PROBLEM 1: PCA & FEATURE SELECTION

PART-1 SVM & PCA

BACKGROUND :

PCA provides a roadmap for how to **reduce a complex dataset to a lower dimension** to reveal the sometimes hidden, simplified dynamics that often underlie it.

PCA is related to mathematical technique SVD (Singular Value Decomposition)

GOAL : To determine the **most meaningful basis** to re express a noisy, garbled data

PX = Y

P {p1, p2, p3 ……pm}, X, Y are (mxn) matrices

Above equation represents **change of basis**

1. P is a matrix that transforms X into Y
2. Geometrically, P is a **rotation and a stretch** which transforms **X into Y**
3. Rows of P are set of new basis vectors for expressing the columns of **X** thus row vectors in this transformation will become the **principal components** of X

Covariance matrix describes all relationships between pairs of measurements in dataset
In order to reduce **redundancy** we would like each variable to co-vary as little as possible with other variables (Ideal choice of covariance matrix to be a diagonal matrix)

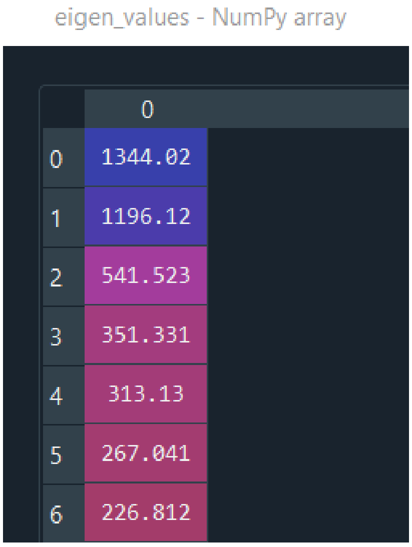
PCA assumptions and limits:

1. Linearity
    a. It frames the problem as change of basis
2. Mean and variance are sufficient statistics
    a. Mean and variance entirely describe a probability distribution
    b. Guarantees that SNR (Signal noise ratio) and covariance matrix fully characterize the noise and redundancies
3. Large variances have important dynamics
    a. Assumption encompasses the belief that the data has a high SNR
    b. Principal components with large associated variances represent interesting dynamics
4. Principal components are orthogonal
    a. Simplification that makes PCA solvable with linear algebra decomposition techniques

SOLUTION :

Find P such that Y = PX and S_Y = (1/n-1) $Y^{*}Y^{T}$ is a diagonal matrix.
We select matric P to be a matrix where each row p_i is an eigenvector of $X^{*}X^{T}$

1. Top six eigen values for data covariance matrix



eigen_values - NumPy array

	0
0	1344.02
1	1196.12
2	541.523
3	351.331
4	313.13
5	267.041
6	226.812

1. Validation errors corresponding to different combinations of K and C are

K/C	1	10	100	1000
1	46.1538	46.1538	46.1538	46.1538
2	30.7692	30.7692	30.7692	30.7692
3	21.1538	21.1538	21.1538	21.1538
4	21.1538	21.1538	21.1538	21.1538
5	25	25	25	25
6	26.9231	26.9231	26.9231	26.9231

1. Test errors for different combinations of K and C are

K/C	1	10	100	1000
1	59.6154	59.6154	59.6154	59.6154
2	44.2308	44.2308	44.2308	44.2308
3	19.2308	19.2308	19.2308	19.2308
4	19.2308	19.2308	19.2308	19.2308
5	17.3077	17.3077	17.3077	17.3077
6	19.2308	19.2308	19.2308	19.2308

Best k/c on validation data corresponds to the values highlighted in green and their corresponding test errors are highlighted below

Best K/C validation error : **21.1538**
Best K/C test error : **19.2308**

Best classifier without feature selection for the same values of C are as follows

c	TRAININIG ERROR	VALIDATION ERROR	TEST ERROR
1	22.1154	23.077	26.9231
10	9.6154	11.5385	**15.3847**
100	4.8077	23.077	23.077
1000	0	19.231	23.077

Best test error : **15.38** obtained for c = 10

OBSERVATIONS :

