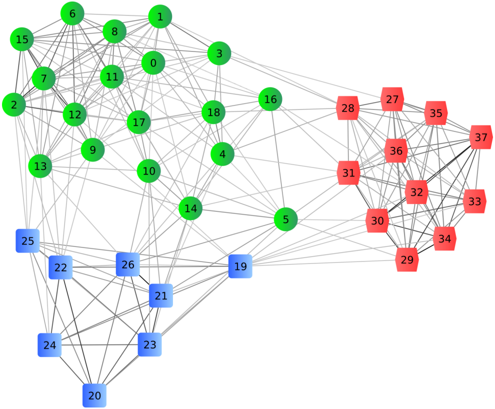
Modern Discrete Probability: A Toolkit

Stochastic blockmodel: Community detection

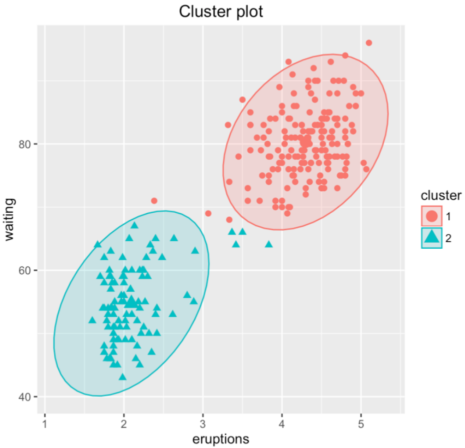
Sébastien Roch
UW–Madison
Mathematics

November 25, 2020

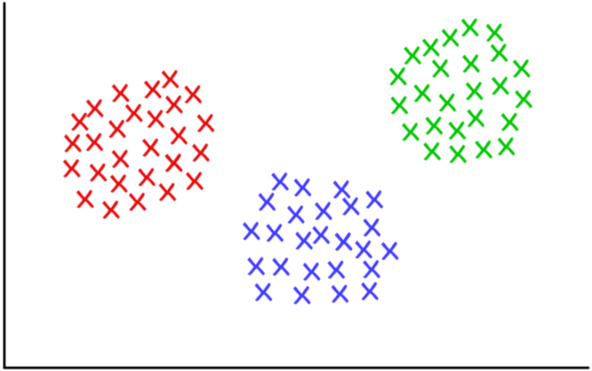
Community detection



Clustering in Euclidean space



Reducing the graph problem to clustering



Recall: Laplacian

Definition (Laplacian Matrix)

Let $G = (V, E)$ be a graph with vertices $V = \{1, \ldots, n\}$ and adjacency matrix $A \in \mathbb{R}^{n \times n}$. Let $D = \mathrm{diag}(\delta(1), \ldots, \delta(n))$ be the degree matrix. The Laplacian matrix associated to G is defined as $L = D - A$. Its entries are

$$l_{ij} = \begin{cases} \delta(i) & \text{if } i = j \\ -1 & \text{if } \{i, j\} \in E \\ 0 & \text{o.w.} \end{cases}$$

Recall: Variational characterization

Corollary (Extremal Characterization of μ_2)

Let $G = (V, E)$ be a graph with $n = |V|$ vertices. Assume the Laplacian L of G has spectral decomposition $L = \sum_{i=1}^n \mu_i \mathbf{y}_i \mathbf{y}_i^T$ with $0 = \mu_1 \leq \mu_2 \leq \cdots \leq \mu_n$ and $\mathbf{y}_1 = \frac{1}{\sqrt{n}}(1, \ldots, 1)^T$. Then

$$\mu_2 = \min \left\{ \frac{\sum_{\{u,v\}\in E} (x_u - x_v)^2}{\sum_{u=1}^n x_u^2} \,:\, \mathbf{x} \neq \mathbf{0}, \sum_{u=1}^n x_u = 0 \right\}.$$

Can think of it as a relaxation of the problem of minimizing the size of the cut between two balanced clusters

$$\min \left\{ \sum_{\{u,v\}\in E} (x_u - x_v)^2 \,:\, \mathbf{x} \in \{-1, +1\}^n, \sum_{u=1}^n x_u = 0 \right\}.$$

