Homework Assignment 3
Due by end of the day Dec 7th via canvas

SDS 385 Statistical Models for Big Data

1. **Spectral Clustering**[1] In class we have seen k-means for estimating GMM's. Since your professor will not get a chance to actually cover her favorite clustering method aka spectral clustering, methinks this is a great opportunity to introduce you to it. There is a class of clustering algorithms, called spectral clustering algorithms, which has recently become quite popular. Many of these algorithms are quite easy to implement and perform well on certain clustering problems compared to more traditional methods like k-means. In this problem, we will try to develop some intuition about why these approaches make sense and implement one of these algorithms.

   Before beginning, we'll review a few basic linear algebra concepts you may find useful for some of the problems.

   - If A is a matrix, it has an v with eigenvalue λ if $Av = \lambda v$.
   - For any $m \times m$ symmetric matrix A, the *Singular Value Decomposition* of A yields a factorization of A into
     $$A = USU^T$$
     where U is an $m \times m$ orthogonal matrix (meaning that the columns are pairwise orthogonal). and $S = diag(|\lambda_1|, |\lambda_2|, \ldots, |\lambda_m|)$ where the λ_i are the eigenvalues of A.

   Given a set of m datapoints $x_1, \ldots, x_m$, the input to a spectral clustering algorithm typically consists of a matrix, A, of pairwise similarities between datapoints often called the *affinity matrix*. The choice of how to measure similarity between points is one which is often left to the practitioner. A very simple affinity matrix can be constructed as follows:

   $$A(i,j) = A(j,i) = \begin{cases} 1 & \text{if } d(x_i, x_j) < \Theta \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

   where $d(x_i, x_j)$ denotes the Euclidean distance between points x_i and x_j.

   The general idea of spectral clustering is to construct a mapping of the datapoints to an eigenspace of A with the hope that points are well separated in this eigenspace so that something simple like k-means applied to these new points will perform well.

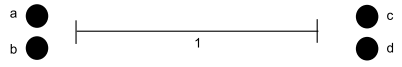

Figure 1: Simple dataset

[1] This problem was designed in collaboration with Jon Huang

As an example, consider forming the affinity matrix for the dataset in Figure 1 using Equation 1 with $\Theta = 1$. We have that

$$A = \left[\begin{array}{c|cccc} & a & b & c & d \\ \hline a & 1 & 1 & 0 & 0 \\ b & 1 & 1 & 0 & 0 \\ c & 0 & 0 & 1 & 1 \\ d & 0 & 0 & 1 & 1 \end{array}\right]$$

Now for this particular example, the clusters $\{a, b\}$ and $\{c, d\}$ show up as nonzero blocks in the affinity matrix. This is, of course, artificial, since we could have constructed the matrix A using any ordering of $\{a, b, c, d\}$. For example, another possibile affinity matrix for A could have been:

$$\tilde{A} = \left[\begin{array}{c|cccc} & a & c & b & d \\ \hline a & 1 & 0 & 1 & 0 \\ c & 0 & 1 & 0 & 1 \\ b & 1 & 0 & 1 & 0 \\ d & 0 & 1 & 0 & 1 \end{array}\right]$$

The key insight here is that the eigenvectors of matrices A and $\tilde{A}$ have the same entries (just permuted). The eigenvectors with nonzero eigenvalue of A are: $e_1 = (.7, .7, 0, 0)^T$, $e_2 = (0, 0, .7, .7)$. And the nonzero eigenvectors of $\tilde{A}$ are: $\tilde{e}_1 = (.7, 0, .7, 0)^T$, $\tilde{e}_2 = (0, .7, 0, .7)$. Spectral clustering embeds the original datapoints in a new space by using the coordinates of these eigenvectors. Specifically, it maps the point x_i to the point $(e_1(i), e_2(i), \ldots, e_k(i))$ where $e_1, \ldots, e_k$ are the top k eigenvectors of A. We refer to this mapping as the *spectral embedding*.

