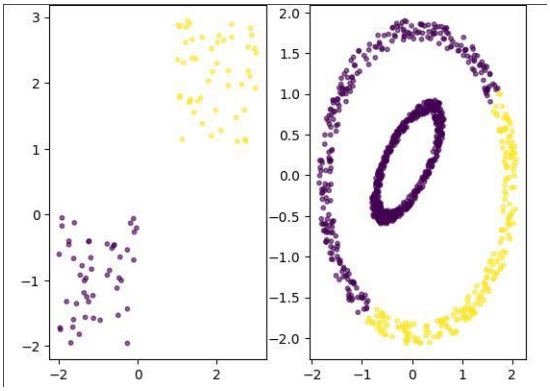
Marcantonio Soda
Project 1 Report
CSE447
12 Mar 2023

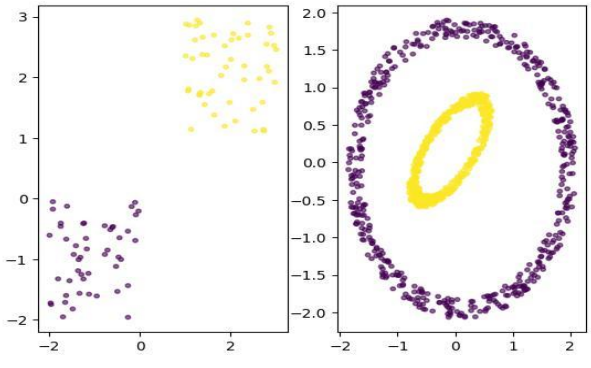
Square and Elliptical Datasets

Implement k-means and spectral clustering algorithms to find 2 clusters on Square and Elliptical datasets and visualize your results. Compare the two methods and discuss their pros and cons.

- K-means and spectral clustering are two different methods by which one can quantify clusters in data. K-means is simpler, faster, and better (in terms of speed) for high-dimensional data than spectral clustering. If good initial cluster centers are chosen, the algorithm is likely to converge much faster (more on this later). Spectral clustering, however, is capable of detecting non-linear structures whereas k-means is incapable of clustering non-linear data. The elliptical dataset is a perfect example of this. K-means does a terrible job of clustering it because the data is non-linear, but spectral clustering does it perfectly. The square dataset is clustered appropriately by both algorithms, but k-means is likely the better option because it is less resource intensive. My understanding is that if k-means works, it should be chosen over spectral clustering.



K-means



Spectral clustering

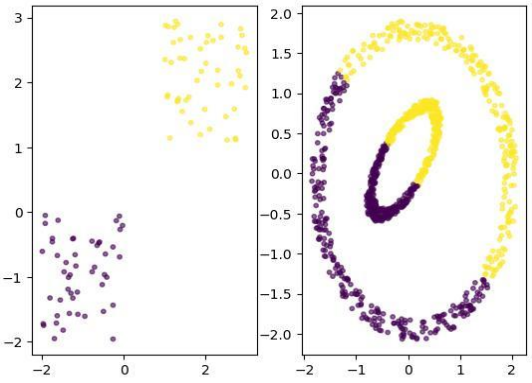
Discuss the effects of centroid initialization on n on k-means clustering results.

- Centroid initialization is pivotal for clustering with k-means. A poor choice of initial centroids can cause the algorithm to take a long time to converge or even cause it to converge incorrectly. Random initialization of centroids is common practice, but can lead to suboptimal results and performance. Another popular technique is called k-means++ that uses a probability

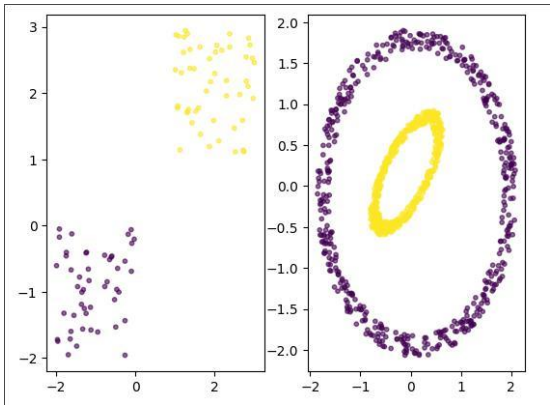
distribution to find initial centroids. It is likely that choosing random centroids is going to take less time than any other centroid initialization mechanism, but spending a little extra time choosing the initial centroids drastically reduces the runtime and error rate of the algorithm.

