10/3 Clustering scribe (cont.)

recall

k-means

Input: $\{x_i\}_{i=1}^n$

k

search for $C = \{C_1, \dots, C_k\}$

$\mu = \{\mu_1, \dots, \mu_k\}$

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

$$\min_{C,\mu} \sum \sum \|x_i - \mu_l\|_+ \quad \text{(k medians)}$$

$$\min_{C,\mu} \sum \sum \|x_i - \mu_l\|_2 \quad (L_2\text{-}L_1 \text{ norm})$$

Initial seeding

① random

② by singular value decomposition

Consistency — can random seeding lead to of kmeans the true μ centroids??

If C^*, μ^* are the true minimizers.

Strong Law of Large of Numbers (1960-80)

$$L_n(\mu) = \int \|x - \mu\|^2 dP_n(x).$$

$dP_n(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}(x) dx$ # empirical measure of data

def $A = \{a_1, \dots, a_n\}$, $x \in \mathbb{R}^d$

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

As $L_n(\mu)$ $n \to \infty$, $L(\mu)$ and

$L(\mu) = \int \|x - \mu\|^2 dP(x)$, but does the minimizer converge?

Consistency of Lloyd's + New analysis

(KK '10, McSherry '01)

If

↳ requirement of "separation"

⇒ bound for misclassification error.

assuming some true centroid partition

$\|\mu_k^* - \mu_l^*\|$ if $k \neq l$ needs to be bigger

than a gap Δ_{kl}

$\|\mu_k^* - \mu_l^*\| > \Delta_{kl}$ where $\Delta_{kl} > c \times \frac{\sigma_k}{\sqrt{w_{min}}}$ and $w_{min} = \min_l \frac{|C_l^*|}{n}$ ← number of points in set C_l ↳ proportion

σ_k^2 = variance of data in k^{th} cluster.

and if we can assume $\sigma_k \approx \sigma_l$.

Thm: Lloyd's and SVD misclassify $\le \varepsilon \cdot n$ points w/ high probability

Practical considerations of k-means?

- what is K?
- how to initial seed
- failing cases of k-means

diameter → σ_k might be too large compared to separation $\|\mu_k - \mu_l\|$

cluster size → too small cluster. If cluster sizes are not evenly balanced

Voronoi cells → cluster must be convex and piecewise. can't do concave.

not the same δ as tolerance.

$\min_{\mu} L_n(\mu) \to \min_{\mu} L(\mu)$

but this doesn't give us the misclassification rate

Spectral Clustering

Given $\{x_i\}_{i=1}^n$ and k,

① Build affinity matrix $W_{n\times n}$
by knn, ε-neighborhood,
or (required) $w_{ij} = e^{-\frac{|x_i - x_j|^2}{\varepsilon}}$

W has to be positive and symmetric

② Consider the eig decomp of $\mathcal{L}$ graph Laplacian

- $L_{un} = D - W$ (unnormalized)
- $L_{rw} = D^{-1}(D-W) = I - P$ (Shi-Malik '00)
- L_{sym} Compute the first k (Ng-Jordan-Weiss '02) eigval and eigvec of the

③ Apply k-means to $\underline{\Psi} = [\psi_1, \dots, \psi_k]_{n\times k}$

y_i is the i^{th} row of $\underline{\Psi}$

Remark: If $k=2$, you can use truncation of truncation of $\operatorname{sign}(\psi_2)$, b/c first eigval is constant. Can use $\operatorname{sign}(\psi_2)$ to indicate clustering

---

Definition (connected components)

On $G=(V,E)$ on each edge $(i,j)\in E$, $w_{ij}>0$. If node i is connected to 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

Prop (Eigenspace of $\lambda=0$ of $\mathcal{L}$ and Connected components in G)

Suppose the graph has k conn. comp $A_1, \dots, A_k$. Then the eigenspace of $\lambda=0$ of dim k is spanned by $\{\mathbb{1}_{A_1}, \dots, \mathbb{1}_{A_k}\}$ where

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

Pf

Suppose $\underline{f}$ is eigenvector of $\lambda=0$. So $\mathcal{L}\underline{f} = \underline{0}$ and $\underline{f}^T\mathcal{L}\underline{f} = 0$. We also know $\underline{f}^T L \underline{f} = \frac{1}{2}\sum_{(i,j)\in E} w_{ij}(f_i - f_j)^2$ and $\underline{f}^T\mathcal{L}\underline{f} = 0 \iff f_i = f_j$ whenever $w_{ij}>0$.

Thus, $\underline{f}$ is piecewise constant in each of conn. comp. Meanwhile, for each $\underline{v}\in \operatorname{Span}\{\mathbb{1}_{A_1}, \dots, \mathbb{1}_{A_k}\}$, $\underline{v}^T\mathcal{L}\underline{v} = 0$

Exercise:

- Does this generalize to L_{rw} and L_{sym}
- How leads to consistency of spectral clustering?