Clustering and Metric Space Magnitude

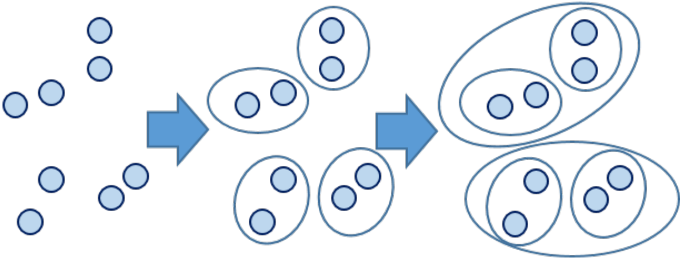
Leo Selker
Pomona College

September 24, 2016

Outline

Hierarchical Clustering

- Points represent data in Euclidean space.
- Idea: Capture structure at various scales



Plan

- There are relatively intuitive ways of clustering: We could connect points within k of each other and then scale k.
- But that's boring!

Definition

- A **metric space** (X, d) is a set X along with a function $d : X \times X \to \mathbb{R}^{\geq 0}$, such that, $\forall x, y, z \in X$:
  - $d(x, y) \geq 0$, with equality iff $x = y$;
  - $d(x, y) = d(y, x)$; and
  - $d(x, y) + d(y, x) \geq d(x, z)$.

  If X is finite then we say (X, d) is a **finite metric space**.
- Useful example: Finite set of points in $\mathbb{R}^2$.

Data as a Metric Space

- Data consists of a finite set of samples over n variables
- Very common idea: Represent data as points in $\mathbb{R}^n$.
- Note: $\mathbb{R}^n$ has far more structure than we need. We only care about distance.

Intuition behind Weighting

- We want to count the number of clusters in a data set. So we'll assign a **weighting** to the points.
- We want each cluster's weight to sum to close to 1
- Points near many other points will have smaller values
- Points which are separated will have larger values
- Then, we'll sum the weights to get the **magnitude**, i.e. number of clusters, in the data set.

Weighting and Magnitude

- A **weighting** on a finite metric space X is a set of weights ω_x in $\mathbb{R}$ such that, for all $x \in X$:

$$\sum_{y \in X} e^{-d(x,y)} \omega_y = 1$$

  Weights may be negative, but it's easier to think of them as nonnegative.

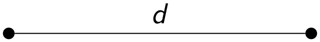
- Then the **magnitude** of X, denoted $|X|$, is defined by:

$$|X| = \sum_{x \in X} \omega_x.$$

- May be various weightings, but magnitude is unique
- Scale-dependent - see examples to come

Two Points

Consider the space of two points separated by distance d:



- If d is small, we expect there to be one cluster, and if d is large, 2 clusters.
- In this case, the weight of each point is equal to

$$\frac{1}{1+e^{-d}}$$

so the magnitude is

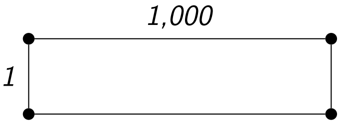
$$\frac{2}{1+e^{-d}}.$$

Two Points Plot

The space's magnitude as we scale d:

Simple Examples

Consider the space below:



- If d is small, we expect there to be one cluster, and if d is large, 2 clusters.
- In this case, the weight of each point is equal to

$$\frac{1}{1+e^{-d}}.$$

Thank you!