Name: Ali Khalid Roll #: MSDS21001

MACHINE LEARNING SPRING 2022

Home Work 2

Question: In Machine learning problem, if we have a very Large number of classes, such as the NADRA database for all Pakistani Citizens. Given a data (image or fingerprint) of an unknown person, have to find its identity in the NADRA database?

ANSWER: Given the problem, one can think of using some multiclass classification algorith like Decision Tree, Random Forest, Support Vector Machine, Naural Network etc. But on further reflection, it is evident that using multiclass classification algorithms when there are million of classes might not be a good approach as there will be subtle difference between class scores and hence the prediction is prone to error.

Therefore, I suggest that instead of using multi-class classification, we should learn feature vector for each class. These feature vectors can be hand engineered or learnt from neural network (like auto encoders). Once we have learnt the feature vector (embedding) for each class, we can convert the test image or fingerprint into a feature vector of same dimensionality (using the trained encoder) and then compare it with feature vector of each class using some distance (like Euclidean) or similarity (like cosine) metric and assign it the label of class which has the minimum distance or highest similarity.

"FaceNet: A Unified Embedding for Face Recognition and clustering" is a paper by Google Researchs, published in 2015. This paper used the similar approach as described before and achieve state of the art accuracy on benchmark datasets.