CS 598: Theoretical Machine Learning

Lecturer: Pranjal Awasthi
Scribe: Vladimir Ivanov

Lecture #
Oct 27th, 2017

1 Graph Clustering

The problem of graph clustering attempts to identify and split a given graph into mutually exclusive graphs, where the subgraphs or 'clusters' are based on the overall connection densities found with in each cluster and between clusters. Specifically, if an arbitrary graph(G) is defined as a set of vertices(V) and a set of edges(E):

$$G = (V, E)$$

A graph clustering algorithm will split G into a set of n subsets, S_i, of vertices of G such that the following holds:

- $V = \bigcup_{i=1}^{n} S_i$
- $\emptyset = \bigcap_{i=1}^{n} S_i$

Specifically, a restricted case for splitting a graph into 2 such clusters can be approached using a stochastic block model.

2 The Stochastic Block Model

The stochastic block model begins with an assumption that there exists a perfect graph $G = (V, E)$ that consists of two disjoint subgraphs. Namely,

$$V = S_1 \bigcup S_2$$

$$\emptyset = S_1 \bigcap S_2$$

If this is was what we had from the start then the task of clustering would be trivial. However, usually, the graphs that must be clustered are not this perfect and contain edges between S_1 and S_2. These edges can be considered noise in the representation of G. Therefore, a clustering algorithm would be attempting to cluster a noisy representation, $G^{'}$, of the perfect graph G. More explicitly, G is the ground truth that a clustering algorithm is attempting to obtain from $G^{'}$.

To better formalize this situation, it is assumed that $G^{'}$ is generated by a random process according to the Erdos-Renyi Model.

2.1 Erdos-Renyi Model

The Erdos-Renyi Model assigns a probability of existence to each edge of a graph G, where the probability of each edge existing is considered independent from all the other edges in the graph. In the case of graph clustering, where $G^{'}$ is generated according to a random process from G, the probability of an edge existing within cluster S_1 or S_2 is assigned a value $p \in [0, 1]$. And, the probability of an edge existing between cluster S_1 and S_2 is $q \in [0, 1]$.

2.2 Bounds on Relationship between p and q

Before proceeding, it is important to establish bounds on probabilities p and q with respect to the theoretical feasibility of extracting G from $G^{'}$. More specifically, p and q determine the noise present in representation of G through $G^{'}$. This means that at some point there can be too much noise present for any meaningful data left from G for the algorithm to extract. It is precisely this threshold that we want to bound with respect to probabilities p,q and $|G| = n$.

To start, we look at an arbitrary vertex v_i and define its expected and variance values for the two different types of edges it may have:

- An edge $e_{i,j}$ connecting v_i to the same cluster. This means that both $v_i, v_j \in S_k$ are in the same cluster k.
  - $E[v_i, v_j \in S_k] = \frac{n}{2}p$
  - $var(v_i, v_j \in S_k) = \pm\sqrt[2]{\frac{n}{2}p(1-p)}$
- An edge $e_{i,j}$ connecting v_i to a different cluster. This means that $v_i \in S_k$ and $v_j \in S_l$ where $k \neq l$ are in different clusters.
  - $E[v_i \in S_k \; and \; v_j \in S_l | k \neq l] = \frac{n}{2}q$
  - $var(v_i \in S_k \; and \; v_j \in S_l | k \neq l) = \pm\sqrt[2]{\frac{n}{2}q(1-q)}$

Now using the above defined statistics about each edge type, we can establish the following inequality:

$$\frac{n}{2}p - \sqrt[2]{\frac{n}{2}p(1-p)} > \frac{n}{2}q + \sqrt[2]{\frac{n}{2}q(1-q)}$$

$$\frac{n}{2}p - \frac{n}{2}q > \sqrt[2]{\frac{n}{2}q(1-q)} + \sqrt[2]{\frac{n}{2}p(1-p)}$$

$$\frac{n}{2}(p-q) > \sqrt[2]{\frac{n}{2}q(1-q)} + \sqrt[2]{\frac{n}{2}p(1-p)}$$

$$p - q > C_1\sqrt[2]{\frac{p\log n}{n}}$$

where C_1 is some constant.