Stochastic blockmodel with two balanced blocks

Definition

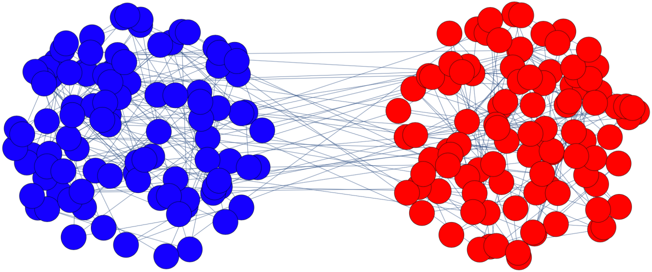
Let $V = [n]$ with n even, let $V_1 = \{1, \ldots, n/2\}$ and $V_2 = \{n/2 + 1, \ldots, n\}$, and let $0 < q < p < 1$. We draw a graph $G = (V, E)$ at random as follows. For each pair $x \neq y$ in V, the edge $\{x, y\}$ is in E with probability:

- p if $x, y \in V_1$, or $x, y \in V_2$;
- q if $x \in V_1$ and $y \in V_2$, or $x \in V_2$ and $y \in V_1$;

independently of all other edges. We write $G \sim \mathrm{SBM}_{n,p,q}$ and we denote the corresponding measure by $\mathbb{P}_{n,p,q}$.

Community detection problem: Given G (without the node labels), output V_1, V_2 (possibly approximately).

Stochastic blockmodel by picture



Expected adjacency matrix

Let $G \sim \mathrm{SBM}_{n,p,q}$ and let A be the adjacency matrix of G.

Theorem

Let $D = \mathbb{E}_{n,p,q}[A]$. Then

$$D = n\,\frac{p+q}{2}\,\mathbf{u}_1\mathbf{u}_1^T + n\,\frac{p-q}{2}\,\mathbf{u}_2\mathbf{u}_2^T - p\,I,$$

where $\mathbf{u}_1 = \frac{1}{\sqrt{n}}(1,\ldots,1)^T$ *and* $\mathbf{u}_2 = \frac{1}{\sqrt{n}}(1,\ldots,1,-1,\ldots,-1)^T$.

Proof: Note that D is a block matrix with diagonal blocks all-p and off-diagonal blocks all-q, all of size $n/2 \times n/2$, with the exception of the diagonal which is all-0. ∎

Idea: Compute the second eigenvector of A and cluster by sign.

Spectral clustering: a positive result

Theorem

Let $G \sim \mathrm{SBM}_{n,p,q}$ and let A be the adjacency matrix of G. Let $\mu = \min\left\{q, \frac{p-q}{2}\right\} > 0$. Clustering according to the sign of the second eigenvector of A identifies the two communities of G with probability at least $1 - e^{-n}$, except for C/μ^2 misclassified nodes for some constant $C > 0$.

Matrix perturbation

Theorem (A version of Davis-Kahan)

Let S and T be symmetric $n \times n$ matrices. Let $\lambda_i(S)$ be the i-th largest eigenvalue of S with corresponding unit eigenvector $\mathbf{v}_i(S)$ (and similarly for T). If

$$\delta := \min_{j \neq i} |\lambda_i(S) - \lambda_j(S)| > 0,$$

then there is $\theta \in \{+1, -1\}$ such that

$$\|\mathbf{v}_i(S) - \theta\, \mathbf{v}_i(T)\|_2 \leq \frac{4\|S - T\|}{\delta}.$$

Bounding the spectral norm

The following lemma is proved in the next section.

Lemma

Let $G \sim \mathrm{SBM}_{n,p,q}$, let A be the adjacency matrix of G and let $D = \mathbb{E}_{n,p,q}[A]$. Then, there is a constant $C > 0$ such that

$$\|A - D\| \leq C\sqrt{n},$$

with probability at least $1 - e^{-n}$.

Spectral clustering: proof I

Proof of spectral clustering theorem: The eigenvalues of D are

$$n\,\frac{p+q}{2} - p, \qquad n\,\frac{p-q}{2} - p, \qquad -p,$$

so $\lambda_2(D) = n\,\frac{p-q}{2} - p$ and

$$\delta = \min_{j \neq 2} |\lambda_2(D) - \lambda_j(D)| = \min\left\{n\,\frac{p-q}{2}, n\,q\right\} =: n\,\mu > 0.$$

By Davis-Kahan and the previous lemma, with probability at least $1 - e^{-n}$, there is $\theta \in \{+1, -1\}$ such that

$$\|\mathbf{v}_2(D) - \theta\,\mathbf{v}_2(A)\|_2 \leq \frac{4C\sqrt{n}}{n\,\mu} \leq \frac{C'}{\sqrt{n}\,\mu}.$$

