Math 690: Topics in Data Analysis and Computation
Lecture notes for October 3, 2017

Scribed by Dev Dabke and Andrew Cho

1 Introduction

The lecture covered the following:

- Discuss KK10 paper on consistency of Lloyd's k-means algorithm
- Start spectral clustering

2 k-means clustering

2.1 Lloyd's algorithm

Recall Lloyd's algorithm from last lecture that given some data points as input $\{x_i\}_{i=1}^n$, we want to search for k clusters $C = \{C_1, \ldots, C_k\}$ with "centers" $\mu = \{\mu_1, \ldots, \mu_k\}$.

Depending on our constraints, also recall we can choose different objective functions and have different interpretations of what these clusters represent.

$$\min_{C,\mu} \sum_{l=1}^{k} \sum_{i \in C_l} \|x_i - \mu_l\|_2^2 \qquad \text{k means}$$

$$\min_{C,\mu} \sum_{l=1}^{k} \sum_{i \in C_l} \|x_i - \mu_l\|_1 \qquad \text{k medians}$$

$$\min_{C,\mu} \sum_{l=1}^{k} \sum_{i \in C_l} \|x_i - \mu_l\|_2 \qquad L_2 \text{ - } L_1 \text{ norm}$$

We can either choose initial seeds by randomly selecting k points or by singular value decomposition (SVD). Ultimately, we are interested in knowing if these results are *consistent*. Can these seeding methods lead to the true centroids?

2.2 Strong Law of Large Numbers

$$\mathcal{L}_n(\mu) = \int \|x - \mu\|^2 dP_n(x)$$

where

$$dP_n(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}(x) dx$$

which is the empirical measure of the data.

Where $A = \{a_1, \ldots, a_n\}$ for all $x \in \mathbb{R}$, we define $\|x - A\|$ to be

$$\|x - A\| = \min_{1 \leq i \leq n} \|x - a_i\|$$

Now, note that as $n \to \infty$ that

$$\mathcal{L}(\mu) = \int \|x - \mu\|^2 dP(x)$$

and that

$$\min_{\mu} \mathcal{L}_n(\mu) \to \min_{\mu} \mathcal{L}(\mu)$$

2.3 Consistency of Lloyd

Ref (Kumar, Amit, and Ravindran Kannan 2010)

If we satisfy some requirements of "separation" to provide some bound for misclassification, assuming some true centroid partition.

$\|\mu_k^* - \mu_l^*\|$ if $k \neq l$ needs to be bigger than a gap Δ_{kl}.

$$\Delta_{kl} > c \frac{\sigma_l}{\sqrt{w_{min}}} \qquad w_{min} = \min_{l} \frac{|C_l^*|}{n}$$

w_{min} denotes proportions of points in cluster l. Letting σ_k^2 be the variance of the data in the k^{th} cluster and if we can assume that $\sigma_k \approx \sigma_l$, then we know that misclassification error from SVD initialization occurs less than some $\epsilon * n$ with high probability.

2.4 In Practice

There are some practical considerations for k-means.

1. what is k?
2. how to choose an initial seed
3. some cases of k-means fails
   (a) σ_k might be too large compared to separation
   (b) clusters might be too small (i.e. cluster sizes may not be balanced)
   (c) cluster might be convex and piecewise can't do concave (Voronoi)

3 Spectral clustering

3.1 Graph Laplacian

Given some data $\{x_i\}_{i=1}^n$ and k, we want to

1. Build positive-definite, symmetric affinity matrix $W_{n \times n}$ by k nearest neighbors, ϵ - neighbor, or the Gaussian kernel $W_{ij} = e^{-\frac{\|x_i - x_j\|^2}{\epsilon}}$.

2. Consider the eigenvalue decomposition of the graph Laplacian $\mathcal{L}$. Note that

$$\begin{aligned} L_{un} &= D - W && \text{unnormalized} \\ L_{rw} &= D^{-1}(D - W) && \text{Shi-Malik '00} \\ L_{sym} &= I - D^{-1/2} W D^{-1/2} && \text{Ng-Jordan-Weiss '02} \end{aligned}$$

3. Apply k means to $\Psi = [\varphi_i, \ldots, \varphi_k]_{n \times k}$ and denote y_i as the i^{th} row of Ψ.

Remark. *If $k = 2$, you can use truncation because the first eigenvalue is constant. Thus, you can use sign(Ψ_2) to indicate clustering.*

Definition (Connected Components)**.** *Let $\mathcal{G} = (V, E)$ such that on each edge $(i, j) \in \mathcal{G}$ that $w_{ij} > 0$. If node i is connected to node j, there is a path from i to j. So then, set A is a connected component if every pair of i and j is connected and A is the maximum set that satisfies this condition to preserve connectivity.*

Proposition (Eigenspace of $\lambda = 0$ of $\mathcal{L}$)**.** *Suppose the graph has k connected components $A_1, \ldots, A_k$. Then the eigenspace of $\lambda = 0$ of $\dim k$ is spanned by*

$$\{\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_k}\}$$

where

$$\mathbf{1}_{A_i}(i) = \begin{cases} 1 & i \in A \\ 0 & otherwise \end{cases}$$

Proof Suppose $\mathbf{f}$ is an eigenvector with $\lambda = 0$. So $\mathcal{L}\mathbf{f} = \mathbf{0}$ and $\mathbf{f}^T \mathcal{L}\mathbf{f} = \mathbf{0}$. We also know $\mathbf{f}^T \mathcal{L}\mathbf{f} = \mathbf{0} = \frac{1}{2} \sum_{(i,j)} w_{i,j}(f_i - f_j)^2$ and $\mathbf{f}^T \mathcal{L}\mathbf{f} = \mathbf{0} \Leftrightarrow f_i = f_j$ whenever $w_{i,j}$>0. Thus, $\mathbf{f}$ is a piecewise constant in each of the connected components. Meanwhile, for each $v \in$ span $\{\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_k}\}, \mathbf{v}^T \mathcal{L}\mathbf{v} = \mathbf{0}$.

Exercise:

- Does this generalize to $\mathcal{L}_{rw}$ and $\mathcal{L}_{sym}$?
- What about consistency of spectral clustering?