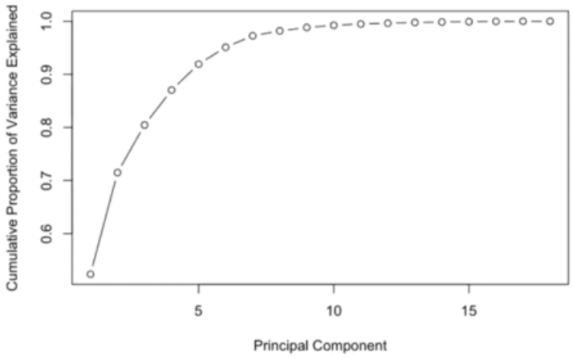
a. SVM without feature transformation **outperforms** SVM with feature selection
b. As the number of eigen vectors taken into consideration for fitting an SVM there is almost **consistent drop in the test error rate**.
c. It seems like k = 6, taking six eigen values is **able to capture most of the information** carried by the dataset.

4. Picking k before evaluating performance on the validation data for an unsupervised setting
As mentioned in the background, PCA assumes large variances have important dynamics and tries to find them so possible options can be following :

a. PERCENTAGE OF VARIANCE EXPLAINED

i. (Eigen value/ Sum of eigen values) gives **percentage of variance explained by it**.
ii. Every eigen value from the top adds its contribution to the variance
iii. If **cumulative percentage of variance explained by k eigen values** is greater than some threshold (lets say 90%) then we fix on that k

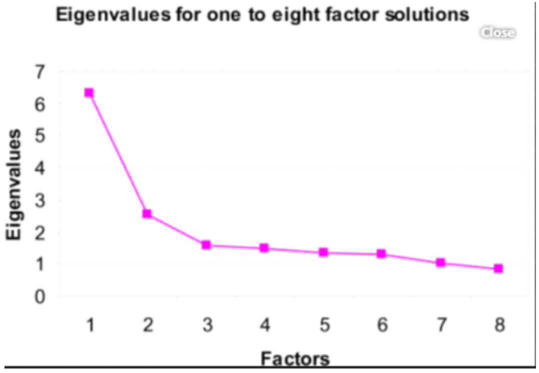
1)



b. SCREE TEST

i. Pick the number of components where slope starts levelling (Pick 3 here)

1)



INTERESTING LINEAR ALGEBRA RESULTS

1. The inverse of an orthogonal matrix is its transpose
2. A^*A^T is symmetric
3. A matrix is symmetric if and only if it is orthogonally diagonalizable
4. A symmetric matrix is diagonalized by a matrix of its orthonormal eigen vectors

Above results may be used in proving few questions

PCA FOR FEATURE SELECTION

1. Probability distribution proof

Computing top (k) eigenvalues and eigen vectors for COVARIANCE MATRIX

$$\Rightarrow \begin{pmatrix} | & | & | & & | \\ v^{(1)} & v^{(2)} & v^{(3)} & \cdots & v^{(k)} \\ | & | & | & & | \end{pmatrix} \Rightarrow \pi = \begin{pmatrix} \frac{1}{k}\sum_{i=1}^{k} v_1^{(i)^2} \\ \frac{1}{k}\sum_{i=1}^{k} v_2^{(i)^2} \\ \vdots \\ \frac{1}{k}\sum_{i=1}^{k} v_k^{(i)^2} \end{pmatrix}$$

for π to be a valid probability distribution it should satisfy

each of $\pi_j \geq 0$ & $\sum_{j=1}^{n} \pi_j = 1$

i) As each of π_j is sum of squares of components.

i.e $\pi_j = \sum_{i=1}^{k} v_j^{(i)^2}$ → [SUM OF SQUARED TERMS ≥ 0]

Squared terms

ii) $\sum_{j=1}^{n} \pi_j = 1 \Rightarrow \sum_{j=1}^{n} \left(\frac{1}{k}\right) \sum_{i=1}^{k} v_j^{(i)^2}$

For covariance matrix (Symmetric) eigen vectors are orthonormal

$$= \left(\frac{1}{k}\right) \sum_{i=1}^{k} \left(\sum_{j=1}^{n} v_j^{(i)^2} \right) = \frac{1}{k} \sum_{p=1}^{k} \left(\|v^{(i)}\|^2 \right)$$

