Report Homework 3

The first thing I did wad to load the dataset and print some information about it. I saw that there were 10 columns latitude, longitude, housing\_median\_age, total\_rooms, total\_bedrooms, population, households, median\_income, median\_house\_value that are numbers(float64) and ocean\_proximity that is an object(string). In each column there are 20640 data values except fot the total\_bedrooms that are 20433, so 207 less than the others. This 207 are null values. I looked also at basic statistics for those columns and saw the mean, the standard deviation, the min, the max and the quartiles. The first thing to do was to handle the missing values for total\_bedrooms, that in my first approach was to delete them. Then I changed this thing when I plotted the frequency distribution of each feature and I though about replacing the missing values with the median value or the mean value. I chose the median because the values are different and the mean couldn’t get an exact approximation of it.

Then was the time to watch for elbow method and finding the best number of clusters for raw data. Here I saw that the column ocean\_proximity gave some problems when running k-means++ for computing the elbow method so I encoded it with a label encoder, giving for each value of the column, that are 5, a number from 0 to 4. Then I computed and plotter the elbow method and saw from the graphs that the best number of clusters were 4. The elbow method is a technique that for a given number of clusters k, compute the inertia which is the sum of squared distances between each data point and its assigned cluster center. After that I looked for silhouette score to see if the values given from it were different or the same as elbow method. The silhouette score is computed by subtracting at the minimum average distance from the data point j to all points in any other cluster the average distance between the data point j and all other points in the same cluster and dividing this by the maximum of those values. Surprisingly I saw that the results were different, so by looking at the graphs I saw different optimal values of number of clusters. For the rest of the code I used the values from elbow method because it was easier and quicker to compute, even if has less precision than silhouette.

Then I ran k-means++ on raw data with the optimal number given by the elbow method, that was 4, and plotted the results with latitude and longitude. In this case I had some problems because by running k-means on a multidimensional dataset then clusters the data on those dimensions and how to plot the results was not clear to me. I thought about using some dimensionality reduction technique like PCA, but didn’t like the idea because it was not mentioned in the exercise and maybe I was going wrong. So, I plotted the results over 2 dimensions, latitude and longitude, even if the cluster was done on multidimensional database. Of course the clusters were not totally separated but I thought that was the way to go. After that I tried running k-means++ only on 2 dimensions, latitude and longitude, and here I had the results I was looking for. Clusters were separated correctly and also the centroids were correct.

Then was the time of doing some feature engineering on the data. For this reason I plotted the frequency distribution of the features to see what I could have used to optimize the dataset. After some try, I thought about one hot encoding the ocean-proximity feature, and scale all the others. I used 2 different scalers, the standard and the min-max, and I separated the features like this:

* Standard Scaler -> longitude, latitude, total\_rooms, total\_bedrooms, population, households and median\_house\_value, because they are features with large scales and lots of different values, for example latitude and longitude have also negative values, while median\_house\_value or population had very distinct and both small and large values.
* Min Max Scaler -> median\_income and housing\_median\_age because they have a relatively uniform distribution with small scale, for example median\_income goes from 0 to 15.

After this process I repeated the elbow method and silhouette score for the processed data and saw that the number of optimal clusters increased a bit. By doing preprocessing k-means++ was able to detect different clusters respect to the raw data so then I ran again it on the optimal value given by the elbow method that was 7. Even in this case I plotted the results of the multidimensional k-means on just latitude and longitude and saw that the clusters were separated better even if the problem of the multi dimensionality. I also repeated the k-means on only 2 features and in this case I plotted the result with centroids and had a very good graph with 7 different clusters of latitude and longitude.

About the running time I didn’t notice big differences between raw and processed data, maybe because the k-means++ algorithm converged fast in both cases. The only thing I noticed about running time was when I used silhouette instead of the elbow method and there I noticed that it was much more expensive in computation, so maybe for even larger dataset it would have taken a lot of time and for this reason I used the optimal value for clustering of the elbow.

The graphs are commented in the notebook.