Learning From Data
Analysis on Programming Assignment 3

Feng Zhao zhaof17@mails.tsinghua.edu.cn

12/11/2020

KMeans Assumption

Problem

Without normalization, use kmeans to cluster the following data and analyze your unexpected result.

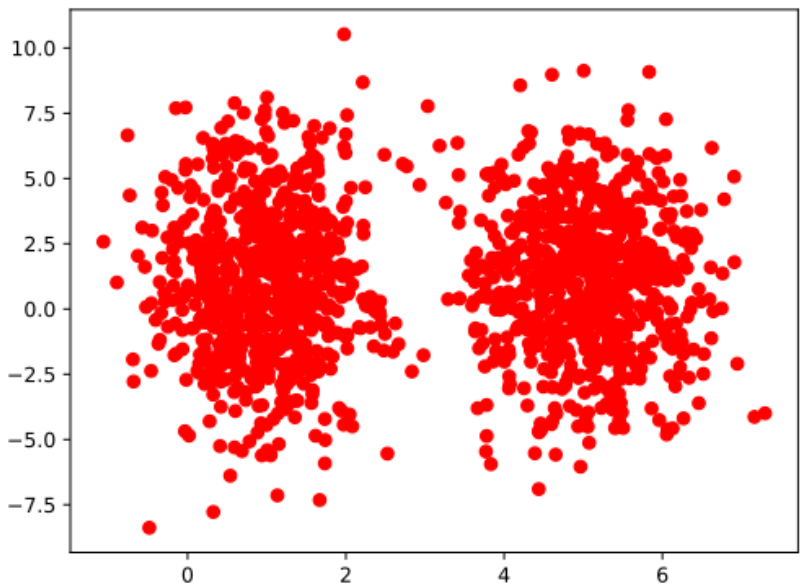


Figure: data blob with two-ellipse contour

KMeans Assumption

Using random initialization:

Result

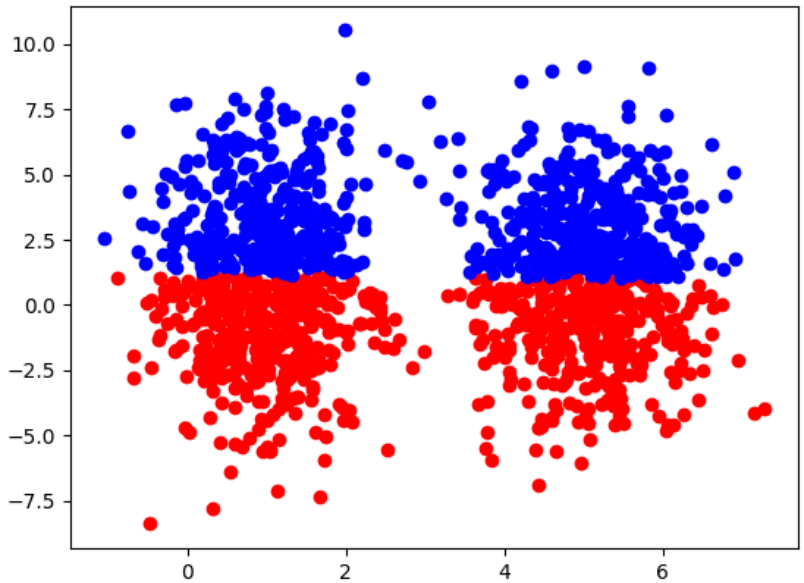


Figure: unexpected clustering result with inertia equal to 11175

inertia of kmeans: $\min_{C,\mu} \sum_{j=1}^{k} \sum_{x \in C_j} ||x - \mu_j||^2$

KMeans Assumption

Choose the initial centroid of KMeans near `[1, 0]`, `[5, 0]`:

