3 Task3

The code for computing 3 validation indexes are below:

```
def Compute_indexs_KM(data,ground,min,max):
    print("Kmeans")
    indexs=[]
    for numc in range(min,max+1):
        KM = KMeans(n_clusters=numc)
        predicted_labels=KM.fit_predict(data)

        silhouette_avg = silhouette_score(data,
           predicted_labels)
        dh=davies_bouldin_score(data, predicted_labels)
        label_true=ground.to_numpy().reshape(1,-1)[0]
        nmi=normalized_mutual_info_score(label_true,
           predicted_labels)
        indexs.append([silhouette_avg,dh,nmi])
    return np.round(pd.DataFrame(indexs,columns=["SI","DB","NMI
       "]),3)

def Compute_indexs_SC(data,ground,min,max,method):
    print("Spectral Cluster with "+method)
    indexs=[]
    for numc in range(min,max+1):
        SC= SpectralClustering(n_clusters=numc,affinity=method)
        predicted_labels=SC.fit_predict(data)

        silhouette_avg = silhouette_score(data,
           predicted_labels)
        dh=davies_bouldin_score(data, predicted_labels)
        label_true=ground.to_numpy().reshape(1,-1)[0]
        nmi=normalized_mutual_info_score(label_true,
           predicted_labels)
        indexs.append([silhouette_avg,dh,nmi])
    return np.round(pd.DataFrame(indexs,columns=["SI","DB","NMI
       "]),3)
```

3.1 Task3.a

Abbreviations: SI=Silhouette index[1], NMI=Normalized Mutual Information, DB=Davies-Bouldin index[2].

In *balls.txt*, we preform cluster method K-means and spectral clustering using a Gaussian kernel and a Laplacian matrix with different cluster number between 2 and 5. The results are in Table5:

From Table above we can see, when $K = 3$, the clustering results are the best among

Table 3: Comparison of 3 clustering methods ($K = [2, 5]$)

method	K=2			**K=3**		
Indexs	SI	DB	NMI	SI	DB	NMI
K-means	0.668	0.523	0.735	0.902	0.136	1.000
SC (Gaussian)	0.577	0.700	0.734	0.902	0.136	1.000
SC (Laplacian)	0.668	0.523	0.735	0.902	0.136	1.000
method	K=4			K=5		
Indexs	SI	DB	NMI	SI	DB	NMI
K-means	0.710	0.653	0.905	0.519	0.941	0.832
SC (Gaussian)	0.711	0.653	0.905	0.518	0.960	0.830
SC (Laplacian)	0.708	0.657	0.904	0.514	0.976	0.827

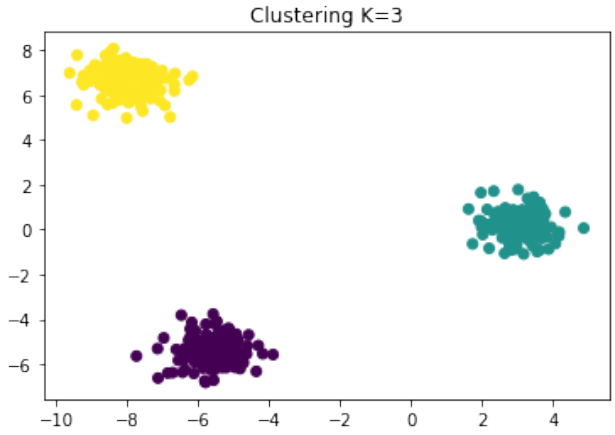

Figure 4: K-means(K=3)

different situation, and the validation index are best. The bigger SI and NMI are, the better the clustering results are. DB are opposite.we can draw the predict labels in Fig.4,5,6 for visually check. When k is not equal to 3, the image shows clearly incorrect predicted labels. So in the classification method, the predetermined K is quite important and determines the performance of the whole classification prediction. But when the amount of data is huge or the difference between the data is not obvious, it is difficult for us to judge the best k value

