2 to 20 and the inertia was calculated.



There is no clear elbow point that was be found from the graph. Since we know that there are supposed to be 6 target values, we'll continue with 6 clusters for now. The results for 6 clusters are given below:

		number
Activity	kmeans	
LAYING	0	12
	1	171
	5	1761
SITTING	0	3
	1	468
	2	1257
	5	49
STANDING	1	672
	2	1234
WALKING	0	704
	3	150
	4	868
WALKING_DOWNSTAIRS	0	187
	3	455
	4	764
WALKING_UPSTAIRS	0	1166
	3	81
	4	297

- Laying was clustered to group 5
- Sitting was majorly clustered to group 2
- Standing was mostly to clustered group 2, followed by comparable clustering to groups 1 and 0
- Walking was clustered to groups 0 and 4
- Walking downstairs was clustered to group 4, followed by 3 and 0 respectively
- Walking upstairs was clustered to group 0

In conclusion the clustering was unsuccesful. The three walking categories were mostly clustered to the same group (0). Standing and Sitting were clustered to group 2 and laying to group 5.

HAC

Hierarchical Agglomerative Clustering was done with average and ward linkages for 6 clusters.

Activity	HAC_average	number
LAYING	0	1944
SITTING	0	1777
STANDING	0	1906
WALKING	1	1
	3	1714
	4	7
WALKING_DOWNSTAIRS	1	24
	2	4
	3	1375
	4	1
	5	2
WALKING_UPSTAIRS	3	1544

Activity	HAC_ward	number
LAYING	0	161
	2	1783
SITTING	0	921
	2	11
	5	845
STANDING	0	1416
	5	490
WALKING	1	1300
	3	371
	4	51
WALKING_DOWNSTAIRS	1	1000
	3	127
	4	279
WALKING_UPSTAIRS	1	373
	3	1171

Both linkages are unsuccessful in clustering as required.

- Average linakge clusters laying, sitting and standing together (group 0), and the three walking groups are clustered together (group 3).
- Ward Linkage clusters laying and standing together in group 0, sitting is split between groups 0 and 5, walking upstairs is group 3, and walking and walking downstairs are in group 1.

In conclusion, both K-Means and HAC were unsuccessful in clustering as required and don't give satisfactory results. The algorithms have difficulty differentiating between the non-movement activities (laying, sitting, standing), and the movement activities (the three walkings). In the next section PCA is done and the new data set with reduced dimensionality is used for K-Means.

PCA

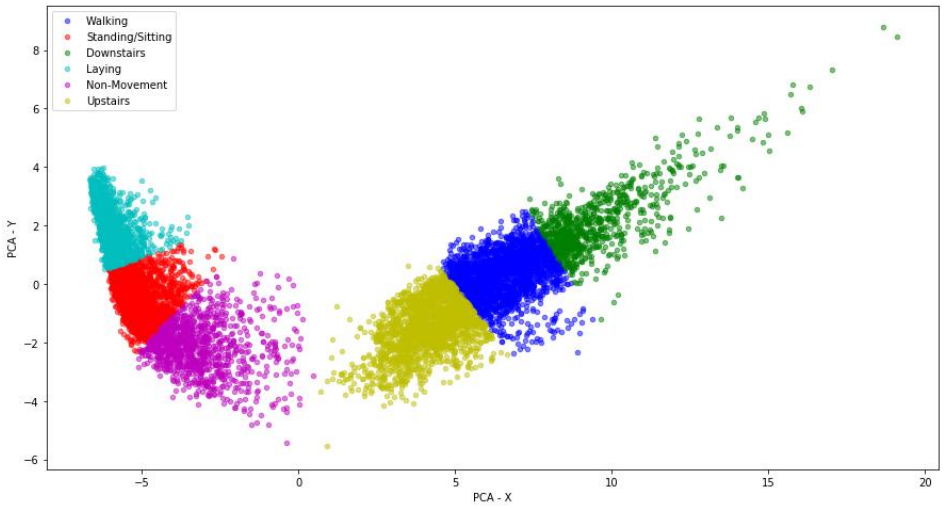
PCA is performed on the dataset to reduce the dimensions and the explained variance of different models are given,