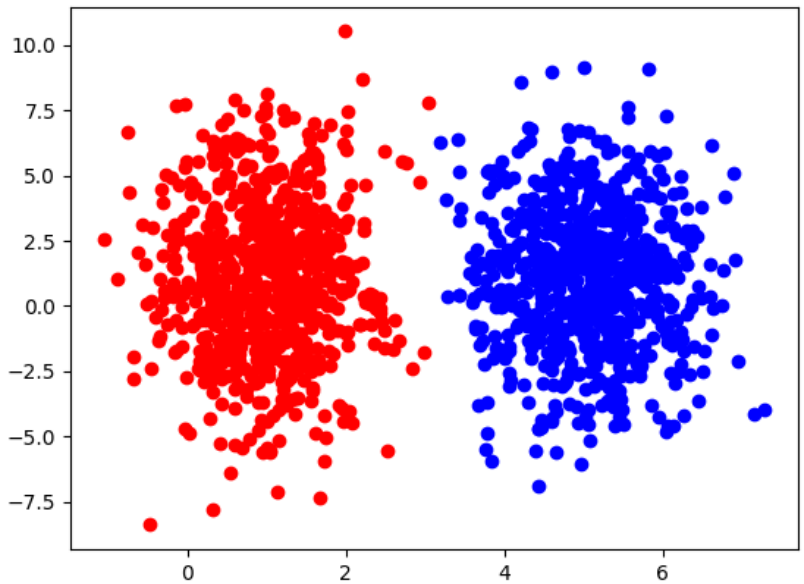
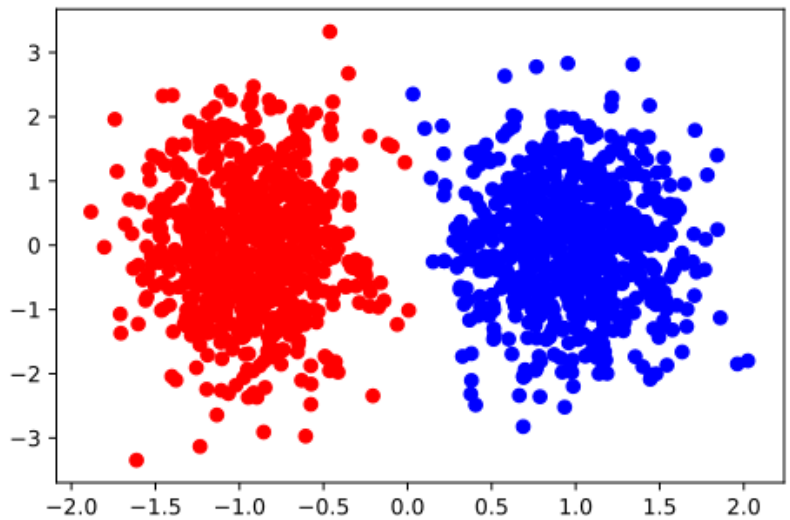


Figure: expected clustering result with inertia equal to 12771

expected result is the local optima (**not** the global optima)

KMeans Assumption

Normalizing the data before using KMeans:
`sklearn.preprocess.scale`



- Major factor: unit variance assumption of KMeans
- Minor factor: random initialization

Gaussian Kernel in Spectral Clustering

Problem

The similarity matrix W is given by $W_{ij} = \exp(-\gamma||x_i - x_j||^2)$. Explore how γ influence the spectral clustering result of the following data:

