Homework 02

Question1:

When training the probabilistic generative classifier, how does the full covariance compare to diagonal covariance in performance for each of the datasets? Why?

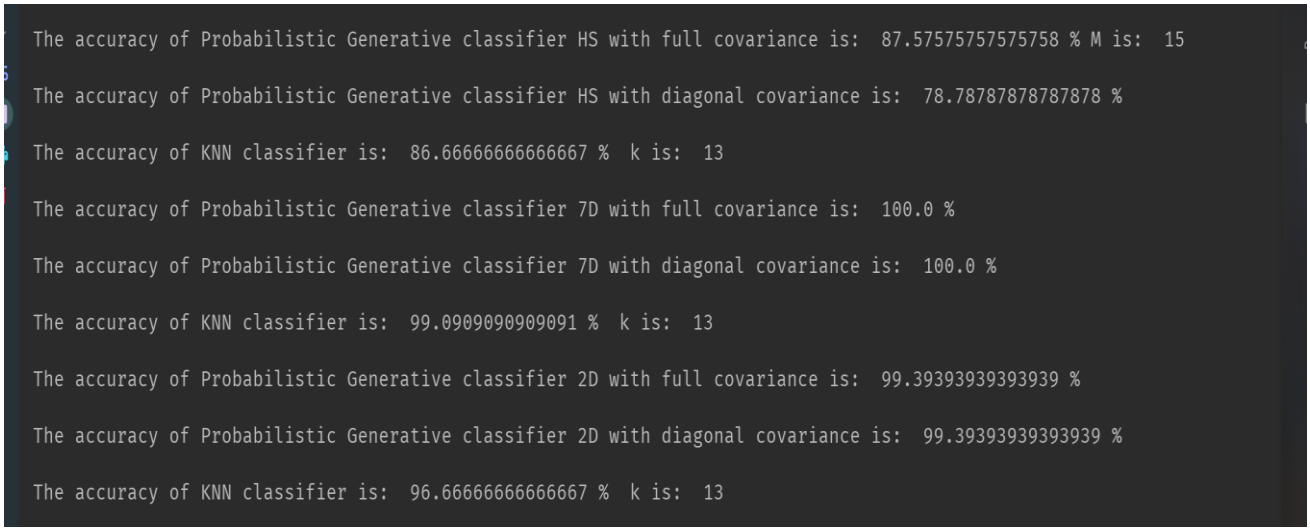

Figure 1

Figure 1 is output of my code. We can see the obviously different between full covariance and diagonal covariance for HS data. The accuracy of full covariance is about 87.6% yet that of diagonal variance is about 78.8%. So full covariance is more accurate. When we look at their definitions, we can find difference. When I use diagonal covariance, in fact I omit other elements which don't locate at diagonal and these elements shows correlation between two dimensions. So if our

dimensions' relations are large, this method can cause huge deviation. But I find for 7D and 2D, the outcome using full covariance and diagonal covariance is same. I guess maybe these data are independent with each other.

Question2:

When training KNN classifier, what happens as you vary k from small to large? Why?