Present the performance analysis of the spectral clustering algorithm using different similarity measures like cosine similarity and Gaussian kernel similarity (set an appropriate bandwidth parameter for Gaussian kernel)

- Gaussian kernel similarity is superior to cosine similarity for spectral clustering for the elliptical dataset. Both accurately cluster the square dataset. This is likely because Gaussian kernel similarity is better for clustering non-linear data structures than cosine similarity.



Cosine Similarity

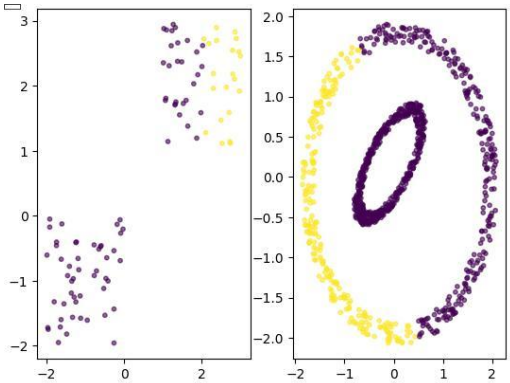


Gaussian Kernel Similarity

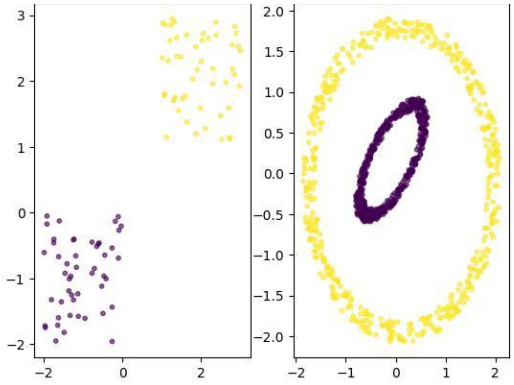
Present the performance analysis of the spectral clustering algorithm using different Laplacian matrices like unnormalized Laplacian and normalized symmetric Laplacian

- The unnormalized laplacian matrix seems to be superior to the normalized laplacian matrix for spectral clustering with both the elliptical and square datasets. The unnormalized laplacian matrix clusters both perfectly, but the normalized one fails on both accounts.

Normalized



Unnormalized



Cho and Iyer Datasets

Validate your clustering results:

- External Index:
    - The accuracy measure for k-means was ~65% for cho and ~50% for Iyer. The accuracy measure for spectral clustering was also ~65% for cho and ~50% for k-means. These results show that there is no clear superior clustering algorithm when it comes to Iyre and Cho.
- Internal Index:
    - The SSE for k-means was ~464 for cho and ~1453 for Iyer. The SSE for spectral clustering was ~461 for cho and ~1510 for Iyer. These results show that there is no clear superior clustering algorithm when it comes to Iyre and Cho.

Discuss the impact of data normalization on `k`-means and spectral clustering results based on Cho and Iyer datasets in terms of clustering accuracy (i.e., Accuracy).

- The accuracy before and after normalization for cho remained about the same with the normalized data having slightly higher accuracy for both k-means and spectral clustering. For Iyer, interestingly, the accuracy was 4% less in k-means for the normalized data. Spectral clustering accuracy remained very close. I believe this is because the gaussian kernel similarity is not very affected by the data being normalized because it bounds euclidean distance between 0 and 1.

Discuss the impact of noise on `k`-means and spectral clustering results based on Iyer dataset in terms of clustering accuracy. Namely, compare the results with and without noise data.

- As you would expect, the removal of noise increased accuracy for both k-means and spectral clustering. The increase was about 3% for k-means and about 4% for spectral. Since cho has no noise, it cannot be compared for this test.