$\Rightarrow \boxed{\|v^{(i)}\|^2 = 1}$

$$= \frac{1}{k}(k) = \boxed{1} \checkmark$$

So, π corresponds to a valid probability distribution

2. AVG VALIDATION ERRORS

S\K	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	39.0769	38.8654	39.5192	38.0962	40.8846	38.9038	39.4231	39.6923	39.5192	39.6538
2	35.7692	37.1923	36.4038	36.2308	35.8654	36.5769	37.3077	35.7115	37.3654	36.5385
3	33.7308	35.1538	33.1731	33.1346	33.4231	34.2115	34.5769	35.3269	33.7692	34.2308
4	32.5192	33.5577	33.1731	32.0577	31.3846	31.8846	33.1731	31.9038	31.9808	32.6154
5	31.25	30.6154	30.75	30.1923	30.6154	30	32.0962	31.0577	30.9808	31.1346
6	30.9231	30	30.4615	29.0769	29.8654	29.5962	30.1731	29.6538	29.1538	29.5385
7	30.6731	30.0192	28.6154	27.7692	28.9231	28.3269	28.9808	29.5	27.9231	28.5769
8	31.0962	29.3654	29.1731	28.3077	27.3846	28.6154	28.2115	28.1538	28.0385	28.8269
9	29.8846	28.1346	27.4423	27.9038	26.4231	27.2115	27.3462	27.3846	27.3846	26.9808
10	29.6346	27.8077	26.7692	27.5577	26.9615	26.9615	26.4038	26.7308	26.8269	26.8077
11	29.4423	27.5385	27.25	26.7692	25.6346	26.6923	26.3077	25.7885	26.6346	26.3846
12	29.4808	26.7692	26.4231	26.3077	24.6346	24.5577	25.8654	26	26.5385	27.1154
13	27.7115	27.0769	25.7692	24.8269	24.8654	24.8654	25.4808	24.9423	25.0385	27.0962
14	28.4615	26.0385	26.7885	24.9808	24.7308	25.8846	26.0385	25.3846	25.6346	24.9038
15	27.2885	25.3077	25.6154	24.6154	24.6923	25.4038	25	24.8654	23.8269	26.0577
16	27.9808	25.2115	25.1731	24.9423	23.8269	25.25	24.2692	23.4615	24.8077	25.2308
17	27.8654	24.5962	24.9231	24.1731	24.9038	24.9808	23.9038	24.4231	24.4808	24.4423
18	27.3269	25.0385	25.2885	23.4038	24.1154	23.4615	25.4615	23.9423	24.3654	24.5385
19	27.1538	26.25	25.1154	24.4615	24.5577	24.4423	23.8462	24.0385	23.1154	24.0577
20	27.6731	24.0577	24.1731	23.8654	23.5962	23.1731	**22.8846**	23.9038	23.75	24.75

Best validation error for all combinations of (k, s) is obtained at k = 7 and S = 20

AVERAGE TEST ERRORS

S\K	1	2	3	4	5	6	7	8	9	10
1	44.1346	43.4808	42.5962	39.6538	41.9615	41.8654	43.3462	43.6538	43.3654	42.9423
2	39.4423	39.6538	38.4038	36.0577	36.3269	36.1923	38.0385	36.6923	38.6346	39.4808
3	35.7308	36.75	33.0962	31.7885	32.8269	33.2308	35.25	35.5	34.4808	35.4423
4	33	34.5577	31.5769	30.4423	31.0192	30.3846	34.0577	32.3654	33.0577	32.6154
5	33	32.8462	29.5385	27.4423	28.5385	30.4231	29.7885	30.5	31.1731	31.1923
6	32.5962	31.6538	28.3654	27.4808	27.1346	29.6154	29.0577	29.0962	27.9038	29.4231
7	31.1538	32.0385	28	27.0192	26.5192	27.7115	28.5769	28.0192	28.5385	28.5769
8	31.9038	29.4808	26.9231	26.2885	26.2885	27.1346	26.6346	28.8846	27.0962	28.0385
9	30.7115	28.9615	27.4615	25.6346	25.4615	26.4808	27.6731	27.4231	28.4615	27.6731
10	30	28.5385	26	24.7885	25.2308	25.4808	26.6538	27.1923	27.4231	27.25
11	29.25	29.1538	26.4615	24.8269	25.1923	26.0192	26.8269	26.9038	25.7692	26.4808
12	29.25	28.0769	26.4231	24.5577	23.9423	25.2115	25.9808	25.5385	26.2308	27.0769
13	28.8077	26.6154	25.6538	23.4038	25.5385	24.3269	26.0769	26.1538	26.5962	27.6154
14	28.9231	27.2885	25.3269	24.8462	24.7885	25.75	25.3077	26.6154	25.9615	26.2885
15	29.2115	27.5192	25.2308	23.8269	24.5577	24.9615	25.5385	25.7115	26.4423	26.9615
16	27.7308	27.6346	24.9808	24.3846	24.6538	24.7885	25.0385	26.2115	25.2885	25.7308
17	29.2115	26.3077	25.4808	23.9808	24.8846	24.0769	26.4038	25.8654	25.3654	26.4231
18	26.9423	26.8462	24.5962	24.4038	24.1346	25	24.9808	25.4423	25.3846	25.5962
19	27.4038	27.3462	23.5	22.5769	23.7692	24.5962	25.0577	24	25.5577	25.0192
20	27.7115	26.3846	24.8654	23.7885	24.1731	24.5385	24.4615	24.7692	25.3269	25.8077