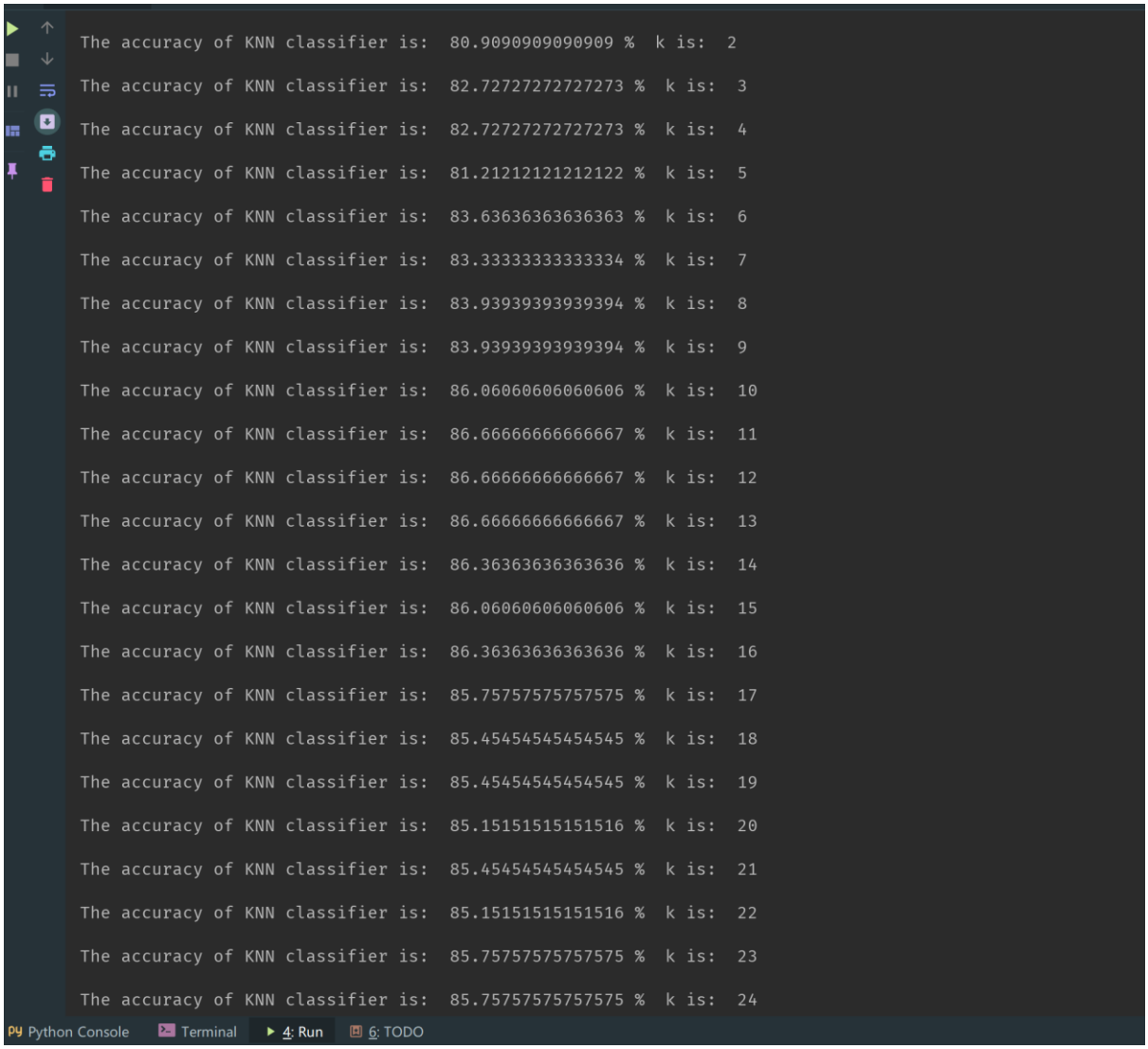

Figure 2

Figure 2 is output when I choose best K for KNN method. Because for 2D and 7D data, the accuracy is almost 100%, the difference is not apparent, here I use HS data. We can get a explicit outcome through this figure. When K change from 1 to 25, the accuracy rises at first then decreases. From the definition of KNN, when K is small, for example 1, our predicted outcome can be easily influenced by some noise or extreme point which can cause overfitting. With K increases, the probability of this situation decreases. But when K is large enough, we get another question, classification inclines to the class who has more train data. When considering extreme situation, K equals to the number of all train data, we always get the class who has most data when classification, and we cause underfitting, so the accuracy decreases.

Question3:

About different parameter setting when use cross validation?

When cross validation, I choose different M, and decide M with highest accuracy, so does the constant of diagonal matrix to solve singular matrix problem, but the loop is not used at final code because it is inconvenient. But I use different M for different dataset through comparing.

Question4:

About choose of classification method?

According to Figure 1, Probabilistic Generative Classification has higher accuracy for all 3 datasets compared to KNN. Although for HS, their accuracy is approximative because KNN is better for high dimension or intersect data, finally I follow the accuracy and choose PGC to test all 3 datasets.