COMP9318 PROJECT

PART 1:

For part 1, we had to implement the product quantization method with L1 distance as the distance function. We were required to write a method pq()takes four arguments as input: data, P no. of partitions, init_centroid array, max_iter. It returns a codebook and codes for the data vectors

The following adjustments were made to the method given in the class:

1. Instead of L2 distance, L1 distance was implemented, i.e. Euclidean distance was replaced by Manhattan distance as the distance function.
2. K-medians clustering algorithm is used to determine the centroid by calculating the median of each cluster (instead of k-means algorithm).
3. Manhattan distance is used again to assign minimum distance values to “codes”.

The gist of the algorithm followed to implement part 1 is as followed:

1. The data is partitioned and kmedians function is run with the given init_centroid for max_iter and k=256, in a way that first partition is allocated with the array of first centroid.
2. L1 Manhattan distance determines the centroid in the clusters, and np.median() is used to update the codebooks.
3. In the function pq(), first the placeholder for codebooks is made. Codebooks are thengenerated for each of the kmedians.
4. A placeholder empty array for codes is created. The data vectors are split again and checked with their corresponding codebooks for the correct code values. Manhattan distance is used for this assignment.