EXERCISE SESSION Spectral Clustering: ADVANCED MACHINE LEARNING COURSE – EPFL – Lecturer A. Billard

A key step in spectral clustering is the assumption that the non-zero entries on the eigenvectors with small eigenvalues code for groups of points that are close to another. In this exercise, we will get an intuition of why this assumption may be correct.

Exercise 1:

Consider a (two-dimensional) dataset composed of two points.

1) Build a similarity matrix using a threshold function on Euclidean (norm-2) distance. The metric outputs 1 if the points are close enough according to a threshold and zero otherwise. Consider two cases: when the two datapoints are close or far.
2) Build the Laplacian in each case and discuss the eigenvalues and eigenvectors.

Solutions (Exercise 1):

Case 1: When the points are close, the similarity matrix is given by $S = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$.

The sum over each line is two. Hence the Laplacian is also:

$$L = D - S = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

where D is a diagonal matrix composed, in its diagonal, of the sum of the entries on the lines of S. One can compute the eigenvalues as follows:

$$\det\left[L - \lambda I\right] = \det\begin{bmatrix} 1-\lambda & -1 \\ -1 & 1-\lambda \end{bmatrix}$$

$$\Rightarrow (1-\lambda)^2 - 1 = 0$$

$$\Rightarrow (1-\lambda)^2 = 1$$

$$\Rightarrow (1-\lambda) = \pm 1$$

$$\Rightarrow \lambda_1 = 0,\ \lambda_2 = 2.$$

The eigenvalue 0 has multiplicity 1. The datapoints are hence grouped in a single cluster (the graph has connectivity 1). Indeed, since we assume that the datapoints were very close to one another, according to our metric *S*, the points were tightly grouped.

Case 2: If the points are very far apart, the similarity matrix is given by $S = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. The sum over each line is 1. Hence the Laplacian is also:

$L = D - S = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$. It has a single eigenvalue $\lambda = 0$ with multiplicity 2. Any pair of orthonormal vectors in $\mathbb{R}^2$ are eigenvectors of this matrix. Hence the associated graph has two clusters, each composed of a single datapoint.

Exercise 2:

Consider a two-dimensional dataset composed of two points (assume again two cases – points are close to one another or are far apart).

1) Build a similarity matrix using a RBF kernel. Build the Laplacian matrix, perform an eigenvalue decomposition and discuss the eigenvalues and eigenvectors, for each of the two cases above.
2) Repeat (1) using a homogeneous polynomial kernel with p=2.

Solutions (Exercise 2):

1) The similarity matrix is given by $S = \begin{bmatrix} 1 & \mathrm{d} \\ \mathrm{d} & 1 \end{bmatrix}$, where d is the distance between the two points given by the RBF kernel. In principle, the RBF kernel is never zero and we have $0 < d \leq 1$.

   The sum over each line is 1+d. Hence the Laplacian is: $L = D - S = \begin{bmatrix} 1+d & 0 \\ 0 & 1+d \end{bmatrix} - \begin{bmatrix} 1 & d \\ d & 1 \end{bmatrix} - = \begin{bmatrix} d & -d \\ -d & d \end{bmatrix}$.

   $\det\left[L - \lambda I\right] = \det\begin{bmatrix} d-\lambda & -\mathrm{d} \\ -\mathrm{d} & d-\lambda \end{bmatrix}$

   $\Rightarrow \left(d-\lambda\right)^2 - d^2 = 0$

   $\Rightarrow \left(d-\lambda\right) = \pm d$

   $\Rightarrow \lambda_1 = 0$

   $\Rightarrow \lambda_2 = 2d$

   We have the same solution as in (1) above. That means that the eigenvector with eigenvalue 0 has multiplicity 1. So, in principle, since the RBF kernel is always positive, one cannot use the decomposition to obtain separate clusters. However, in practice, when the datapoints are very far from each other, *d* becomes zero and we have again a single eigenvalue $\lambda = 0$ with multiplicity two, indicating two connected components, each of which is composed of a single of the two datapoints. The closer the points, the larger *d*.

   Hence, in the case of a matrix of similarity full, one can use the absolute value of the eigenvalues to determine whether the points are close or not. If an eigenvalue is close to zero, this indicates the existence of a cluster.

2) The similarity matrix is given by $S = \begin{bmatrix} a & d \\ d & b \end{bmatrix}$, where a and b are the length of the vector of the two points elevated to power 2p, i.e. $a = \left\| x^1 \right\|^{2p}$ and $b = \left\| x^2 \right\|^{2p}$, and $d = \left\| x^1 \right\|^p \left\| x^2 \right\|^p \cos^p \left(x^1, x^2 \right)$ is the distance between the two points given by the dot product (which may be positive or negative depending on the sign of the cos function) elevated to the power of p, where p is the degree of the polynomial kernel. The Laplacian is:

$$L = D - S = \begin{bmatrix} a+d & 0 \\ 0 & b+d \end{bmatrix} - \begin{bmatrix} a & d \\ d & b \end{bmatrix} = \begin{bmatrix} d & -d \\ -d & d \end{bmatrix}.$$

We are back to the solution found for the RBF kernel.
Note however that here one can get the clear partitioning that one had for the binary case when $\cos\left(x^1, x^2 \right) = 0$, i.e. when the two vectors x^1, x^2 are orthogonal to one another with respect to the origin. A true partitioning can hence be derived when the similarity matrix entails non-binary entries but is sparse (some of the entries are zero). With the polynomial kernel, the clustering is done by partitioning the space into groups of datapoints living in each quadrant (see solution for kernel K-means).

In general, it is correct to say that the smaller *d*, the further apart the datapoints in the polynomial kernel.

These simple 2D examples support the claim that the smaller the eigenvalue, the more split the points are.