With the bounds established on the relationship between p and q, now we know whether it is theoretically possible for any algorithm to extract G from $G^{'}$. This bound is known as the fundamental limit.

3 An Algorithm for Graph Clustering

In the previous section, theoretical bounds were established telling use whether it is at all feasible for an algorithm to extract the ground truth graph G from noisy graph $G^{'}$ generated using a stochastic block model. There are many different algorithms for approaching this problem. One way is through semi-definite programming which can provide solid results. However, a more general and practical approach is to use the currently very popular and

widely used spectral clustering algorithms. In addition to satisfying the fundamental limit, spectral clustering algorithms require one other constraint on p in order to guarantee results:

$$p > \frac{C_2 log^4(n)}{n}$$

where C_2 is some constant. Thus, assuming both the constrain imposed on p above and the fundamental limit derived in the section above, the spectral clustering algorithm takes the following form:

Algorithm Spectral Clustering

1: Obtain $G^{'}$ with the assumption that it was generated by a stochastic block model.
2: Obtain the singular value decomposition of the adjacency matrix of $G^{'}$.

$$M^{'} = \sum_{i=1}^{n} \sigma_i v_i v_i^T$$

3: Project rows of $M^{'}$ into Euclidean space
4: Run an arbitrary clustering algorithm, such as k-means, on $\hat{M}$.

3.1 An Example of Spectral Clustering

To illustrate how this algorithm works, in the simplest case one can apply it to $G = E[G^{'}]$. Since, G is the ground truth graph, its adjacency matrix has the following form:

$$M = E[M^{'}] = \begin{bmatrix} p_{1,1} & \cdots & p_{1,i} & q_{1,i+1} & \cdots & q_{1,n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ p_{j,1} & \cdots & p_{j,i} & q_{j,i+1} & \cdots & q_{j,n} \\ p_{j+1,1} & \cdots & p_{j+1,i} & q_{j+1,i+1} & \cdots & q_{j+1,n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ p_{n,1} & \cdots & p_{n,i} & q_{n,i+1} & \cdots & q_{n,n} \end{bmatrix}$$

From the above it is clear that M is of rank 2. The singular value decomposition can be expressed as:

$$M = \sum_{i=1}^{n} \sigma_i u_i v_i^T = \frac{p+q}{2} v_1 v_1^T + \frac{p-q}{2} v_2 v_2^T$$

where v_1 and v_2 are n by 1 orthonormal vectors:

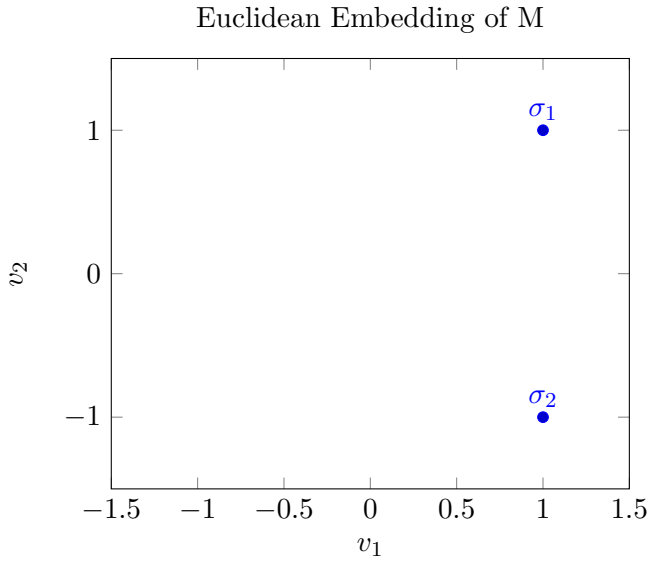
$$v_1 = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt[2]{n}} \\ \vdots \\ \frac{1}{\sqrt[2]{n}} \end{bmatrix}$$

$$v_2 = \begin{bmatrix} b_1 \\ \vdots \\ b_j \\ -b_{j+1} \\ \vdots \\ -b_n \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt[2]{n}} \\ \vdots \\ \frac{-1}{\sqrt[2]{n}} \end{bmatrix}$$

