1 Spectral clustering [50 points]

1. (20 points) Consider an undirected graph with non-negative edge weights w_{ij} and graph Laplacian L. Suppose there are m connected components $A_1, A_2, \ldots, A_m$ in the graph. Show that there are m eigenvectors of L corresponding to eigenvalue zero, and the indicator vectors of these components $I_{A_1}, \ldots, I_{A_m}$ span the zero eigenspace.

Answer:

Step 1. Proof of graph Laplacian property:

For every vector $f \in R^n$:

$$f'Lf = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2$$

By the definition of D, W, we know that:

$$d_i = \sum_{j=1}^{n} w_{ij}$$

Here is the proof:

$$f'Lf = f'Df - f'Wf = \sum_{i=1}^{n} d_i {f_i}^2 - \sum_{i,j=1}^{n} w_{ij} f_i f_j$$

$$= \frac{1}{2}\left(\sum_{i=1}^{n} d_i {f_i}^2 - 2\sum_{i,j=1}^{n} w_{ij} f_i f_j + \sum_{j=1}^{n} d_i {f_j}^2\right) = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2$$

Step 2. Proof of m=1 case.

In this situation, the graph is connected as one component. Assume f is the eigen vector:

$$0 = f'Lf = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2$$

Because w_{ij} is bigger than zero (non-negative weights). Then for any connected vertices v_i and v_j, $f_i = f_j$.

Because all vertices are connected, then $f_1 = f_2 = \ldots = f_n$

According to the class slide, when m = 1, constant one vector is the eigenvector with eigen value as zero.

Then $f_1 = f_2 = \ldots = f_n = 1$.

So there are only one eigenvector of eigenvalue as zero. And the indicator vector tells the connected components.

Step 3. Proof of m components case. We assume that vertices are ordered by the connected components. So adjacency matrix W and matrix L have a block diagonal form.

$$L = \begin{pmatrix} L_1 & & \\ & \ddots & \\ & & L_k \end{pmatrix}$$

Because each block L_i is a graph Laplacian on its own, corresponding to the i-th connected components. L is formed by the union of all L_i, and L have eigenvectors corresponding to L_i eigenvectors, filled with 0 at other blocks positions. From step 2 proof, we know that each L_i has constant one eigenvector on the i-th connected components, with eigenvalue 0 and multiplicity 1. So we have proved that:

There are m eigenvectors of L corresponding to eigenvalue zero, and the i-th indicator vectors are constant one eigenvector on the i-th connected components, filled with 0 at other blocks positions.

Ref:

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, *17*(4), 395-416.

2. (30 points) Real data: political blogs dataset. We will study a political blogs dataset first compiled for the paper Lada A. Adamic and Natalie Glance, "The political blogosphere and the 2004 US Election", in Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem (2005). The dataset nodes.txt contains a graph with $n = 1490$ vertices ("nodes") corresponding to political blogs. Each vertex has a 0-1 label (in the 3rd column) corresponding to the political orientation of that blog. We will consider this as the true label and try to reconstruct the true label from the graph using the spectral clustering on the graph. The dataset edges.txt contains edges between the vertices.

   Here we assume the number of clusters to be estimated is $k = 2$. Using spectral clustering to find the 2 clusters. Compare the clustering results with the true labels. What is the false classification rate (the percentage of nodes that are classified incorrectly).

Answer:

Step 1: Remove isolated nodes and connected tiny components

Isolated nodes have no edges with any other nodes and are considered as their own cluster. 266 nodes are removed and details are in the Jupiter notebook.

One connected tiny component, consisted by node 181 and 665, is isolated from other big connected components. The two nodes are considered as one cluster and are removed.

Step 2: Relabeling names, labels and edges.

Because we are using sparse matrix and we need to remove 268 nodes found in step 1 and relabel names, labels and edges.

The lists of names and labels are easy to update by removing elements using 268 indexes found in step 1.

The edges are harder to relabel. Here is my method to relabel it.

Step 2a: Build the list of new index to old index

Step 2b: Build the dictionary of old index to new index

Step 2c: Update all nodes/indexes in the edge using step 2b dictionary

Step 3: Build adjacency matrix A (symmetric), degree matrix D and symmetric normalized Laplician L.

$$A_{sym} = \frac{1}{2}(A + A^T)$$

Though Wikipedia define (symmetric) normalized Laplacian as follows. The latter one is the right L that produce positive eigenvalues.

