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.

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

c) Show that L is positive semidefinite.

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

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).

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?

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, $p(x^{(i)})$?

   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?

   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?