Best validation error : (k, s) = (7, 20)
Test error for the combination of (k, s) = (7, 20) = 24.4615

Although least test error is obtained at (4, 19) = 22.57

3. Is this a reasonable alternative to SVM with slack formation without feature selection ?

    Best average test error : **24.4**
    Best test error to SVM with feature selection : **15.384**

    Considering the accuracies calculated for this dataset, **I don't think it is reasonable** to use this instead of SVM without feature selection.

    Eigen vectors are linear combinations of original feature vectors. By defining the probability distribution as above we are trying to weigh important features more.
    Sampling using the above resulting distribution allows to include these important features with higher probability in the learned model

    PROS :

1. As we are restricting the number of features, **model training can be faster.**
2. With the determination of features that explain data, model can turn out to be **more interpretable**

    CONS :

1. **Grid search** for best combination of k and s can be **time consuming.**
2. Less features can impact performance
3. As it based on probability to derive best results for each combination of (k,s) experiment has to be performed many times (similar to suggestion in the problem to perform 100 times)

PROBLEM 2 SPECTRAL CLUSTERING

PART-1 THE BASIC ALGORITHM

1. Arguing Laplacian Matrix is positive semidefinite

POSITIVE SEMIDEFINITE : A symmetric $n \times n$ real matrix M is said to be positive semi definite if for any non-zero column vector z of n-real numbers. $(z^T M z \geq 0)$

$$A_{ij} = A_{ji} = e^{-\frac{1}{2\sigma^2}\|x_i - x_j\|^2}$$

$$\boxed{A_{ij} \geq 0}$$

$$D_{ii} = \sum_j A_{ij} \rightarrow D \text{ (SYMMETRIC)}$$

Consider Laplacian matrix (Difference of two symmetric matrices → SYMMETRIC)

$$L = D - A$$

Similarity matrix

Diagonal matrix

SYMMETRIC (A)

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ & & \vdots \; \cdots \end{pmatrix} \qquad \begin{pmatrix} (a_{11}+a_{12}+\cdots) & 0 & 0 \\ 0 & (a_{21}+a_{22}+\cdots) & 0 \\ 0 & 0 & \ddots \end{pmatrix}$$

$$\Rightarrow z^T L z \qquad (z^T L z)^T = (Lz)^T z^{T^T}$$
$$= z^T L^T z = \underline{z^T L z}$$

$$= z^T (D - A) z$$

$$= z^T D z - z^T A z \quad \text{—} \; \textcircled{1}$$

Consider $\begin{pmatrix} a \\ b \end{pmatrix}^T \begin{pmatrix} p & 0 \\ 0 & q \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$

$$= \begin{pmatrix} ap & bq \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \underline{pa^2 + qb^2}$$

$$\begin{pmatrix} a \\ b \end{pmatrix}^T \begin{pmatrix} p & q \\ r & s \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$
$$= \begin{pmatrix} ap+br & aq+bs \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$
$$= \left[a^2p + abr + abq + b^2s\right]$$

$$= \sum_{i=1}^{n} D_{ii} (z_i)^2 - \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} (z_i)(z_j)$$

$$= \sum_{i=1}^{n} \left(\sum_{j=1}^{n} A_{ij}\right)(z_i)^2 - \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} (z_i)(z_j)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} z_i^2 - \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} (z_i)(z_j)$$