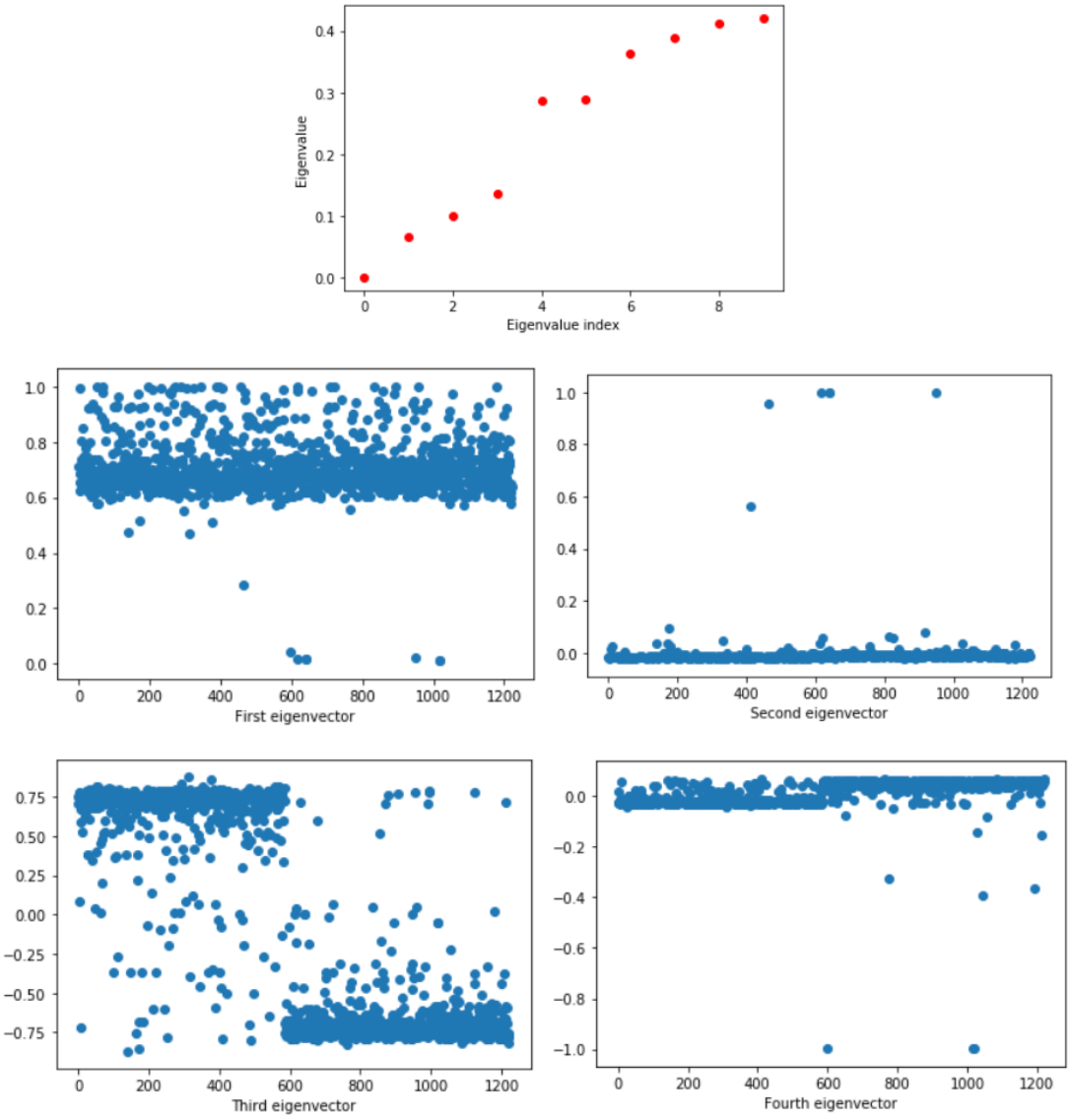
$$L^{\text{sym}} := D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$$

Step 4: Use np.linalg.eigh(L) function to get sorted eigenvalues (ascending) and corresponding eigenvectors.

According to eigenvalues, index 4 eigenvalue has a gap from the beginning ones. So beginning four eigenvectors are chosen, normalized as the new coordinates for the 1222 data points. (The gap also indicates that there are likely four clusters)

According to true labels, we know the first half is one group and the second half is the other group. The third eigenvector showed the best label assignment and must be included. Other eigenvectors wrongly assign 1222 data points to one major group.

The third eigenvector is called Fiedler vector, corresponding to the second nonzero eigenvalue (Fiedler value). After one graph cut to separate the graph into half, Fiedler vector gives the information about which side of the cut that node belongs to.



Step 5: Summary and conclusion.

Finally, beginning 3 eigenvectors are used to give 1162 correct label assignment out from 1222 data points, with a false classification rate 4.9%. Other number of eigenvectors (2 and 4) are also tested with results as follows.

