CS 189: Introduction to Machine Learning - Discussion 12

1. Spectral clustering. In this question we will provide some intuition on spectral clustering in the context of simple undirected and regular graphs. Consider a $d-$regular graph $G = (V, E)$ of n vertices and m edges. The adjacency matrix of a graph is $A \in \mathbb{R}^{n\times n}$ matrix such that:

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

The normalized Laplacian of the graph G is $L = I - \frac{1}{d}A$.

a) Using the notation from lecture. If we set $w_{j,i} = w_{i,j} = \frac{1}{d}$ for all $(i,j) \in E$. Check that the following is an alternative definition for L:

$$L_{i,j} = \begin{cases} -w_{i,j} & \text{if } (i,j) \in E \\ \sum_{j|(i,j)\in E} w_{i,j} & \text{if } i = j \\ 0 & \text{o.w.} \end{cases}$$

Show also that the all ones vector 1 is an eigenvector for eigenvalue 0 of L.

Solution:
Clearly the off diagonal entries agree, since these equal the negative of the entries of the Adjacency matrix multiplied by the weights $\frac{1}{d}$. For the diagonal entries notice that since the degree for every vertex is d, then the sum $\sum_{j|(i,j)\in E} w_{i,j}$ is always over d terms each equal to $\frac{1}{d}$. From this representation it follows immediately that $L1 = 0$.

b) Show that for any vector $x \in \mathbb{R}^n$, $x^T L x = \frac{1}{d}\sum_{(i,j)\in E}(x_i - x_j)^2$.

Solution:
Consider $x^T(Id - A)x$. Expanding yields: $d\sum_{i=1}^n x_i^2 - x^T A x = d\sum_{i=1}^n x_i^2 - 2\sum_{(i,j)\in E} x_i x_j = \sum_{(i,j)\in E}(x_i - x_j)^2$. The last equality follows because the degree of each node equals d. The desired result follows.

c) Show that L is positive semidefinite.

Solution:
It follows immediately from part b).

d) Show that the number of zero eigenvalues of L equals the number of connected components of G. How does this relate to clustering?

Solution:
First we show the first direction. If L has k connected components then L has at least k eigenvalues equal to zero.

Observation 1: Since L is positive semidefinite, x is an eigenvector for eigenvalue 0 iff $x^T Lx = 0$. Indeed, taking a orthonormal basis of eigenvectors of L, $u_1, \cdots, u_n$ with eigenvalues $\lambda_1, \cdots, \lambda_n$, we can write $x = \sum_{i=1}^n \alpha_i u_i$ for some coefficients α_i. Then $x^T Lx = \sum_{i=1}^n \lambda_i \alpha_i^2$. Since all $\lambda_i \geq 0$ because L is PSD, the only time when $x^T Lx = 0$ is when $\alpha_i = 0$ for all eigenvalues $\lambda_i \neq 0$. In other words, x is in the eigenspace of 0.

We show that if the connected components are $C_1, \cdots, C_k$, then the indicator vectors $\{1_{C_i}\}$ have all eigenvalue 0. Indeed, $1_{C_i}^T L 1_{C_i} = 0 \ \forall i$, which, by the observation in the previous paragraph means that all 1_{C_i} are eigenvectors with eigenvalue 0. Since all vectors $\{1_{C_i}\}$ are independent (they are orthogonal), then their span is k dimensional, implying that the dimension of the eigenspace of 0 is at least k dimensional.

We have proven:

$$\text{num connected components} \ \leq \ \text{dim eigenspace of eigenval 0 of } L$$

Now we show the second direction. If L has k eigenvalues equal to 0 then it has at least k connected components.

By b) and Observation 1, an eigenvector x of 0 must have $x_a = x_b$ for all pairs of vertices $a, b \in C_i$. In other words, the eigenvectors of 0 are constant within each connected component of G. The effective dimension of the 0 eigenvectors is at most the number of connected components. Since there are k connected components, the dimension of the eigenspace cannot exceed k. This proves:

$$\text{num connected components} \ \geq \ \text{dim eigenspace of eigenval 0 of } L$$

The desired result follows.

It is reasonable to expect each connected component to be a cluster.

e) (Optional) Recall the variational representation of the eigenvalues of L:

$$\lambda_k = \min_{S \text{ } k\text{dimensional subspace of } \mathbb{R}^n} \max_{x \in S - \{0\}} \frac{x^T L x}{x^T x}$$

Show that the eigenvalues of L are between 0 and 2. This justifies the use of the normalized Laplacian (the eigenvalues do not blowup with degree/dimension).

Solution: By positive semidefiniteness of L all eigenvalues are nonnegative. Notice that scaling of x doesn't affect the objective $\frac{x^T L x}{x^T x}$. So we can restrict ourselves to the unit ball.
The objective equals by part b) $\frac{1}{d} \sum_{(i,j) \in E} (x_i - x_j)^2$ with $\| x \| = 1$.
To be finished.

f) You are given a connected $d-$regular graph $G = (V, E)$ and are told that there is a partition (V_0, V_1) of the vertices $|V_0| = |V_1| = |V|/2$ such that every node in V_j has d_{in} neighbors within V_j and d_{out} neighbors in V_{1-j} with $d_{in} > d_{out}$, for $j = 0, 1$. You are also told that, if $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ are the normalized Laplacian eigenvalues, $\lambda_3 > 2d_{out}/d$. Describe an algorithm to find (V_0, V_1) in polynomial time. Why choosing V_0 and V_1 as our two clusters is a reasonable cluster partition for the two clusters case?

