UPI: yjia970
Name: Yuwei Jiang

Assignment 5 asks us to classify the dataset to distinguish genuine and forged banknotes, which is unsupervised learning. My public leaderboard score is 0.97500. The following is what I have done:

Feature processing

In order to make the data visible, I chose any two of four features and any three of four features to explore the data. Then, I noticed that when I plot V1, V3, and V4, the plot shows the data is significant linearly separable by a plate. Next, I normalized the data and used PCA method to reduce the dimension from three to two, which benefits us from choosing clustering algorithms visibly. Instead of reducing the dimension to the largest two variance variables, we need to project the data to a plate on which the data contains the maximum variance as Fig.1 shows.

Clustering algorithm

Based on Fig.1, the 2-dimensional data shows a linear separable pattern. I tried three types of clustering methods and evaluated each of them by the plots, like Fig.4, and the accuracy in leader board. I finally chose the k-means as clustering algorithm: I cluster the dataset into six classes and then gather different parts together and then get binary classes. I think Hierarchical clustering and K-means have similar effects, so, I tried both of them and the predicting accuracy of k-means is 97.5% higher than 96.39% of hierarchical clustering (Fig.2). DBSCAN algorithm is smart to distinguish bounders and noise for multiple classes classifier after choosing the right eps and min_samples. However, it clusters all the noise into a cluster as purple points on Fig.3. Therefore, it's not ideal for our binary class classifier.

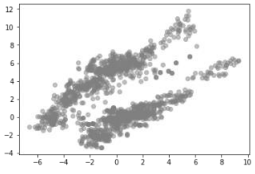

Fig.1

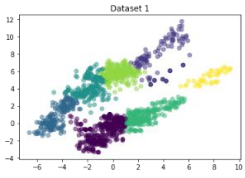

Fig.2

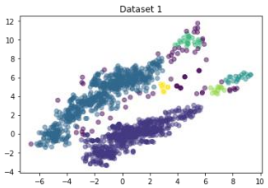

Fig.3

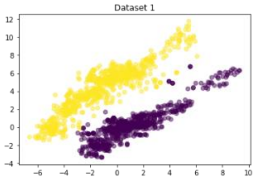

Fig.4