Algorithm description

Frequently, the affinity matrix is constructed as

$$A_{ij} = \begin{cases} 1 & \text{If } i, j \text{ are amongst } k \text{ nearest neighbors of each other} \\ 0 & \text{Otherwise} \end{cases} \tag{2}$$

The best that we can hope for in practice is a near block-diagonal affinity matrix. It can be shown in this case, that after projecting to the space spanned by the top k eigenvectors, points which belong to the same block are close to each other in a euclidean sense. We won't try to prove this, but using this intuition, you will implement one (of many) possible spectral clustering algorithms. This particular algorithm is described in

`On Spectral Clustering: Analysis and an algorithm`
`Andrew Y. Ng, Michael I. Jordan, Yair Weiss (2001)`

We won't try to justify every step, but see the paper if you are interested. The steps are as follows:

- Construct an affinity matrix A using Equation 2.
- Symmetrically 'normalize' the rows and columns of A to get a matrix N: such that $N(i, j) = \frac{A(i,j)}{\sqrt{d(i)d(j)}}$, where $d(i) = \sum_k A(i, k)$.

- Construct a matrix Y whose columns are the first k eigenvectors of N.
- Normalize each row of Y such that it is of unit length.
- Cluster the dataset by running k-means on the set of spectrally embedded points, where each row of Y is a datapoint.

(a) Generate a dataset with 10000 points, with 5000 coming from a Gaussian centered at $\mu_1(i) = 3/\sqrt{p}$ with $\Sigma_1 = I_p$ and the rest from a mean $\mu_2(i) = -3/\sqrt{p}$, for $i = 1, \ldots, p$, gaussian with $\Sigma_2 = I_p$. Create a $\tilde{k}$-nearest neighbor graph from this dataset and do Spectral clustering. For $p \in \{250, 500, 1000, 2000\}$, plot the time taken and the clustering accuracy of the following algorithms, averaged over 10 randomly generated datasets. You just have to write your own code for Spectral clustering, you can use available software for kdtree for the rest. There are three algorithms, *Exact*, *JL+Exact*, *JL+KDtrees*. Use $\tilde{k} = 5$. You can also try out different $\tilde{k}$, its interesting to try this because too small or too large $\tilde{k}$ can lead to a bad clustering accuracy. Remember that you can calculate the clustering accuracy because you generated the data and hence know the "latent" ground truth memberships.

i. (Exact) Use the brute force k-nearest neighbor algorithm. If this is taking too long, you may omit the results, but please write explicitly how long it took, e.g. "my $\tilde{k}$-nn graph took over yyy hours to build when $p = xxx$ and so I gave up."

ii. (JL+Exact, JL+KDtrees) Use the Johnson Lindenstrauss lemma to reduce the dimensionality of the data. Remember, for n points JL lets you project the datapoints into a $O(\log n/\epsilon^2)$ dimensional space with a multiplicative distortion of ϵ of the distances. Try out different ϵ values to generate the curves. For small ϵ, you would have higher accuracy and longer processing time, whereas for larger ϵ you would have lower accuracy and less computation time. Now use the KDtree algorithm and exact $\tilde{k}$-nearest neighbors to build the nearest neighbor graph. There is a builtin function in Matlab using knnsearch which has an option of using the KDtree.

2. We are going to learn and implement the power method in this problem. Let the eigenvalues of a square symmetric matrix $A \in \mathbb{R}^{n \times n}$ be given by $|\lambda_1| > |\lambda_2| > \ldots$. We will assume that the eigenvalues are all different for convenience of analysis. Start with some vector $\boldsymbol{q}^{(0)}$. Now compute the following:

$$\boldsymbol{z}^{(k)} = A\boldsymbol{q}^{(k-1)} \tag{3}$$

$$\boldsymbol{q}^{(k)} = \frac{\boldsymbol{z}^k}{\|\boldsymbol{z}^k\|} \tag{4}$$

$$\nu^{(k)} = (\boldsymbol{q}^{(k)})^T A\boldsymbol{q}^{(k)} \tag{5}$$

(a) Prove that for $k > 1$,

$$\boldsymbol{q}^{(k)} = \frac{A^k\boldsymbol{q}^{(0)}}{\|A^k\boldsymbol{q}^{(0)}\|}.$$

You can use induction to do this.