Solution:
We see that $\frac{2d_{out}}{d}$ is an eigenvalue of the normalized Laplacian.
Let $v = [v_0^T, v_1^T]^T$ where $v \in \mathbb{R}^{|V|}$ and $v_0 \in \mathbb{R}^{|V_0|}$ and $v_1 \in \mathbb{R}^{|V_1|}$.
Let $v_0 = [a, \cdots, a]^T$ and $v_1 = [b, \cdots, b]^T$.
By looking at the adjacency matrix A of G, one can see that the first $|V|/2$ entries of $(Av)_0 = d_{in} v_0 + d_{out} v_1$ and $(Av)_1 = d_{in} v_1 + d_{out} v_0$.
And therefore if v is an eigenvector of λ it must be the case that:

$$\lambda a = d_{in} a + d_{out} b$$
$$\lambda b = d_{in} b + d_{out} a$$

and therefore if we assume $\lambda \neq 0$:
$\lambda = d_{in} - d_{out}$
and after a bit of algebra $a^2 = b^2$ so that to obtain the second eigenvalue, we want that $a = -b$. And since normalization is irrelevant we can choose for example $a = 1$, $b = -1$

If $d_{in} - d_{out}$ is an eigenvalue of the adjacency matrix, then $1 - \frac{d_{in} - d_{out}}{d} = \frac{2d_{out}}{d}$ is an eigenvalue for the normalized Laplacian with eigenvalue $v = [v_0^T, v_1^T]^T$ with $v_0 = [1, \cdots, 1]^T$ and $v_1 = [-1, \cdots, -1]^T$.
Because $\lambda_3 > \lambda_2 = \frac{2d_{out}}{d}$ the eigenspace for $\frac{2d_{out}}{d}$ has dimension 1, so all eigenvectors for this eigenvalue are unique up to scaling. Solving for $\hat{v}$, ($L\hat{v} = \frac{2d_{out}}{d}\hat{v}$) we can discern between V_0 and V_1 by splitting the nodes of V along the positive and negative coordinates of the eigenvector $\hat{v}$.

2. K-means.

Recall the K-means algorithm:

1. Initialize k cluster centers c_k.
2. For each $x^{(i)}$, assign cluster with closest center $c_{\hat{k}}$ s.t. $\hat{k} = \arg\min_k d(x, c_k)$ for some distance function d.
3. For each cluster, recompute center $c_k = \frac{1}{n_k}\sum_{x \in C_k} x$ where n_k is the number of points currently assigned to cluster k.
4. Check convergence. If not converged, go to 2.

Now assume data generated using the following procedure:

1. Pick one of k $m-$dimensional mean vectors $z_1, \cdots, z_k$ according to probability distribution $p(j)$. This selects a (hidden) class label j. Suppose that $p(j) = \frac{1}{k}$.
2. Generate a data point by sampling from $p(x|i) \sim N(z_i, \sigma^2 I_n)$.

a) Under the data generation procedure described above. What is the probability distribution of a single point, $p(x^{(i)})$?

Solution:

$$p(x^{(i)}) = \sum_{j=1}^{k} p(j)p(x^{(i)}|j)$$

b) Suppose $z_1, \cdots, z_k$ are not known. We are given independent samples $x^{(i)}$ along with their corresponding generating class $y^{(i)} \in \{1, \cdots, k\}$. What is $\log(P(\{x^{(i)}\}|z_1, \cdots, z_k))$? What is the ML estimator of the means and how does it relate to previous topics in the course? What is the relationship between this and k means?

Solution:
Let C_j be the set of points $x^{(i)}$ that have class $y^{(i)} = j$.

$$\log(P(\{x^{(i)}\}|\text{ guess for } z_1, \cdots, z_k)) = \text{const} - \frac{1}{2}\sum_{i=1}^{k}\sum_{x \in C_i}(x - z_i)^T(x - z_i)$$

The ML estimator for $z_1, \cdots, z_k$ corresponds to $z_i = \frac{1}{|C_i|}\sum_{x \in C_i} x$. Same as in LDA. This corresponds to the averaging step of the K-means algorithm.

c) Now suppose we are not given the generating class $y^{(i)}$ but the means $z_1, \cdots, z_k$ are known. What is $\log(P(x^{(1)}, \cdots, x^{(n)}|$ guess for $y^{(1)}, \cdots, y^{(n)}))$? What is the ML estimator of the class labels? What is the relationship between this and k means?

Solution:

$$\log(P(x^{(i)}|\text{ guess for } y^{(1)}, \cdots, y^{(n)})) = \text{const} - (x^{(i)} - z_{y^{(i)}})^2$$

Therefore:

$$\log(P(\{x^{(i)}\}|\text{ guess for } y^{(1)}, \cdots, y^{(n)})) = \text{const} - \sum_{i=1}^{n}(x^{(i)} - z_{y^{(i)}})^2$$

The ML estimator for $y^{(1)}, \cdots, y^{(n)}$ is: $y^{(i)}_{ML} = \arg\min_j (x^i - z_j)^2 \ \forall i$. This corresponds to finding the closest mean step of the K-means algorithm.