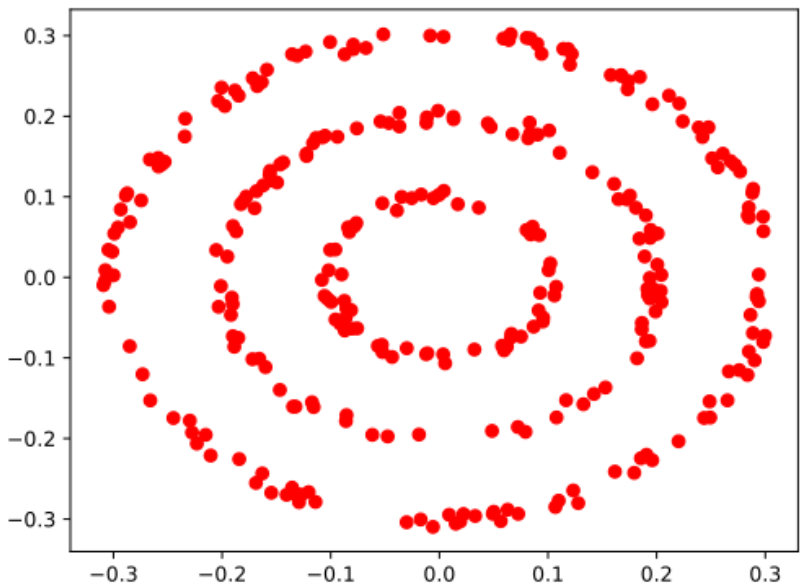
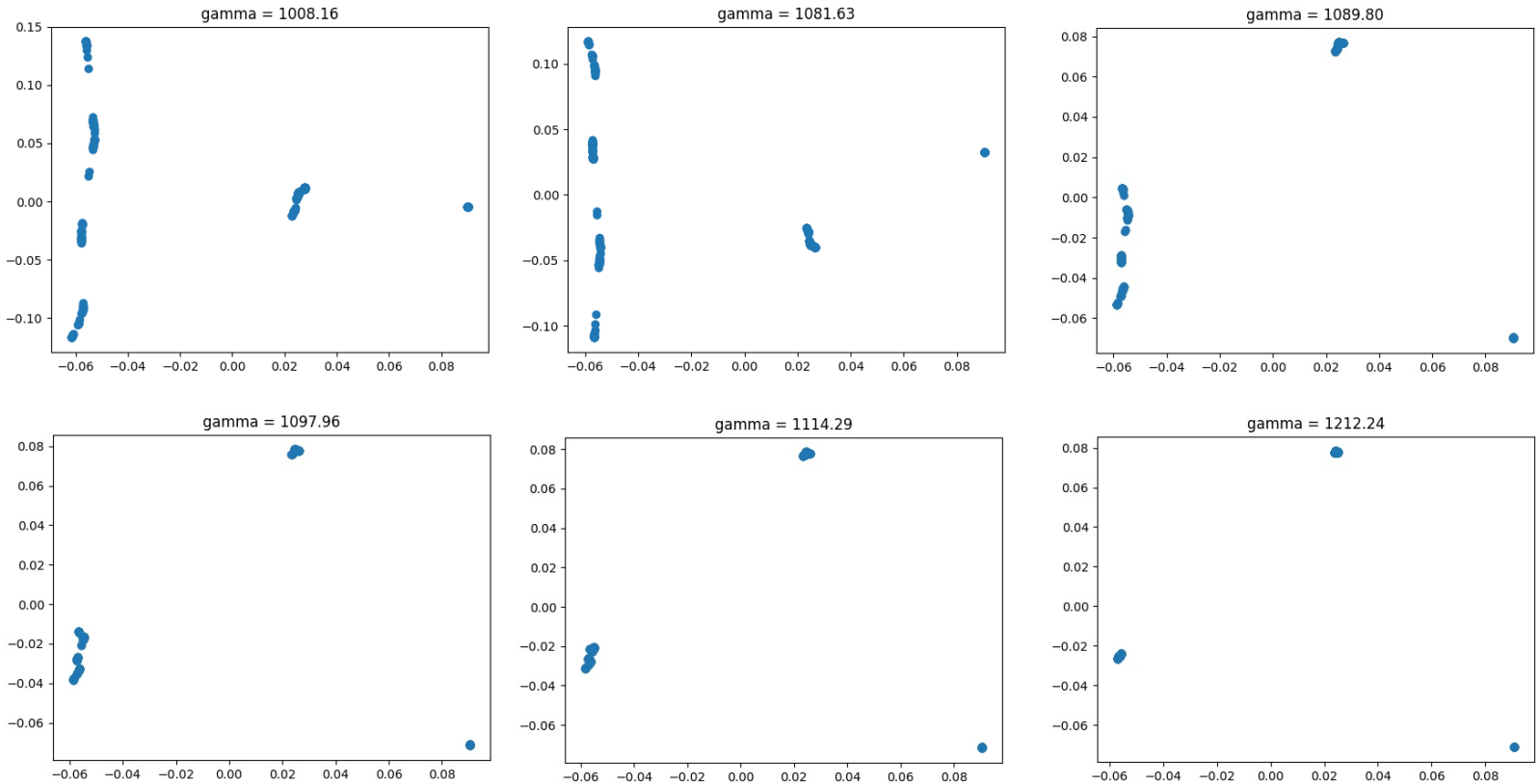


Figure: three-circle dataset

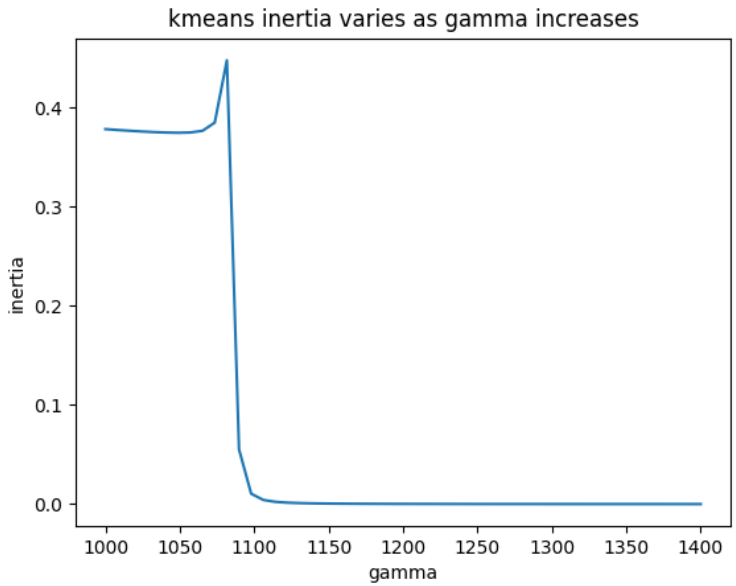
Gaussian Kernel in Spectral Clustering

When γ increases from 1000 to 1400, draw the embedded features (the second and third smallest eigenvector) in Euclid plane:



The transition occurs when *gamma* changes from 1081 to 1089.

Gaussian Kernel in Spectral Clustering



The inertial drops to zero in a neighborhood of $\gamma = 1086$.

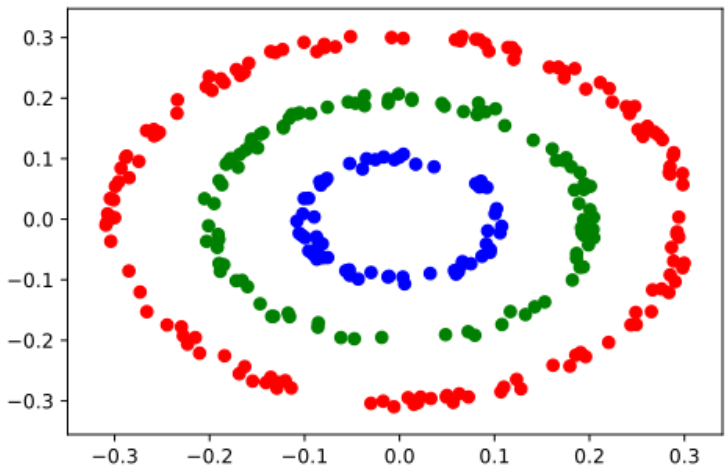
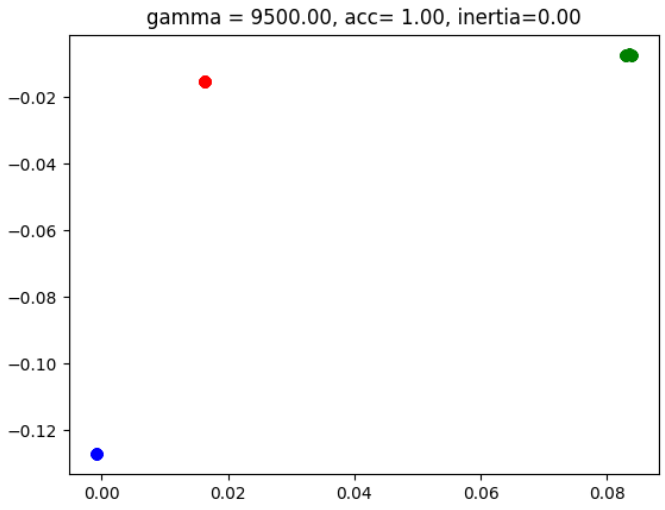
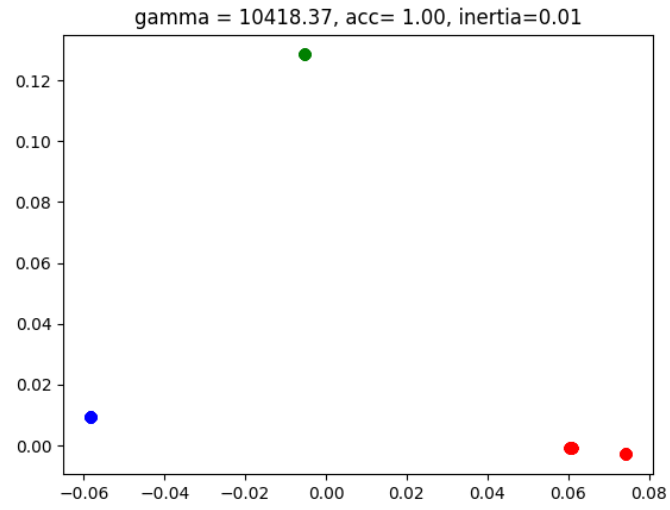
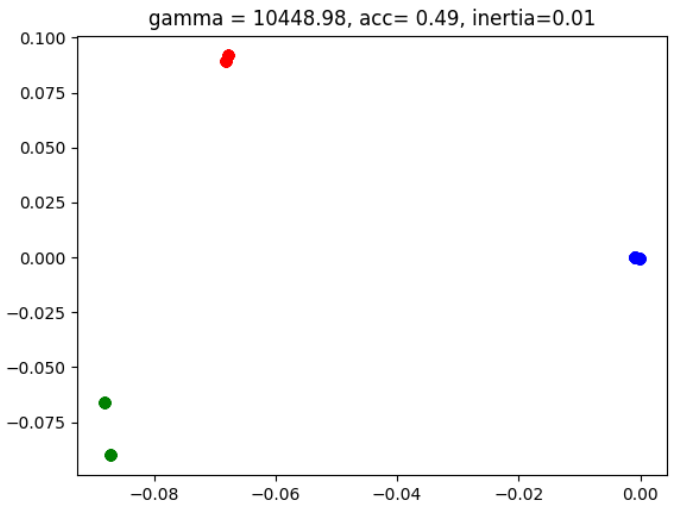