Using symmetry of $[A_{ij} = A_{ji}]$

$$= \frac{1}{2}\left(\sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} z_i^2 - 2\sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} z_i z_j + \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ji} z_j^2\right)$$

$$= \frac{1}{2}\left(\sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} \left(z_i^2 - 2 z_i z_j + z_j^2\right)\right)$$

$[A_{ij} \geq 0]$

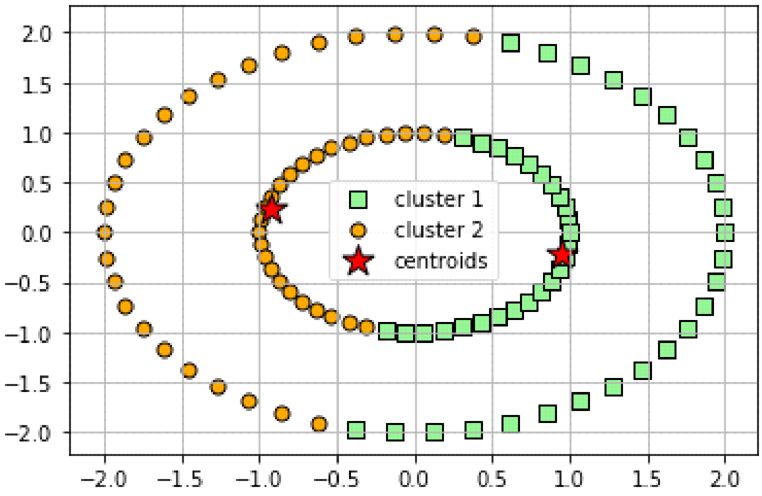
$$= \frac{1}{2}\left(\sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} (z_i - z_j)^2\right) \Rightarrow \geq 0$$

[SQUARES]

(THUS POSITIVE SEMIDEFINITE)