Beginning n eigenvectors used	Correct assignment count	False classification rate
2	634	48.1%
3	1162	4.9%
4	1160	5.1%

2 PCA: Food consumption in European area [50 points]

The data food-consumption.csv contains 16 countries in the European area and their consumption for 20 food items, such as tea, jam, coffee, yoghurt, and others. There are some missing data entries: you may remove the rows "Sweden", "Finland", and "Spain". The goal is to perform PCA analysis on the data, i.e., find a way to perform linear combinations of features across all 20 food-item consumptions, for each country. If we extract two principal components, that means we use two singular vectors that correspond to the largest singular values of the data matrix, in combining features. You will need to implement PCA by writing your own code.

1. (15 points) Write down the set-up of PCA for this setting. Explain how the data matrix is set-up in this case (e.g., each dimension of the matrix correspond to what.) Explain in words how PCA is performed in this setting.

Answer:

Step 1: Data cleaning.

Remove 3 rows containing NaN value. After cleaning, raw data has 13 data points (countries) with 20 variables (food items).

Step 2: Normalize data

Ensure each variable/food item is compared equally, raw data is normalized.

Step 3: Compute mean and covariance matrix

Raw data is transposed to have a shape of (20, 13) as matrix X. After obtaining mean matrix, matrix X is centered and used to obtain covariance matrix C that has the shape of (20, 20).

Step 4: Perform PCA

From covariance matrix C, obtain eigenvectors w^1, w^2 . . . with largest eigenvalues λ_1, λ_2 ... These eigenvectors w^1, w^2 are principle directions that contain most variance and are orthogonal to each other.

2. (15 points) Suppose we aim to find top k principal components. Write down the mathematical optimization problem involved for solving this problem. Explain the procedure to find the top k principal components in performing PCA.

Step 1. Mathematical optimization problem.

① Given m data points, $\{x^1, x^2, \dots x^m\} \in R^n$, with mean $\mu = \frac{1}{m}\sum_{i=1}^{m} x^i$

② Find a direction $w \in R^n$ when $\|w\| \le 1$

③ The variance of data along direction w is maximized.

$$\max_{w:\|w\|\le 1} \frac{1}{m}\sum_{i=1}^{m} (w^T x^i - w^T \mu)^2$$

Step 2. Manipulate the objective.

$$\frac{1}{m}\sum_{i=1}^{m} (w^T x^i - w^T \mu)^2$$
$$= \frac{1}{m}\sum_{i=1}^{m} (w^T (x^i - \mu))^2$$
$$= \frac{1}{m}\sum_{i=1}^{m} w^T (x^i - \mu)(x^i - \mu)^T w$$
$$= w^T \underbrace{\left(\frac{1}{m}\sum_{i=1}^{m} (x^i - \mu)(x^i - \mu)^T\right)}_{\text{covariance matrix } C} w$$

Step 3. Then it becomes a eigen-value problem.

① Given a symmetric matrix $C \in R^{n \times n}$

② Find a vector $w \in R^n$ and $\|w\| = 1$

③ Such that $Cw = \lambda w$

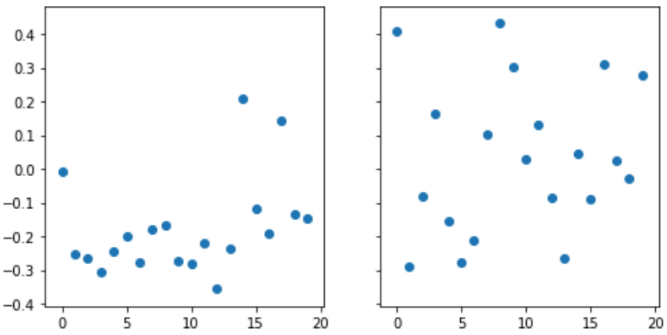
④ Multiple solution of $w^1, w^2 \dots$ with different $\lambda_1, \lambda_2 \dots$

⑤ Variance in principle direction is $w^T C w = w^T \lambda w = \lambda$,
If we find biggest eigen value $\lambda_1, \lambda_2 \dots$, we will find top principle components $w^1, w^2 \dots$

3. (10 points) Find the top two principal component vectors for the dataset and plot them (plot a value of the vector as a one-dimensional function). Describe do you see any pattern.

Answer: The top two principal component vectors are plotted as follows. In this picture, we can find the biggest absolute value and then find most important food items.

In the first vector, real coffee (index 0) is the most insignificant food and tinned fruit (index 12) is the most significant food. In the second vector, real coffee (index 0) and frozen fish (index 8) are most significant food items.



4. (10 points) Now project each data point using the top two principal component vectors (thus now each data point will be represented using a two-dimensional vector). Draw a scatter plot of two-dimensional reduced representation for each country. What pattern can you observe?

Answer: According to the scatter plot, food consumption in Ireland is quite different from other countries. Then food consumption in England, Denmark and Norway is slightly different from rest countries. Rest countries have similar food consumption.