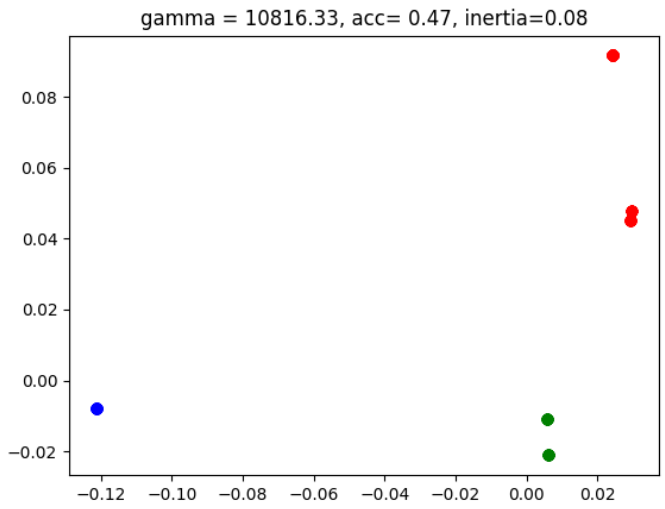
Figure: spectral clustering result when $\gamma = 1100$

Gaussian Kernel in Spectral Clustering









Gaussian Kernel in Spectral Clustering

The number of clusters is more than 3 when *gamma* is larger than 10418. Using $k = 3$ will produce incorrect result:

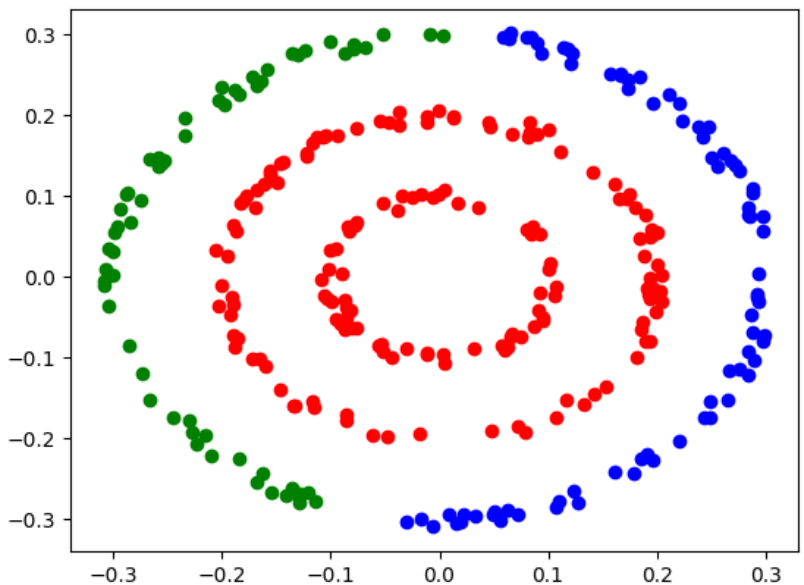


Figure: spectral clustering result when $\gamma = 10448$

Gaussian Kernel in Spectral Clustering

Conclusion

- $\gamma < 1086$: similarity between different clusters are larger than that of the same cluster
- $\gamma > 10418$: similarity within the same cluster is smaller than that between different clusters
- $\gamma \in (1086, 10418)$: Using proper eigen-decomposition and kmeans initialization strategy can achieve accuracy 100%