Spectral clustering: proof II

Proof of spectral clustering theorem (continued): Put differently,

$$\sum_i \left|\sqrt{n}\,(\mathbf{v}_2(D))_i - \sqrt{n}\,\theta\,(\mathbf{v}_2(A))_i\right|^2 \leq \frac{(C')^2}{\mu^2}.$$

If the signs of $(\mathbf{v}_2(D))_i$ and $\theta\,(\mathbf{v}_2(A))_i$ disagree, then the i-th term in the sum above is ≥ 1. So there can be at most $(C')^2/\mu^2$ of those. That establishes the desired bound on the number of misclassified nodes. ■

1. Data science application: Community detection

2. Bounding the spectral norm

Recall: Sub-Gaussian variables

We say that a centered random variable X is *sub-Gaussian with variance factor* $\nu > 0$ if for all $s \in \mathbb{R}$

$$\Psi_X(s) \leq \frac{s^2 \nu}{2},$$

which is denoted by $X \in \mathcal{G}(\nu)$. By the Chernoff-Cramér bound

$$\mathbb{P}[X \leq -\beta] \vee \mathbb{P}[X \geq \beta] \leq \exp\left(-\frac{\beta^2}{2\nu}\right),$$

where we used that $X \in \mathcal{G}(\nu)$ implies $-X \in \mathcal{G}(\nu)$.

Recall: Hoeffding's inequality

Theorem (General Hoeffding inequality)

Let $X_1, \ldots, X_n$ *be independent centered random variables with* $X_i \in \mathcal{G}(\nu_i)$ *for* $0 < \nu_i < +\infty$ *and let* $(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$*. Let* $S_n = \sum_{i \leq n} \alpha_i X_i$*. Then* $S_n \in \mathcal{G}(\sum_{i=1}^n \alpha_i^2 \nu_i)$ *and for all* $\beta > 0$*,*

$$\mathbb{P}[S_n \geq \beta] \leq \exp\left(-\frac{\beta^2}{2\sum_{i=1}^n \alpha_i^2 \nu_i}\right).$$

Proof: By independence,

$$\Psi_{S_n}(s) = \sum_{i \leq n} \Psi_{\alpha_i X_i}(s) = \sum_{i \leq n} \Psi_{X_i}(s\alpha_i) \leq \sum_{i \leq n} \frac{(s\alpha_i)^2 \nu_i}{2} = \frac{s^2 \sum_{i \leq n} \alpha_i^2 \nu_i}{2}.$$

■

Recall: Epsilon-nets

Definition (ε-net)

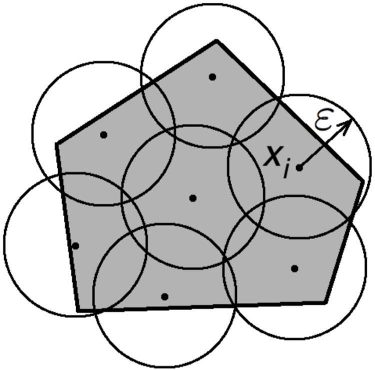
Let T be a subset of a pseudometric space (M, ρ) and let $\varepsilon > 0$. The collection of points $N \subseteq M$ is called an *ε-net of T* if

$$T \subseteq \bigcup_{t \in N} B_\rho(t, \varepsilon),$$

where $B_\rho(t, \varepsilon) = \{s \in T : \rho(s, t) \leq \varepsilon\}$, that is, each element of T is within distance ε of an element in N. The smallest cardinality of an ε-net of T is called the *covering number*

$$\mathcal{N}(T, \rho, \varepsilon) = \inf\{|N| : N \text{ is an } \varepsilon\text{-net of } T\}.$$

Recall: Epsilon-nets by picture



(a) This covering of a pentagon K by seven ε-balls shows that $\mathcal{N}(K, \varepsilon) \leq 7$.