Exercise 3:

Consider a dataset composed of four points with two pairs of points that are close to each other, one pair being far from the other.

More formally, assume that the similarity matrix looks as follows:

$$S = \begin{bmatrix} 1 & 0.8 & 0 & 0 \\ 0.8 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.5 \\ 0 & 0 & 0.5 & 1 \end{bmatrix}$$

1) What are the eigenvalues and eigenvectors of L = D – S? How many connected components do you obtain?
2) What are the eigenvalues and eigenvectors of S? What do you notice? How would infer clusters of points?
   Hint: Look at the ratio of the eigenvalues

Solutions (Exercise 3):

1) To compute the Laplacian matrix, we first compute the sum of each row (or column) and define:

$$D = \begin{bmatrix} 1.8 & 0 & 0 & 0 \\ 0 & 1.8 & 0 & 0 \\ 0 & 0 & 1.5 & 0 \\ 0 & 0 & 0 & 1.5 \end{bmatrix}$$

Hence, we obtain the Laplacian matrix by subtracting S to D, i.e.

$$L = \begin{bmatrix} 0.8 & -0.8 & 0 & 0 \\ -0.8 & 0.8 & 0 & 0 \\ 0 & 0 & 0.5 & -0.5 \\ 0 & 0 & -0.5 & 0.5 \end{bmatrix}$$

The eigenvectors and eigenvalues can be determined directly if we notice the peculiar structure of this Laplacian matrix. Indeed, it is made of two blocks of symmetric matrices with off-diagonal values being the opposite of the diagonal values. Hence, the obvious solutions are the eigenvectors:

$$e^1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad e^2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \quad e^3 = \frac{1}{\sqrt{2}}\begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \end{bmatrix} \quad e^4 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}$$

for the eigenvalues (in **increasing** order):

$$\lambda_1 = 0 \quad \lambda_2 = 0 \quad \lambda_3 = 1 \quad \lambda_4 = 1.6$$

We can also perform the eigenvalue decomposition in a more systematic way using the determinant:

$$\det\left[L-\lambda I\right]=\begin{vmatrix} 0.8-\lambda & -0.8 & 0 & 0 \\ -0.8 & 0.8-\lambda & 0 & 0 \\ 0 & 0 & 0.5-\lambda & -0.5 \\ 0 & 0 & -0.5 & 0.5-\lambda \end{vmatrix}$$

$$\Rightarrow ((0.8-\lambda)^2-0.8^2)((0.5-\lambda)^2-0.5^2)=0$$
$$\Rightarrow \lambda(\lambda-1.6)\lambda(\lambda-1)$$
$$\Rightarrow \lambda_1=0,\quad \lambda_2=0,\quad \lambda_3=1.6,\quad \lambda_4=1$$

and then compute the eigenvectors which are solutions of $L-\lambda I=0$, for all eigenvalues.

As seen in the previous exercises, the zero eigenvalue of multiplicity 2 indicates that we have 2 connected components. As indicated by the first two eigenvectors, the two components are made of the pairs (x_1,x_2) and (x_3,x_4), respectively.

2) We now consider computing the eigenvalue decomposition of the similarity matrix S directly:

$$S=\begin{bmatrix} 1 & 0.8 & 0 & 0 \\ 0.8 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.5 \\ 0 & 0 & 0.5 & 1 \end{bmatrix}$$

Once again, we can notice that S is a block diagonal matrix and deduce the rather simple solutions of the eigenvalue decomposition or use the determinant as shown above. We obtain the following eigenvectors:

$$e^1=\frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\\0\\0\end{bmatrix}\quad e^2=\frac{1}{\sqrt{2}}\begin{bmatrix}0\\0\\1\\1\end{bmatrix}\quad e^3=\frac{1}{\sqrt{2}}\begin{bmatrix}0\\0\\1\\-1\end{bmatrix}\quad e^4=\frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\\0\\0\end{bmatrix}$$

for the eigenvalues (in **decreasing** order):

$$\lambda_1=1.8\quad \lambda_2=1.5\quad \lambda_3=0.5\quad \lambda_4=0.2$$

Interestingly enough, we obtain the same eigenvectors as in the eigenvalue decomposition of the corresponding Laplacian matrix L, with (different) decreasing values for the eigenvalues. However, we no longer observe the convenient eigenvalues (close or equal to zero) that we had in the former case and cannot directly infer clusters of points from the eigenvalues…Or can we?

In fact, we may be able to get a sense of the distribution of points (i.e. clusters) by looking at the ratio between the eigenvalues. Indeed, the first 2 eigenvalues are very comparable and relatively high when compared to the others. This would suggest the two connected components that we would expect. Contrary to the eigenvalue decomposition of the Laplacian matrix, we could consider that the largest eigenvalues provide information on the partitioning of the graph.

The specific ordering of the eigenvalues (decreasing order) and the discarding of the smallest eigenvalues may already ring a bell. Indeed, this whole scheme amounts to the application of PCA on the Similarity matrix S, which in turn may be seen as a Gram matrix. Using this analogy with Kernel PCA, we can deduce that the eigenvalues computed by Kernel PCA can be used as an indication of the number of connected components (or clusters in the clustering/classification context) for a given set of hyperparameters. Typically, one may compute the eigenvalues obtained through Kernel PCA, sort them in decreasing order and compare their ratio in order to infer the number of components. Therefore, Kernel PCA can be used in the initial choice of the hyperparameters of Kernel K-Means (both the number of cluster K and the kernel hyperparameters).

NB: There is actually a strong relation between Kernel K-Means, Kernel PCA and Spectral Clustering, and the interested reader may refer to the brief explanation given in the description of the second Practical session (TP2 on Manifold and Clustering) and to the related documentation available on the LASA website and Moodle (see Kernel K-Means section).