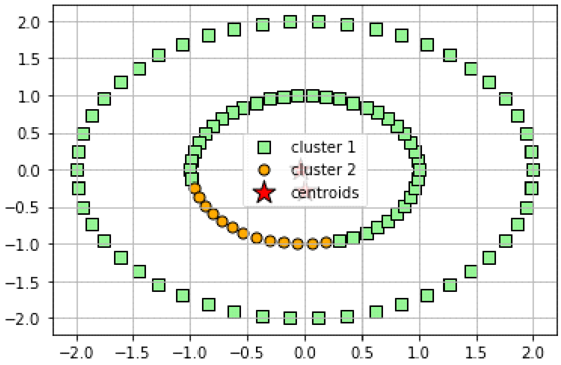
PART-2 A SIMPLE COMPARISON

2. KMEANS CLUSTERING

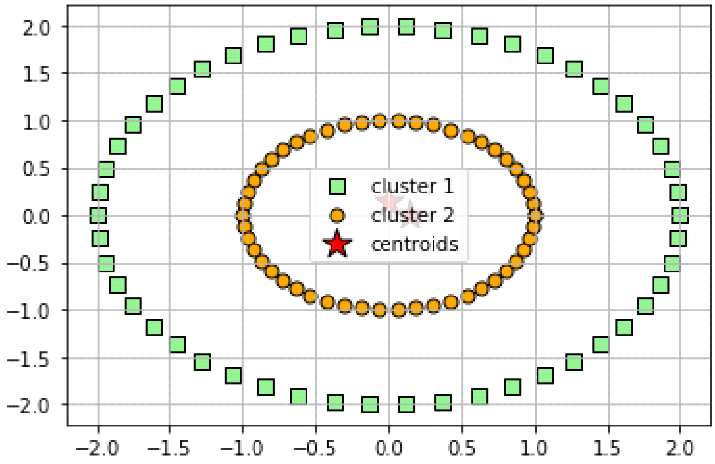


SPECTRAL CLUSTERING FOR DIFFERENT VALUES OF SIGMA

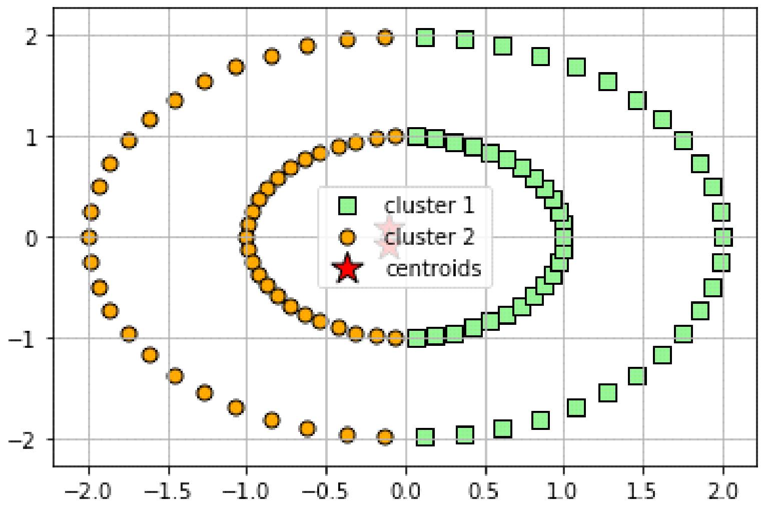
SIGMA = 0.01



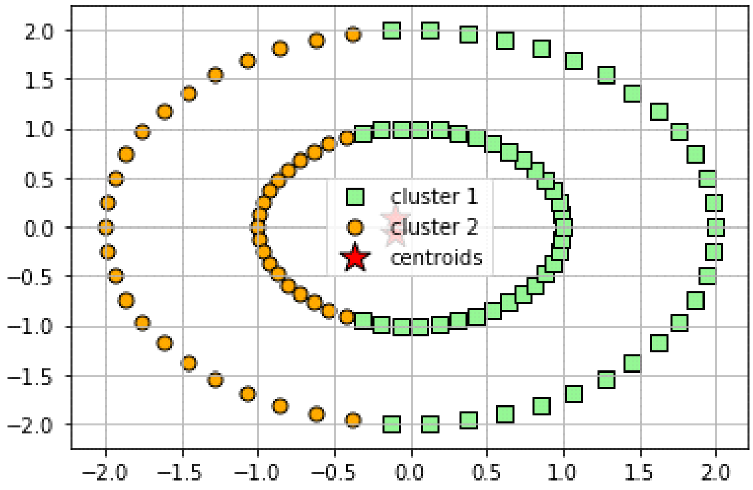
SIGMA = 0.1



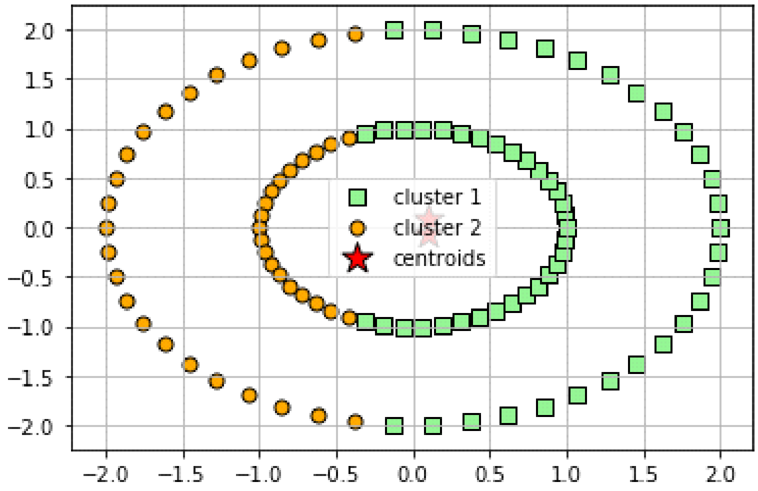
SIGMA = 1



SIGMA = 10



SIGMA = 100



3. Performing a grid search over sigma's for the circle dataset shows

   sigma = 0.1 with spectral clustering produces much better clustering algorithm

   KMEANS

   a. It tries to separate samples into groups of equal variance minimizing **INERTIA** or **WITHIN CLUSTER SUM OF SQUARES**
   b. Inertia makes the assumption that clusters are **CONVEX** and **ISOTROPIC.**

   GOAL : To prove there exists no choice of centres and clusters that performs as good as spectral clustering in categorizing two concentric ellipses into two groups.