Recall: Epsilon-net on sphere

Let $\mathbb{S}^{k-1}$ be the sphere of radius 1 centered around the origin in $\mathbb{R}^k$ with the Euclidean metric. Let $0 < \varepsilon < 1$. We claim that

$$\mathcal{N}(S, \rho, \varepsilon) \leq \left(\frac{3}{\varepsilon}\right)^k.$$

Let N be any ε-net of S. The balls of radius $\varepsilon/2$ around points in N, $\{\mathbb{B}^k(x_i, \varepsilon/2) \,:\, x_i \in N\}$, satisfy two properties:

1. Pairwise disjoint: if $z \in \mathbb{B}^k(x_i, \varepsilon/2) \cap \mathbb{B}^k(x_j, \varepsilon/2)$, then $\|x_i - x_j\|_2 \leq \|x_i - z\|_2 + \|x_j - z\|_2 \leq \varepsilon$, a contradiction.
2. Contained in $\mathbb{B}^k(0, 3/2)$: if $z \in \mathbb{B}^k(x_i, \varepsilon/2)$, then $\|z\|_2 \leq \|z - x_i\|_2 + \|x_i\| \leq \varepsilon/2 + 1 \leq 3/2$.

The volume of a ball of radius is $\varepsilon/2$ is $\frac{\pi^{k/2}(\varepsilon/2)^k}{\Gamma(k/2+1)}$ and that of a ball of radius $3/2$ is $\frac{\pi^{k/2}(3/2)^k}{\Gamma(k/2+1)}$. Divide one by the other.

Spectral norm of random matrix I

For a $m \times n$ matrix $A \in \mathbb{R}^{m \times n}$, recall that the spectral norm is defined as

$$\|A\| := \sup_{\mathbf{x} \in \mathbb{R}^n \setminus \{0\}} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \sup_{\mathbf{x} \in \mathbb{S}^{n-1}} \|A\mathbf{x}\|_2 = \sup_{\substack{\mathbf{x} \in \mathbb{S}^{n-1} \\ \mathbf{y} \in \mathbb{S}^{m-1}}} \langle A\mathbf{x}, \mathbf{y} \rangle,$$

where $\mathbb{S}^{n-1}$ is the sphere of radius 1 around the origin in $\mathbb{R}^n$.

(To see the rightmost equality above, note that Cauchy-Schwarz implies $\langle A\mathbf{x}, \mathbf{y} \rangle \leq \|A\mathbf{x}\|_2 \|\mathbf{y}\|_2$ and that one can take $\mathbf{y} = A\mathbf{x}/\|A\mathbf{x}\|_2$ for any $\mathbf{x}$ such that $A\mathbf{x} \neq 0$ in the rightmost expression.)

Spectral norm of random matrix II

Theorem

Let $A \in \mathbb{R}^{m \times n}$ be a random matrix whose entries are centered, independent and sub-Gaussian with variance factor ν. Then there exist a constant $0 < C < +\infty$ such that, for all $t > 0$,

$$\|A\| \leq C\sqrt{\nu}(\sqrt{m} + \sqrt{n} + t),$$

with probability at least $1 - e^{-t^2}$.

Without independence of the entries, the spectral norm can be much larger. Say A is all-$(+1)$ or all-(-1) with equal probability. Taking the vector $\mathbf{x} = (1/\sqrt{n}, \ldots, 1/\sqrt{n})$ shows that $\|A\| \geq n$ with probability 1.

Spectral norm of random matrix III

Proof: We seek to bound

$$\|A\| = \sup_{\substack{\mathbf{x}\in\mathbb{S}^{n-1}\\ \mathbf{y}\in\mathbb{S}^{m-1}}} \langle A\mathbf{x}, \mathbf{y}\rangle = \sup_{\substack{\mathbf{x}\in\mathbb{S}^{n-1}\\ \mathbf{y}\in\mathbb{S}^{m-1}}} \sum_{i,j} x_i y_j A_{ij},$$

where we note that the last quantity is a linear combination of independent variables. Fix $\varepsilon = 1/4$. We proceed in two steps:

1. We first apply the general Hoeffding inequality to control the deviations of the supremum *restricted to* ε*-nets* N *and* M *of* $\mathbb{S}^{n-1}$ *and* $\mathbb{S}^{m-1}$.
2. We then extend the bound to the full supremum by continuity.