n	model	var
2	PCA(copy=True, iterated_power='auto', n_compon...	0.67
27	PCA(copy=True, iterated_power='auto', n_compon...	0.88
52	PCA(copy=True, iterated_power='auto', n_compon...	0.93
77	PCA(copy=True, iterated_power='auto', n_compon...	0.96
102	PCA(copy=True, iterated_power='auto', n_compon...	0.97
127	PCA(copy=True, iterated_power='auto', n_compon...	0.98
152	PCA(copy=True, iterated_power='auto', n_compon...	0.99
177	PCA(copy=True, iterated_power='auto', n_compon...	0.99
202	PCA(copy=True, iterated_power='auto', n_compon...	1.00
227	PCA(copy=True, iterated_power='auto', n_compon...	1.00

We see that as we reach near 200 features, the explained variance reaches 100%. For our analysis we use the first model with 2 features and do K-means clustering with 6 clusters.

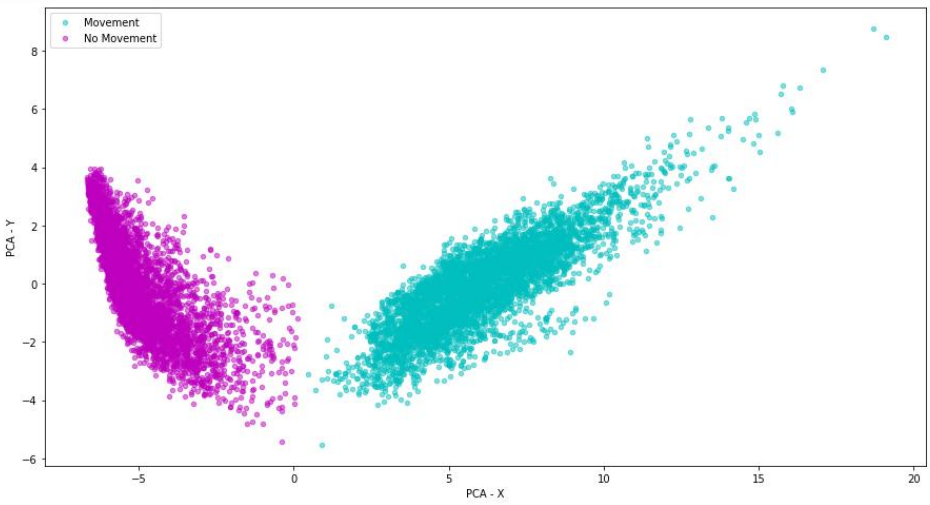
Activity	pca	
LAYING	1	527
	3	1237
	4	168
	5	12
SITTING	1	890
	3	552
	4	332
	5	3
STANDING	1	1112
	3	230
	4	564
WALKING	0	870
	2	198
	5	654
WALKING_DOWNSTAIRS	0	690
	2	533
	5	183
WALKING_UPSTAIRS	0	442
	2	39
	5	1063

The movement activites are clustered in groups 0, 2 and 5, whereas the non-movement activites are clustered across groups 1, 3 and 4.



On repeating the same thing with 2 clusters instead, movement (group 1) and non-movement (group 0) activities, the clustering works significantly well.

Activity	pca	number
LAYING	0	12
	1	1932
SITTING	0	3
	1	1774
STANDING	0	1
	1	1905
WALKING	0	1722
WALKING_DOWNSTAIRS	0	1406
WALKING_UPSTAIRS	0	1544



Conclusion

In conslusion, the chosen clustereing methods did not work as expected on the dataset. Different algorithms such as DBSCAN could be tried and the hyperparameters could be tuned. Data could also be reduced to slightly higher dimensions before applying clustering methods, or a non-liner PCA method (kernal PCA) could be used. The demensionaly reduced data could also be used for supervised learning for classification.