3.2 Task3.b

In *spirals.txt*, we preform cluster method K-means and spectral clustering using a Gaussian kernel and a Laplacian matrix with different cluster number between 2 and 5. The results are in Table5:

The NMI index can better captured the performance of Spectral Cluster with Gaussian kernel. the data in *spirals.txt* are spiral shape ,which is hard for K-means and Spectral Cluster with Laplacian to distinguish successfully. But with RBF kernel, we may get easily distinguishable results by mapping the data of the preserved features to the space of other

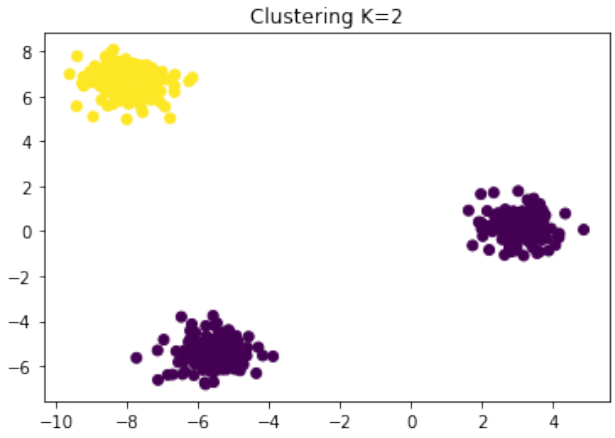

Figure 5: K-means(K=2)

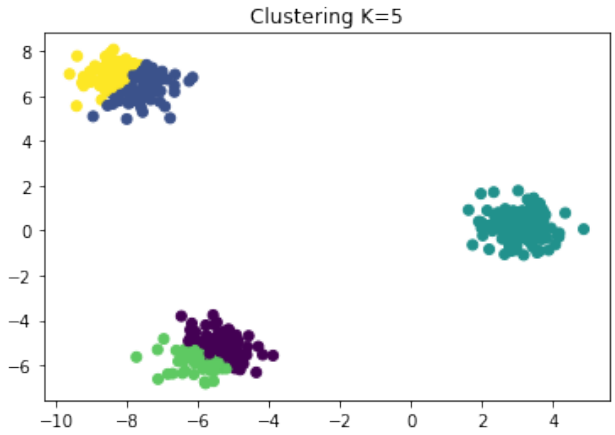

Figure 6: K-means(K=5)

Table 4: Comparison of 3 clustering methods ($K = [2, 5]$)

method	K=2			**K=3**		
Indexs	SI	DB	NMI	SI	DB	NMI
K-means	0.348	1.168	0.001	0.360	0.880	0.000
SC (Gaussian)	0.025	6.314	**0.729**	0.001	5.882	**1.000**
SC (Laplacian)	0.345	1.173	0.000	0.362	0.896	0.002
method	K=4			K=5		
Indexs	SI	DB	NMI	SI	DB	NMI
K-means	0.354	0.881	0.003	0.347	0.895	0.009
SC (Gaussian)	-0.012	6.914	**0.910**	0.015	5.459	**0.833**
SC (Laplacian)	0.330	0.927	0.052	0.279	1.818	0.134

dimensions. The graphs in Fig.7,8,9 shows the clustering results.

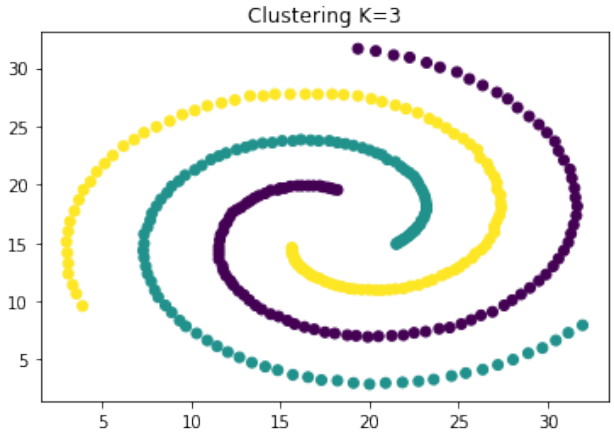

Figure 7: Spectral Cluster with RBF(K=3)