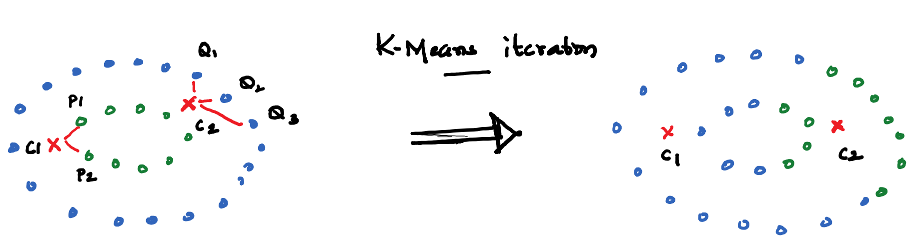
   PROOF BY CONTRADICTION :

   Assume there exists a choice of two centres which performs the same kind of clustering as spectral minimizing withing cluster sum of squares.

   Let the choice be C1 and C2 for clusters.
   C1 is the cluster describing points on outer ellipse and C2 for points on inner ellipse.

   Observations :

   a. For cluster centre selection any point outside the ellipse does not minimize within cluster sum of squares. So the **cluster centre has to be inside** the chosen ellipse.
   b. Now if cluster centre is inside the chosen ellipse then **there exists at least one point** on the other ellipse which when included in this cluster further minimizes the overall within cluster sum of squares.



DISTORTING CLUSTER CHOICE BY MINIMIZING WITHIN CLUSTER SUM OF SQUARES

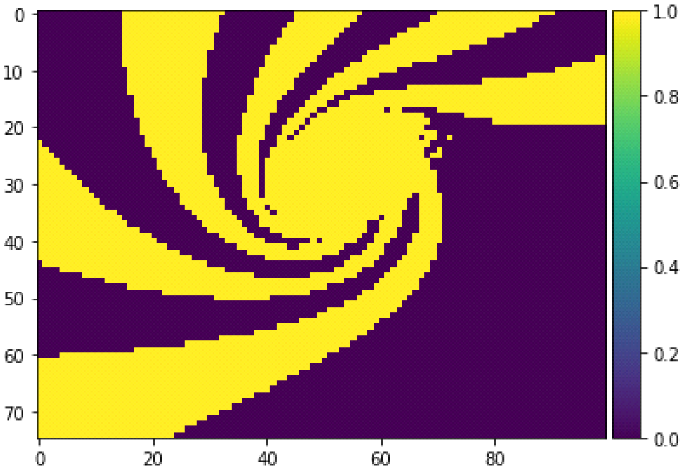
C1 → Outer ellipse, C2 - Innerellipse

Thus **distorting our cluster choice**.
So, there exists **no choice of two centres** which can form equivalent clustering as spectral clustering does.

PART-3 PARTITIONING IMAGES

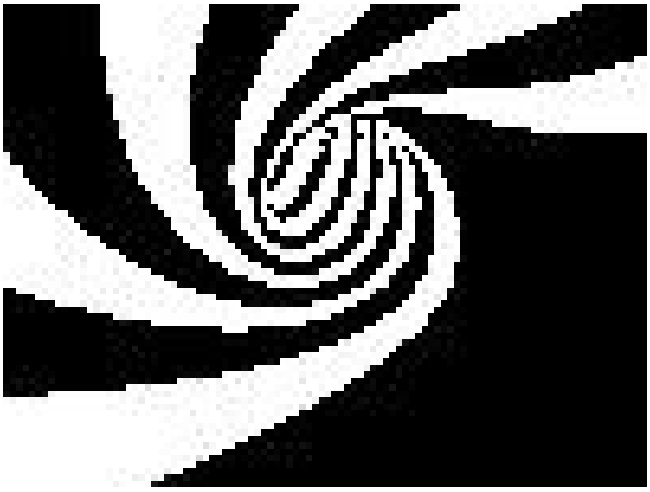
1. Performing the same comparison for spectral and k means clustering algorithms over the image partitioning example gives the following results

KMEANS CLUSTERING



Spirals at centers are not recognized properly.

