Problem Set 1

Spectral clustering and community detection

1. Show that Lloyd's k-mean algorithm converges in finitely many steps.

2. Let $x_1, x_2, \ldots, x_n$ be points in $\mathbb{R}^d$. Define an $n \times n$ weight matrix W with

$$W(i,j) = \exp\left(-\|x_i - x_j\|_2^2/2\right).$$

   Show that $W \succcurlyeq 0$, that is, W is a positive semi-definite matrix.

3. Let G be a finite undirected, unweighted graph. Let $\mathcal{L} = D^{-1/2}(D-A)D^{-1/2}$ be its normalized Laplacian, where A is the adjacency matrix of the graph. Suppose λ_n be its maximum eigenvalue.

   (a) Show that

$$\lambda_n = \max_{x \neq 0} \frac{\sum_{i \sim j}(x_i - x_j)^2}{\sum_i d_i x_i^2} \leq 2,$$

   where d_i is the degree of the vertex i.

   (b) Prove that $\lambda_n = 2$ if and only if G has a bipartite connected component.

   (c) Give an example of a non-bipartite (and disconnected) graph with $\lambda_n = 2$.

4. (a) Let G be a connected, unweighted graph, λ_2 the second smallest eigenvalue of the normalized Laplacian $\mathcal{L}$ and $\mathrm{diam}(G)$ the diameter of G. Then

$$\lambda_2 \geq \frac{1}{\mathrm{diam}(G)\mathrm{vol}(G)},$$

   where $\mathrm{vol}(G) = \sum_i d(i)$.
   Hint: Use the Courant-Fischer characterization for λ_2:

$$\lambda_2 = \inf_{x \neq 0 : \sum_i d_i x_i = 0} \frac{\sum_{i \sim j}(x_i - x_j)^2}{\sum_i d_i x_i^2}.$$

   (b) Consider the dumbbell graph of $2n$. It is defined as the disjoint union of two complete graphs K_n connected by a single edge. Use part (a) to show that for this graph,

$$\lambda_2 \geq c n^{-2},$$

   for some constant $c > 0$ independent of n.

5. Let G be an undirected d-regular graph. The Cheeger's inequality (hard direction) can be generalized as follows. Let z be any vector orthogonal to 1. Let S be subset obtained by performing the sweep cut on z. Then

$$\phi(S) \leq \sqrt{2R(z)},$$

   where

$$R(z) = \frac{\sum_{i \sim j}(z_i - z_j)^2}{d \sum_i z_i^2}.$$

   The following modifications of the proof are needed to obtain $y \geq 0$ such that $\mathrm{supp}(y) \leq n/2$ and $R(y) \leq R(z)$, on which we apply the key lemma (randomized rounding) as before.

   (a) Show that $R(z - c1) \leq R(z)$.

(b) Define $x = z - m\mathbf{1}$ where m is the median of the values of z. By definition, x has at most $n/2$ positive values and at most $n/2$ negative values.

(c) Set $x^+ = \max(x, 0)$ and $x^- = \max(-x, 0)$. Note that $x = x^+ - x^-$. Prove that

$$\min(R(x^+), R(x^-)) \leq R(x).$$

If $R(x^+) \leq R(x)$ take $y = x^+$. Otherwise, we must have $R(x^-) \leq R(x)$ and we take $y = x^-$.