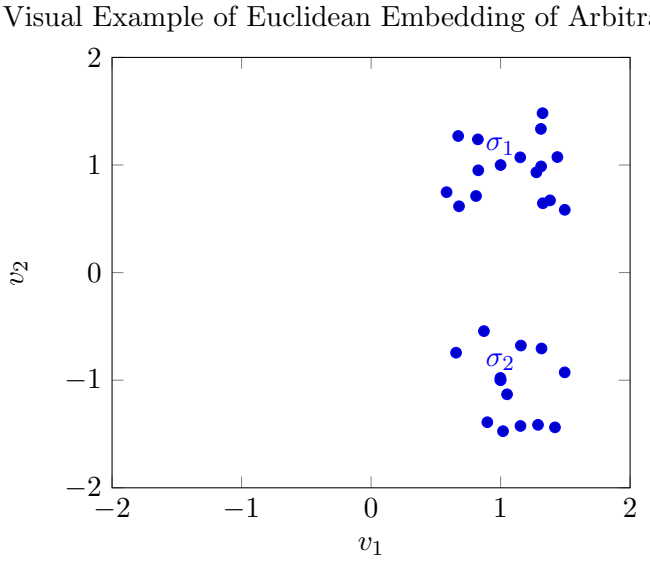
Then, if we substitute in $\sigma_1 = \frac{n}{2}(p+q)$ and $\sigma_2 = \frac{n}{2}(p-q)$, we get the SVD of M:

$$G = \sigma_1 v_1 v_1^T + \sigma_2 v_2 v_2^T$$

Next, each row of M is projected into a Euclidean space:



As a last step, we can run any simple clustering algorithm over this Euclidean embedding to obtain S_1 and S_2 clusters. In a real world case where this spectral clustering will be applied to a noisy version of G, $G^{'}$, the projections will not fall neatly onto points σ_1 and σ_2, but instead will cluster around these central points with some variance looking something like this:



4 Bounding the Spectral Clustering Error

The next logical question to ask is how many of the projections from $G^{'}$ into this Euclidean embedding can be expected to be misclassified. It turns out that if the above bounds on p

and q are satisfied, a further bound can be established on the ultimate number of projections that will be misclassified by spectral clustering. This is shown in the theorem below.

Theorem 1. *If* $p - q > C_1 \sqrt[2]{\frac{p \log n}{n}}$ *is satisfied, then with high probability spectral clustering will output* $S_1^{'}$ *and* $S_2^{'}$ *such that:*

$$|S_1 \Delta S_1^{'}| + |S_2 \Delta S_2^{'}| \leq \frac{n}{\sqrt[2]{log(n)}}$$

Proof. To begin the proof, lets first define all of our objects:

- The perfect graph G is represented by connectivity matrix M of size n×n:

$$M = \sigma_1 v_1 v_1^T + \sigma_2 v_2 v_2^T$$

where:

$$\sigma_1 = \frac{n}{2}(p+q)$$

$$\sigma_2 = \frac{n}{2}(p-q)$$

$$v_1 = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt[2]{n}} \\ \vdots \\ \frac{1}{\sqrt[2]{n}} \end{bmatrix}$$

$$v_2 = \begin{bmatrix} b_1 \\ \vdots \\ b_j \\ -b_{j+1} \\ \vdots \\ -b_n \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt[2]{n}} \\ \vdots \\ \frac{-1}{\sqrt[2]{n}} \end{bmatrix}$$

- The noisy representation of M is represented by matrix $M^{'}$ also of size n×n:

$$M^{'} = \sum_{i=1}^{n} u_i w_i w_i^T = u_1 w_1 w_1^T + u_2 w_2 w_2^T$$

Now, using perturbation theory of matrices, we need to show the following:

1. That $||\sigma_i - u_i||$ is small for all singular values of M,$M^{'}$.
2. That $||v_2 - w_2||$ is small for all singular vectors of M,$M^{'}$.

First we rewrite $M^{'}$ as the combination of the perfect matrix M and a noise matrix R:

$$M^{'} = M + R$$

Now we want to bound $||R||_2$ since inherently this is where the variation in projection is coming from. Therefore, we want to show that $||R||_2 \leq \epsilon$ for some $\epsilon > 0$, which would imply that singular values also exhibit little variation. Starting with what we know:

$$\sigma_i(M) = \max_{||v||=1} ||Mv||$$

$$= \max_{||v||=1} ||M^{'}v - Rv||$$

$$= \max_{||v||=1} ||M^{'}v|| - \max_{||v||=1} ||Rv||$$

$$\leq \max_{||v||=1} ||M^{'}v|| + \max_{||v||=1} ||Rv||$$

Now, we observe the structure of R. For an edge (i, j) where $i, j \in S_1$ or $i, j \in S_2$, $R_{i,j} = 1-p$ or $R_{i,j} = p$. And,

$$E[R_{i,j}] = p(1-p) - p(1-p) = 0$$

$$var(R_{i,j}) = p(1-p)^2 + (1-p)^2 = p(1-p)$$

For an edge (i, j) where $i \in S_1$ and $j \in S_2$ or $j \in S_1$ and $i \in S_2$, $R_{i,j} = 1 - q$ or $R_{i,j} = q$. And,

$$E[R_{i,j}] = q(1-q) - q(1-q) = 0$$

$$var(R_{i,j}) = q(1-q)^2 + (1-q)^2 = q(1-q)$$

Now, we have shown everything to make use of the following theorem:

Theorem 2. *Let $R_{n\times n}$ be a random matrix such that*

1. *E[$R_{i,j}$]=0*
2. *$|R_{i,j}| \leq 1$*
3. $$\max_{i,j} var(R_{i,j}) = \sigma^2$$

If $\sigma^2 \geq \frac{clog^4 n}{n}$, then with high probability $||R||_2 \leq 3\sigma\sqrt[2]{n}$.

Since $\sigma^2 \leq p(1-p)$, then by theorem 2:

$$||R_2|| \leq 3\sqrt[2]{np(1-p)}$$

Thus, we have shown part 1 of the proof. Now we want to bound the deviations between the singular vectors of M and $M^{'}$. For this we can use the following theorem:

Theorem 3 (Davis-Kahar Theorem)**.** *Let $M_{n\times n}$, $M^{'}_{n\times n}$ be two matrices such that $||M - M^{'}||_2 \leq \delta$. Let $v_1, ..., v_n$ be singular vectors of M. Let $w_1, ..., w_n$ be singular vectors of $M^{'}$. Then, the following holds $\forall i \in [n]$:*

$$||v_i - (\pm w_i)||_2 \leq \frac{2\delta}{\min_{j=i} |\sigma_j - \sigma_i|}$$

Here we notice that δ equals the bound we just derived for noise:

$$||M - M^{'}||_2 \leq \delta = 3\sqrt[2]{np(1-p)}$$

Thus, by the theorem above we obtain that the angular deviation between singular vectors of M and $M^{'}$ is:

$$||v_i - (\pm w_i)||_2 \leq \frac{6\sqrt[2]{np(1-p)}}{\min(\frac{n}{2}(p-q), nq)}$$

$$\leq \frac{6\sqrt[2]{np(1-p)}}{\frac{n}{2}(C_1\sqrt[2]{\frac{p\log n}{n}})}$$

where C_1 is some constant.

$$\leq O(\frac{1}{\sqrt[2]{log(n)}})$$

Now, suppose w_i makes k misclassified projections. Then,

$$||v_i - (\pm w_i)||_2 \geq \sqrt[2]{\frac{k}{n}}$$

This implies that the bound on number of misclassified projections from the noisy matrix $M^{'}$,and by extension graph $G^{'}$, is:

$$k \leq \frac{n}{log(n)} = O(1)$$

□

5 Reference Material

- Analysis of spectral clustering on dense graphs: http://www.cs.yale.edu/homes/spielman/561/lect21-15.pdf
- General k-partitioning: https://www.cc.gatech.edu/~mihail/D.8802readings/mcsherrystoc01.pdf
- General survey of spectral clustering methods: https://www.cs.cmu.edu/~aarti/Class/10701/readings/Luxburg06_TR.pdf