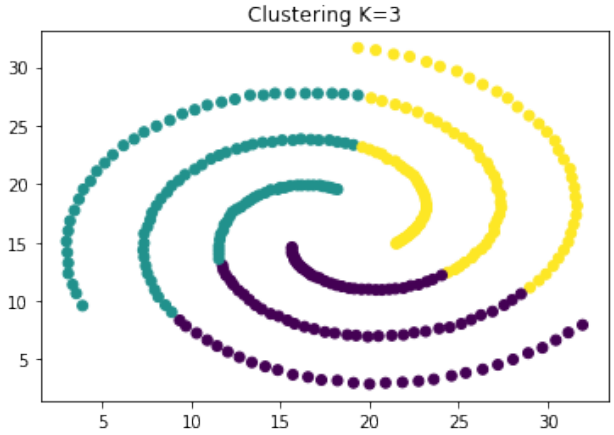

Figure 8: Spectral Cluster with Laplacian(K=3)

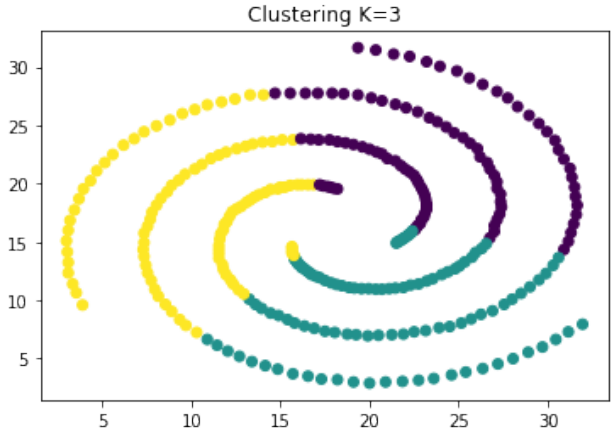

Figure 9: K-means(K=3)

3.3 Task3.c

1. SI is a measure of how similar an object is to its own cluster compared to other clusters. The silhouette ranges from -1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. We average SI on all points to show the performance in general. If most points have a high value, then the clustering configuration is appropriate.

2. DH is defined as the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances. Clusters which are farther apart and less dispersed will result in a better score. The DB index is generally higher for convex clusters than other concepts of clusters, such as density based clusters. The usage of centroid distance limits the distance metric to Euclidean space.

3. NMI is a good external measure for determining the quality of clustering, we need the class labels of the data to determine. Since it's normalized we can measure and compare NMI between different clusterings having different number of clusters. Mutual Information tells us the reduction in the entropy of class labels that we get if we know the ground cluster labels. But external measures require the knowledge of the ground truth classes while almost never available in practice or requires manual assignment by human annotators.

NMI score in spectral cluster with gaussian kernel perform better in data of spiral shape, we may get easily distinguishable results by mapping the data of the preserved features to the space of other dimensions. NMI is an external index and gain more information, in which the comparison are between the ground truth labels and predicted labels. But SI and DB are both internal indexes, which mean the positions of data themselves are also important. They measure the performance based on the distance between data and similarity between clusters.So in plot of spirals, even though the correct classification result is obtained using spectral cluster with gaussian kernel, because the aggregation of points inside the class is not gathered together as in the ordinary case, but presents a curve. This leads to a completely wrong evaluation result if following SI and DB. Clusters which are farther apart and less dispersed will result in a better score. We can not use internal indices to determine optimal K for spectral clustering. In the exercise, the internal indices performed wrong evaluation on spiral data.