Lecture 14

8.2 k-Means clustering

Adapted from [2, Chapter 3.1.1]

In the previous section we motivated spectral graph theory by looking at some interesting embeddings of graphs. Hidden in that exposition were some more general properties of the eigenvalues and eigenvectors of the graph Laplacian that are related to clustering graphs. We now motivate the systematic development of these unstated ideas by first describing a popular clustering algorithm, which is *k-means*.

Suppose we have k "centers" $\mu_1, \dots, \mu_k \in \mathbb{R}^p$ and we want to cluster (or partition) all of $\mathbb{R}^p$ into k sets $S_1, \dots, S_k$ such that if $x \in S_i$, then x is closer to μ_i than it is to μ_j for any other $j \neq i$. Such a partition is called a *Voronoi diagram*. Suppose by closer we mean the standard Euclidean norm; the sets (or cells as they are often called) S_i in the Voronoi diagram are then defined as:

$$\forall\, i = 1, \dots, k, \quad S_i = \{x \in \mathbb{R}^p : \|x - \mu_i\| \leq \|x - \mu_j\|, \ \ \forall\, j \neq i\}.$$

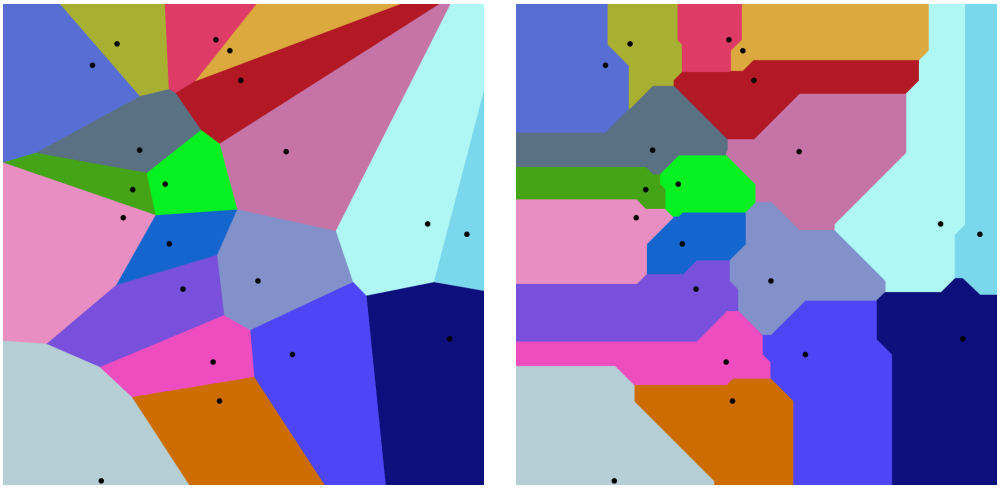
Figure 23(a) visualizes a particular Voronoi diagram. Notice that the Voronoi diagram depends upon our notion of distance. If we replace the Euclidean norm with the ℓ^1 norm,

$$\|x - x'\|_1 = \sum_{i=1}^{p} |x[i] - x'[i]|,$$

then the resulting Voronoi diagram is given in Figure 23(b). So different metrics give different clusters! This is perhaps obvious, but it again reminds us that our notion of similarity greatly affects how are algorithm will behave.

The k-means algorithm aims to find k clusters of a finite amount of data $\mathcal{X}_n = \{x_1, \dots, x_n\} \subset \mathbb{R}^p$ by using a Voronoi-like partition of space. It partitions the n data points into clusters $S_1, \dots, S_k \subset \{1, \dots, n\}$ with centers $\mu_1, \dots, \mu_k \in \mathbb{R}^p$ as the solution to:

$$\min_{\substack{\text{partition } S_1,\dots,S_k \\ \text{centers } \mu_1,\dots,\mu_k}} \sum_{i=1}^{k} \sum_{j \in S_i} \|x_j - \mu_i\|^2. \tag{45}$$



(a) Voronoi diagram with Euclidean distance

(b) Voronoi diagram with ℓ^1 distance

Figure 23: Voronoi diagrams with different metrics

Notice that, if the partition $S_1, \ldots, S_k$ is fixed, then the optimal centers are give by:

$$\mu_i = \frac{1}{|S_i|} \sum_{j \in S_i} x_j,$$

which is simply the mean of the set S_i, hence the notation μ_i.