Spectral norm of random matrix IV

Lemma

Let N and M be as above. For C large enough, for all $t > 0$,

$$\mathbb{P}\left[\max_{\substack{\mathbf{x}\in N\\ \mathbf{y}\in M}} \langle A\mathbf{x}, \mathbf{y}\rangle \geq \frac{1}{2} C\sqrt{\nu}(\sqrt{m} + \sqrt{n} + t)\right] \leq e^{-t^2}.$$

Proof: By the general Hoeffding inequality, $\langle A\mathbf{x}, \mathbf{y}\rangle$ is sub-Gaussian with variance factor

$$\sum_{i,j} (x_i y_j)^2 \nu = \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 \nu = \nu,$$

for all $\mathbf{x} \in N$ and $\mathbf{y} \in M$. In particular, for all $\beta > 0$,

$$\mathbb{P}[\langle A\mathbf{x}, \mathbf{y}\rangle \geq \beta] \leq \exp\left(-\frac{\beta^2}{2\nu}\right).$$

Spectral norm of random matrix V

Proof of lemma (continued): Hence, by a union bound over N and M,

$$\begin{aligned}
&\mathbb{P}\left[\max_{\substack{\mathbf{x}\in N\\ \mathbf{y}\in M}} \langle A\mathbf{x}, \mathbf{y}\rangle \geq \frac{1}{2} C\sqrt{\nu}(\sqrt{m}+\sqrt{n}+t)\right] \\
&\quad\leq \sum_{\substack{\mathbf{x}\in N\\ \mathbf{y}\in M}} \mathbb{P}\left[\langle A\mathbf{x}, \mathbf{y}\rangle \geq \frac{1}{2} C\sqrt{\nu}(\sqrt{m}+\sqrt{n}+t)\right] \\
&\quad\leq |N||M| \exp\left(-\frac{1}{2\nu}\left\{\frac{1}{2} C\sqrt{\nu}(\sqrt{m}+\sqrt{n}+t)\right\}^2\right) \\
&\quad\leq 12^{n+m} \exp\left(-\frac{C^2}{8}\left\{m+n+t^2)\right\}\right) \\
&\quad\leq e^{-t^2},
\end{aligned}$$

for $C^2/8 = \log 12 \geq 1$, where in the third inequality we ignored all cross-products since they are non-negative. ∎

Spectral norm of random matrix VI

Lemma

For any ε-nets N and M of $\mathbb{S}^{n-1}$ and $\mathbb{S}^{m-1}$ respectively, the following inequalities hold

$$\sup_{\substack{\mathbf{x}\in N\\ \mathbf{y}\in M}} \langle A\mathbf{x}, \mathbf{y}\rangle \leq \|A\| \leq \frac{1}{1-2\varepsilon} \sup_{\substack{\mathbf{x}\in N\\ \mathbf{y}\in M}} \langle A\mathbf{x}, \mathbf{y}\rangle.$$

Proof: The first inequality is immediate. For the second inequality, we will use the following observation

$$\langle A\mathbf{x}, \mathbf{y}\rangle - \langle A\mathbf{x}_0, \mathbf{y}_0\rangle = \langle A\mathbf{x}, \mathbf{y}-\mathbf{y}_0\rangle + \langle A(\mathbf{x}-\mathbf{x}_0), \mathbf{y}_0\rangle.$$

Fix $x \in \mathbb{S}^{n-1}$ and $y \in \mathbb{S}^{m-1}$ such that $\langle A\mathbf{x}, \mathbf{y}\rangle = \|A\|$, and let $\mathbf{x}_0 \in N$ and $\mathbf{y}_0 \in M$ such that

$$\|\mathbf{x}-\mathbf{x}_0\|_2 \leq \varepsilon \qquad \text{and} \qquad \|\mathbf{y}-\mathbf{y}_0\|_2 \leq \varepsilon.$$

Spectral norm of random matrix VII

Proof of lemma (continued): Then the inequality above, Cauchy-Schwarz and the definition of the spectral norm imply

$$\|A\| - \langle A\mathbf{x}_0, \mathbf{y}_0\rangle \leq \|A\|\|\mathbf{x}\|_2\|\mathbf{y} - \mathbf{y}_0\|_2 + \|A\|\|\mathbf{x} - \mathbf{x}_0\|_2\|\mathbf{y}_0\|_2 \leq 2\varepsilon\|A\|.$$

Rearranging gives the claim. ∎

Application: Back to the SBM

Lemma

Let $G \sim \mathrm{SBM}_{n,p,q}$, let A be the adjacency matrix of G and let $D = \mathbb{E}_{n,p,q}[A]$. Then, there is a constant $C > 0$ such that

$$\|A - D\| \leq C\sqrt{n},$$

with probability at least $1 - e^{-n}$.

Proof: The entries of R are centered, independent and sub-Gaussian with variance factor $1/4$. ■

Go deeper

Course website:

http://www.math.wisc.edu/~roch/mdp/

For more on community detection, see e.g. (available online):

- *Community Detection and Stochastic Block Models* by Abbé