SPECTRAL CLUSTERING



Edges of separation are recognised properly

OBSERVATIONS :

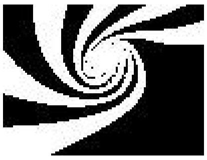
1. Performance of **KMEANS degraded** in detecting separations between white and black spirals as they approach towards the centre.
2. **SPECTRAL** clustering could still identify the separations between white and black lines as they spiral towards the centre.

As KMEANS just tries to minimize within cluster sum of square distance. SPECTRAL clustering solves convex relaxation of the normalized cuts problem on the similarity graph.

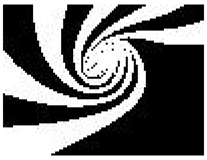
It tries to cut the similarity graph in such a way that weight of edges cut is smaller compared to weights of edges inside the cluster.

Following are the trials with different sigma values for Image Partitioning problem:

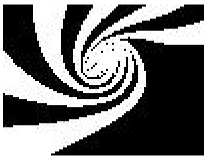
SIGMA : 0.001



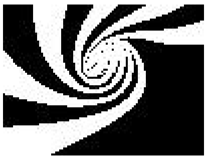
SIGMA : 0.005



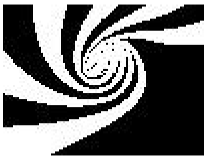
SIGMA : 0.015



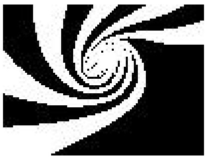
SIGMA : 0.055



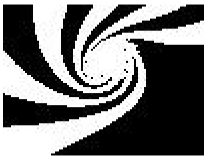
SIGMA : 0.06



SIGMA : 0.075

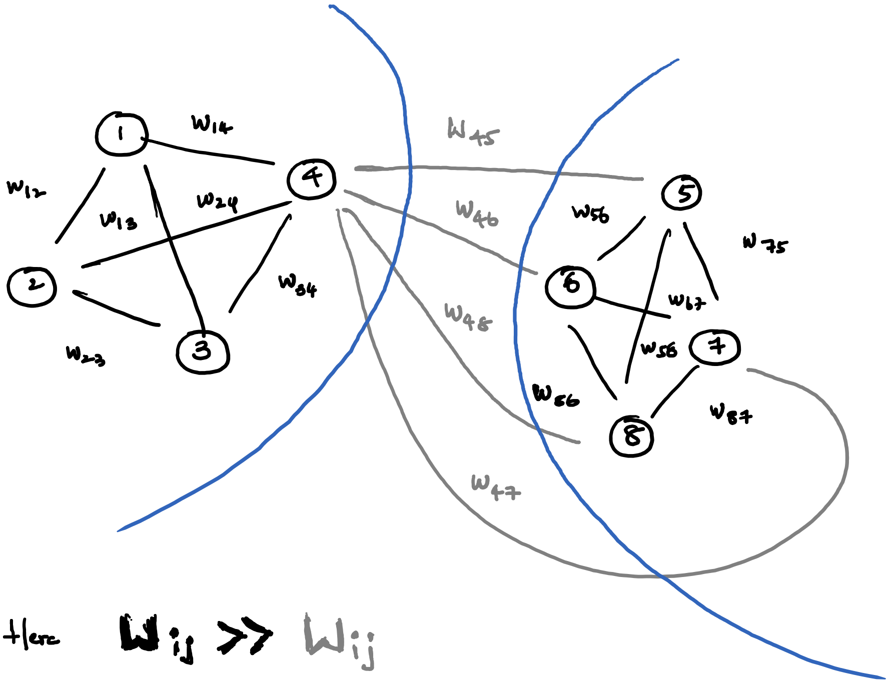


SIGMA : 0.5



SPECTRAL CUSTERING

Similarity matrix → $A_{ij} = \boxed{w_{ij}}$ weights corresponding to x_i, x_j



Here $\mathbf{W_{ij}} >> W_{ij}$

(weights described in BLACK) (Weights described in GREY)

Spectral Clustering constructs the complete graph formed by SIMILARITY MATRIX then searches for a cut

Division of vertices V_1, $S \setminus V_1$ such that sum of all GREY edges is minimized

It solves the problem by considering nearest neighbour graph [GRAPH DISTANCE] Not [DISTANCE B/W POINTS]