The k-means algorithm is an iterative algorithm. It is initialized by choosing k random centers $\mu_1, \ldots, \mu_k$. It then alternates between:

1. Given centers $\mu_1, \ldots, \mu_k$, assign each points $x_i \in \mathcal{X}_n$ to the cluster:

$$j = \arg \min_{\ell=1,\ldots,k} \|x_i - \mu_\ell\|.$$

2. Update the centers: $\mu_i = \frac{1}{|S_i|} \sum_{j \in S_i} x_j$.

The algorithm continues until the centers and clusters stabilize. The algorithm is guaranteed to converge, but it is *not* guaranteed to find the solution to (45). Indeed, it often gets stuck in local minima, and in fact optimizing (45) is NP-hard. Other issues that can arise using k-means are:

- One needs to set k in advance, and because the algorithm can get stuck in local minima, it is sensitive to the initial random choice of $\mu_1, \ldots, \mu_k$ (see the slides).

- As it stands now, it is formulated in terms of Euclidean space $\mathbb{R}^p$, and so would need to be reformulated if all we have are some notions of similarity (like a metric or kernel). This can be done though.
- The solutions to k-means are always convex clusters, which means it will have trouble finding clusters that are not convex (imagine winding tubes).

All of these issues motivate the study of spectral clustering.

8.3 The graph Laplacian

Adapted from [10, Lecture 2]

We now give a more systematic treatment of the graph Laplacian and prove our first fundamental result relating the connectivity of a graph to the spectrum of its graph Laplacian.

8.3.1 Review of some basics of graph theory

We first give some definitions regrading graphs.

Recall that a graph $G = (V, E)$ consists of vertices and edges. Without loss of generality, we can take the vertices to be $V = \{1, \dots, n\}$. Edges are then denoted by $(i, j) \in E$, which is an unordered pair.

A graph $G = (V, E)$ is *connected* if for every $i, j \in V$ there exists a sequence $i = k_1, \dots, k_m = j$ such that $(k_\ell, k_{\ell+1}) \in E$ for all $\ell = 1, \dots, m$. The sequence $k_1, \dots, k_m$ is called a *walk* between i and j. If the vertices $k_1, \dots, k_m$ are distinct, then the sequence is called *path* and the length of the path is $m - 1$. If G is connected then every $i, j \in V$ are connected by a path.

Graphs G that are not connected are called *disconnected*. Each maximal connected piece of a graph is called a *connected component*. Connected graphs have one connected components, and disconnected graphs have two or more connected components.

The degree of the vertex i is the number of vertices $j \in V$ connected to i by an edge:

$$\deg(i) = \#\{j \in V : (i, j) \in E\}.$$

References

[1] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning. The MIT Press, 2002.

[2] Afonso S. Bandeira. Ten lectures and forty-two open problems in the mathematics of data science. MIT course *Topics in Mathematics of Data Science*, 2015.

[3] Jon Shlens. A tutorial on principal component analysis. arXiv:1404.1100, 2014.

[4] Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine, Series 6*, 2(11):559–572, 1901.

[5] V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb. (N.S.)*, 72(114):507–536, 1967.

[6] J. Baik, G. Ben-Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

[7] Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642, 2007.

[8] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer-Verlag New York, 2nd edition, 2009.

[9] Athanasios Tsanas and Angeliki Xifara. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, 49:560–567, 2012.

[10] Daniel A. Spielman. Spectral graph theory. *Yale Course Notes*, Fall, 2009.