8.1 Clustering

We will look at three techniques for data clustering:

1. k-means
2. hierarchical clustering
3. spectral clustering

8.1.1 k-means

We have studied this technique in the previous lecture. As a reminder, it is a descent on a Least Squares objective function.

8.1.2 Hierarchical Clustering

Given a set of n points in a metric space $S$, $x_1, x_2, \ldots, x_n \in S$, we follow an iterative algorithm to find the clusters. For the first iteration, we merge the two closest points into a single cluster. As a result, we are left with n-1 clusters. We need to iterate, but we need to know how to compare clusters.

$$d(C_1, C_2) = \max_{x \in C_1} d(x) \quad \text{where} \quad \forall x \in C_1, d(x) = \min_{y \in C_2} d(x, y)$$

Alternatively, we can define it as an average:

$$\sum_{x \in C_1} d(x) = \sum_{y \in C_2} d(x, y) \quad \text{where} \quad \sum_{x \in C_1} \sum_{y \in C_2} d(x, y)$$

At each level, we can associate a cost. For example, for some "goodness" function $g$, $\sum_{j=1}^{n} g(j)$. If we let $g$ be the average distance between two points in a cluster, we have:

$$\frac{1}{|C_j|} \sum_{x, y \in C_j} d(x, y)$$

But, what happens in some pathological cases? Imagine two rings of points, where a smaller ring is inside the larger ring. How would hierarchical clustering classify this set of points? Ultimately, clustering is a topological feature of the data set.
8.1.3 Spectral Clustering

Again, we are given a set of \( n \) points, \( x_1 \ldots x_n \).

We have a symmetric matrix \( W \) defined as \( W_{ij} = \) "association" or "similarity" between \( x_i \) and \( x_j \).

We want to make a graph by connecting close points. Choose some \( \epsilon > 0 \), and connect all points within \( \epsilon \) of each other.

For a geometric random graph \( G(n, \epsilon) \), randomly sample \( n \) points and connect all points within \( \epsilon \) of each other.

Define \( W_{ij} = 1 \) if \( ||x_i - x_j|| < \epsilon \).

Consider cutting the graph in two. That is, we want to find a map \( b : V \to \{ -1, 1 \} \) to find \( S = b^{-1}(1) \) and \( \bar{S} = b^{-1}(-1) \)

But, we want very few links between the two clusters. That is we want:

\[
\min_{b} \sum_{i \in S, j \in \bar{S}} W_{ij} = 0 \quad \text{balanced cuts}
\]

\[
\min_{b} \sum_{i \in S, j \in \bar{S}} W_{ij} = \frac{1}{2} \sum_{i,j=1}^{n} W_{ij} (b_i - b_j)^2 = b^T L b
\]

where \( L = D - W \). \( L \) is the Laplacian of the graph and \( D \) is a diagonal matrix. \( D_{ii} = \sum_{j} W_{ij} \)

To prove this last equality,

\[
\sum_{i,j=1}^{n} W_{ij} (b_i - b_j)^2 = \sum_{i,j=1}^{n} W_{ij} (b_i^2 + b_j^2 - 2b_i b_j) = \\
\sum_{i,j=1}^{n} W_{ij} b_i^2 + \sum_{i,j=1}^{n} W_{ij} b_j^2 - 2 \sum_{i,j=1}^{n} W_{ij} b_i b_j = \\
\sum_i b_i^2 \sum_j W_{ij} + \sum_j b_j^2 \sum_i W_{ij} - 2b^T W b = \\
\sum_i b_i^2 D(i, i) + \sum_j b_j^2 D(j, j) - 2b^T W b =
\]

The first two terms are equal since \( W \) is symmetric. Thus, we have

\[
2b^T Db - 2b^T W b = 2b^T (D - W) b = 2b^T L b
\]

Finding the min cut is the same as minimizing the Laplacian.

The Laplacian is symmetric (both \( D \) and \( W \) are symmetric).

\( L \) is positive semi-definite : \( b^T L b = \sum W_{ij} (b_i - b_j)^2 \geq 0 \)

\( L \) has real eigen values \( \lambda_1 \leq \lambda_2 \ldots \lambda_n \) with associate eigen vectors \( v_1 \ldots v_n \)

Notice that the smallest eigen value \( \lambda_1 = 0 \) and \( v_1 = 1 \)

\( (D - W)1 = D1 - W1 = 0 \cdot 1 \)

\( v_2 \perp v_1 \)

Claim: \( \min_{v \perp v_1} v^T L v = \lambda_2 \)

\[
v^T v_1 = \alpha_1 = 0 \\
L v = \sum_{i=2}^{n} \alpha_i L v_i = \\
\sum_{i=2}^{n} \alpha_i^2 \lambda_i = v^T L v = (\sum \alpha_i v_i)^T (\sum \alpha_i \lambda_i v_i)
\]

Note: the multiplicity of \( \lambda = 0 \) is the number of connected components.

The difficult question is what is the value of \( k \) (number of clusters)?

1. \( p = \sum_{i=1}^{k} \alpha_i N(\mu_i, \epsilon) \)

mixture of Gaussians.
2. P has support on the manifold $M = \bigcup_{i=1}^{k} M_i$
   if 2 connected components, you can discover the number of components if you sample enough

Suppose you have a manifold $M \in \mathbb{R}^n$. If we compare with a graphs (discrete):

Graphs: $G(V,E)$

1. $f : V \rightarrow R$
2. $Lf = g$
3. $f^T Lf = \Sigma W_{ij} (f_i - f_j)^2$ (Stoke’s theorem on graphs)
4. Random walk on graph

Manifolds:

1. $f : M \rightarrow R$
2. $\Delta f = g$
   
   $f : \mathbb{R}^n \rightarrow R$
   
   $\Delta f = \Sigma_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2}$

3. $\int_M f \Delta f = \int_M < \nabla f, \nabla f > = \int_M ||\nabla f||^2$ (Stoke’s theorem on graphs)
4. Brownian motion, or heat flow