(b) Now let the eigenvectors of A be $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ and let

$$\boldsymbol{q}^{(0)} = \sum_{i=1}^{n} \alpha_i \boldsymbol{x}_i, \quad \boldsymbol{y}^{(k)} = \sum_{i=2}^{n} \frac{\alpha_i}{\alpha_1} \left(\frac{\lambda_i}{\lambda_1}\right)^k \boldsymbol{x}_i.$$

Show that

$$\|\boldsymbol{q}^{(k)} - \langle \boldsymbol{q}^{(k)}, \boldsymbol{x}_1 \rangle \boldsymbol{x}_1\| \leq C \left| \frac{\lambda_2}{\lambda_1} \right|^k$$

3. Last, but not the least, we will learn kernel PCA to deal with non-linear decision boundaries. Recall PCA? There you first centered your data to get $\tilde{X}$, computed covariance matrix, and then compute top K eigenvectors V_k of this and project a datapoint $\tilde{x}$ to get $\tilde{x}^T V_k$. Now we will directly compute these projections .

   (a) Assume that you are mapping the datapoints to a different feature space $\phi(\boldsymbol{x}) \in \mathbb{R}^N$ and $\sum_i \phi(\boldsymbol{x}_i) = 0$. While we will not do this explicitly, let us follow through the steps of PCA applied on this new feature space. Create a new matrix $\phi(X) \in \mathbb{R}^{n \times N}$. Let $\boldsymbol{v}$ be the first eigenvector of the covariance matrix in this feature space, i.e.

   $$\sum_{i=1}^{n} \phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_i)^T \boldsymbol{v} = \lambda \boldsymbol{v}. \tag{6}$$

   Show that $\boldsymbol{v}$ can be written as

   $$\boldsymbol{v} = \phi(X)^T \boldsymbol{a} \tag{7}$$

   So finding $\boldsymbol{v}$ boils down to finding $\boldsymbol{a}$.

   (b) Now show that if you could get $\boldsymbol{a}$, the projection of the data (in the new space) on this direction is given by:

   $$\Phi(X)\boldsymbol{v} = K\boldsymbol{a},$$

   where $K(i,j) = \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j)$. Now we will get $\boldsymbol{a}$.

   (c) Now plug in Eq (7) to Eq (6), and left multiply by $\phi(\boldsymbol{x}_k)^T$ to get:

   $$K\boldsymbol{a} = c_\lambda \boldsymbol{a},$$

   where c_λ is a scalar depending on λ. You can assume that K is invertible.

   (d) But typically we don't have centered features. So instead of $\phi(\boldsymbol{x}_i)$ we should work with $\psi(\boldsymbol{x}_i) = \phi(\boldsymbol{x}_i) - \dfrac{\sum_j \phi(\boldsymbol{x}_j)}{n}$ Show that the kernel matrix built from these centered feature vectors equal $\tilde{K} = K - K\mathbf{1}\mathbf{1}^T/n - \mathbf{1}\mathbf{1}^T/nK + \mathbf{1}^T K \mathbf{1}/n^2 \mathbf{1}\mathbf{1}^T$

   (e) So the final algorithm is to build $\tilde{K}$ matrix from the data using your choice of a kernel, and then compute top eigenvector of this matrix and project K on that direction. Download the two parabola dataset. Plot the first principal component using PCA. Use your power iteration algorithm to do this.

   (f) Now do Kernel PCA with the RBF kernel $K(i,j) = \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/2\sigma^2)$. Change your power method algorithm so that you do not need to compute the $\tilde{K}$ matrix, but only the K matrix. Provide the Pseudocode.

   (g) Try different values of σ, and report the one which returns a first kernel PC such that the two classes are separated along this direction.

   (h) What do you think is the relationship of Spectral Clustering with Kernel PCA? Give your answer in 4-5 lines.