CS247: Advanced Data Mining (Winter 2024) February 7, 2024

Assignment 4 (Part 2)

Due Date: **February 16, 2023**

Instructions

- Submit your answer on Gradescope as a PDF file. Both typed and scanned handwritten answers are acceptable.
- Submit your solutions to Part 1 and Part 2 through GradeScope in BruinLearn separately.
- Late submissions are allowed up to 24 hours post-deadline with a penalty factor of $\mathbf{1}(t \leq 24)e^{-(\ln(2)/12)t}$.
- Ensure all sources are cited appropriately; plagiarism will be reported.

Problems

In this part, we follow the notations below:

- Let $G = (V, E)$ be a simple (that is, no self- or multi- edges) undirected, connected graph with $n = |V|$ and $m = |E|$.
- $\boldsymbol{A}$ is the adjacency matrix of the graph G, i.e., $\boldsymbol{A}_{ij}$ is equal to 1 if $(i, j) \in E$ and equal to 0 otherwise.
- $\boldsymbol{D}$ is the diagonal matrix of degrees: $\boldsymbol{D}_{ii} = \sum_j \boldsymbol{A}_{ij} = d_i$, where d_i is the degree of node i.
- We define the graph Laplacian of G by $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$.

For a set of nodes $S \subset V$, we will measure the quality of S as a cluster with a "cut" value and a "volume" value. We define the cut of the set S to be the number of edges that have one end point in S and one end point in the complement set $\bar{S} = V \backslash S$:

$$\text{cut}(S) = \sum_{i \in S, j \in \bar{S}} \boldsymbol{A}_{ij}. \tag{1}$$

Note that the cut is symmetric in the sense that $\text{cut}(S) = \text{cut}(\bar{S})$. The volume of S is simply the sum of degrees of nodes in S:

$$\text{vol}(S) = \sum_{i \in S} d_i. \tag{2}$$

Problem 1: Normalized Cuts (15 points)

Consider the problem of partitioning a graph G into two subgraphs with similar sizes. Let $S \subset V$ and $\bar{S} = V \backslash S$ denote the disjoint node sets of the two clusters. The normalized cut between two clusters is defined as:

$$\text{ncut}(S) = \frac{\text{cut}(S)}{\text{vol}(S)} + \frac{\text{cut}(\bar{S})}{\text{vol}(\bar{S})}. \tag{3}$$

Intuitively, a set S with a small normalized cut value must have few edges connecting to the rest of the graph (making the numerators small) as well as some balance in the size of the clusters (making the denominators large).

Define the assignment vector x for some set of nodes S such that

$$x_i = \begin{cases} \sqrt{\dfrac{\mathrm{vol}(\bar{S})}{\mathrm{vol}(S)}}, & i \in S \\ -\sqrt{\dfrac{\mathrm{vol}(S)}{\mathrm{vol}(\bar{S})}}, & i \in \bar{S} \end{cases} \tag{4}$$

Please prove the following properties.

1. $\boldsymbol{L} = \sum_{(i,j)\in E}(\mathbf{1}_i - \mathbf{1}_j)(\mathbf{1}_i - \mathbf{1}_j)^\intercal$, where $\mathbf{1}_k$ is an n-dimensional column vector with a 1 at position k and 0's elsewhere. Note that we are not summing over the entire adjacency matrix and only count each edge once.
2. $\boldsymbol{x}^\intercal \boldsymbol{L} \boldsymbol{x} = \sum_{(i,j)\in E}(x_i - x_j)^2$.
3. $\boldsymbol{x}^\intercal \boldsymbol{L} \boldsymbol{x} = c \cdot \mathrm{ncut}(S)$ for some constant c. Hint: Rewrite the sum in terms of S and $\bar{S}$.
4. $\boldsymbol{x}^\intercal \boldsymbol{D} \mathbf{1} = 0$, where $\mathbf{1}$ is the vector of all ones.
5. $\boldsymbol{x}^\intercal \boldsymbol{D} \boldsymbol{x} = 2m$.

Answer:

Problem 2: Solution to Normalized Cut Minimization (15 points)

Since $\boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{x}$ is just a constant ($2m$), we can formulate the normalized cut minimization problem in the following way:

$$\begin{aligned} \underset{S\subset V,\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \quad & \frac{\boldsymbol{x}^\top \boldsymbol{L}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{x}} \\ \text{subject to} \quad & \boldsymbol{x}^\top \boldsymbol{D}\mathbf{1} = 0, \\ & \boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{x} = 2m, \\ & \boldsymbol{x} \text{ is defined as in Equation (4).} \end{aligned} \tag{5}$$

The constraint that $\boldsymbol{x}$ takes the form of Equation (4) makes the optimization problem NP-hard. We will instead use the "relax and round" technique, where we relax the problem to make the optimization problem tractable and then round the relaxed solution back to a feasible point for the original problem.

Our relaxed problem will eliminate the constraint that $\boldsymbol{x}$ takes the form of Equation (4) which leads to the following relaxed problem:

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \quad & \frac{\boldsymbol{x}^\top \boldsymbol{L}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{x}} \\ \text{subject to} \quad & \boldsymbol{x}^\top \boldsymbol{D}\mathbf{1} = 0, \\ & \boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{x} = 2m. \end{aligned} \tag{6}$$

Please prove that the minimizer of Problem (6) is $\boldsymbol{D}^{-1/2}\boldsymbol{v}$, where $\boldsymbol{v}$ is the eigenvector corresponding to the second smallest eigenvalue of the normalized graph Laplacian $\tilde{\boldsymbol{L}} = \boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2}$.

Finally, to round the solution back to a feasible point in the original problem, we can take the indices of all positive entries of the eigenvector to be the set S and the indices of all negative entries to be $\bar{S}$.

Hints:

- Make the substitution $\boldsymbol{z} = \boldsymbol{D}^{1/2}\boldsymbol{x}$.
- Note that $\mathbf{1}$ is the eigenvector corresponding to the smallest eigenvalue of $\boldsymbol{L}$.
- The normalized graph Laplacian $\tilde{\boldsymbol{L}}$ is symmetric, so its eigenvectors are orthonormal and form a basis for $\mathbb{R}^n$. This means we can write any vector $\boldsymbol{x}$ as a linear combination of orthonormal eigenvectors of $\tilde{\boldsymbol{L}}$.

Answer: