Clustering

Hierarchical Agglomerative Clustering (HAC)

Learning Objectives

- What's the **Hierarchical Agglomerative Clustering (HAC) Algorithm**?
  - What's hierarchical clustering? What's agglomerative? How does it work?
  - How does it compare to *k*-Means?
- What is a measure of **dissimilarity between points**? How about **dissimilarity between clusters**? What's **linkage**?
- How does **(high) dimensionality of data** impacts **metrics-based clustering techniques** such as HAC and *k*-Means?
- Afternoon Assignment
  - Topic modeling with *k*-Means
  - Implement HAC with *scipy*
  - Topic modeling with HAC

Overview

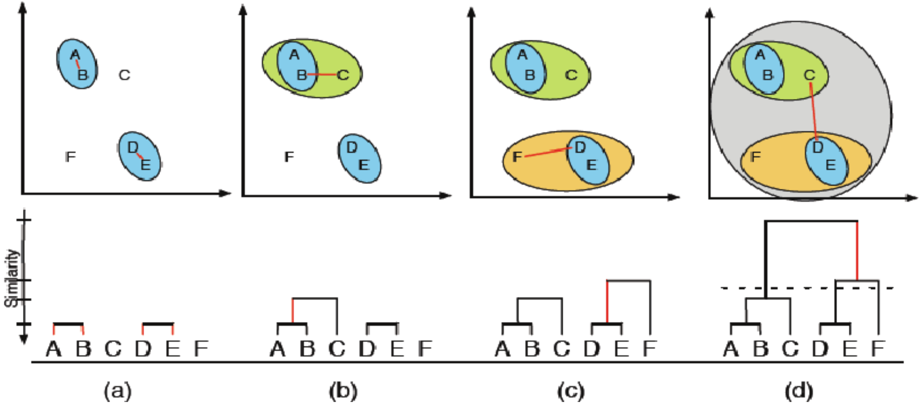
Overview

Hierarchical Clustering

- Type of **agglomerative** clustering
  - I.e., we will **iteratively group** observations together based on their **distance** from one another
  - As we continue to group observations together we form a hierarchy of their **similarity** to one another
- This will force us to answer different questions then we did in k-Means
  - No longer do we have to choose the number of clusters up front
  - Instead we'll have to define the **nature of successive grouping** of observations

Overview

Example of Hierarchical Agglomerative Clustering



Overview

HAC

The algorithm in all its glory:

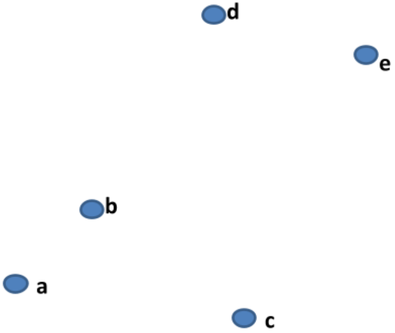
1. Each point as its own cluster
2. Merge "closest" clusters
3. End when all data points are in a single cluster

Like k-Means, this training algorithm may look pretty simple...
and that's because it is

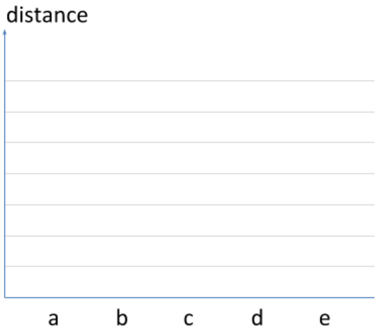
Overview

Step-by-step Execution: DATA!!

1 - Compute distances between observations
2 – Identify/choose a minimum
3 – Fuse observations

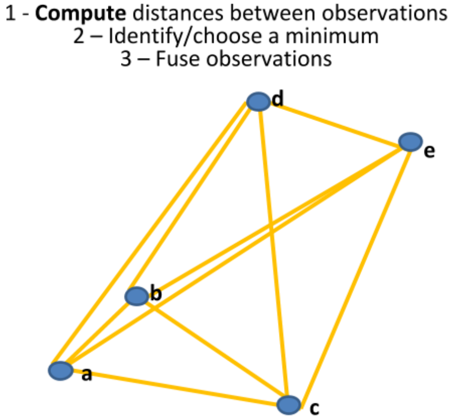


Observations

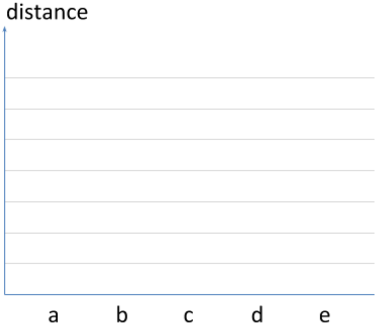


Dendrogram

Step-by-step Execution: Iteration 1 - Compute



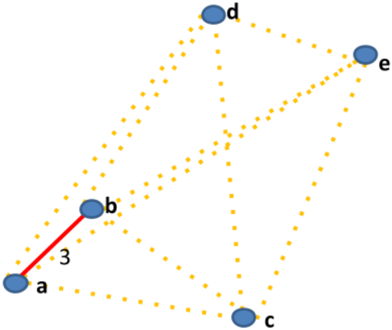
Observations



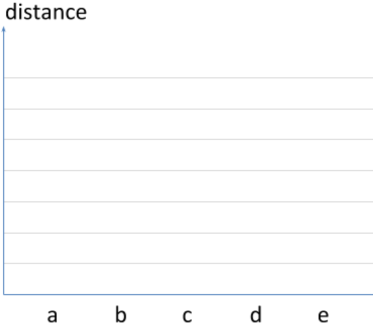
Dendrogram

Step-by-step Execution: Iteration 1 - Identify

1 - Compute distances between observations
2 – **Identify**/choose a minimum
3 – Fuse observations

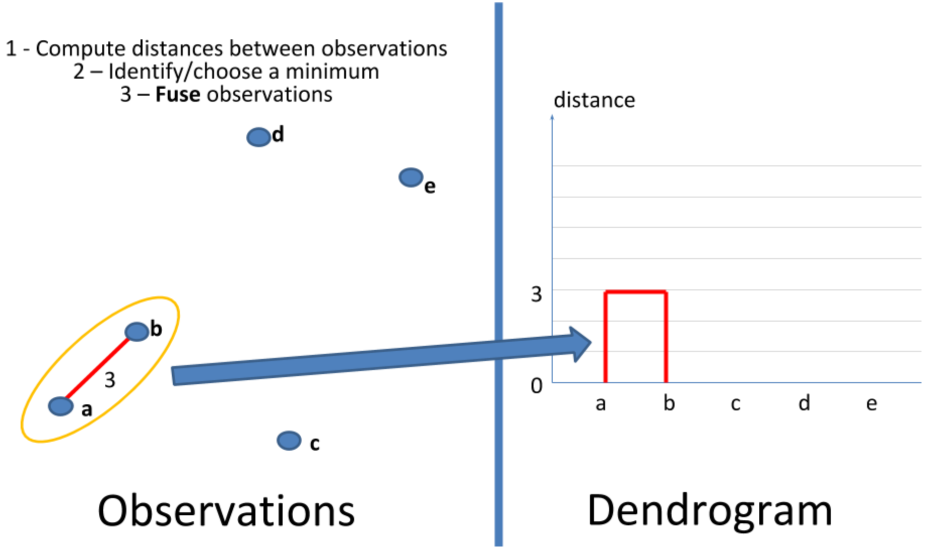


Observations



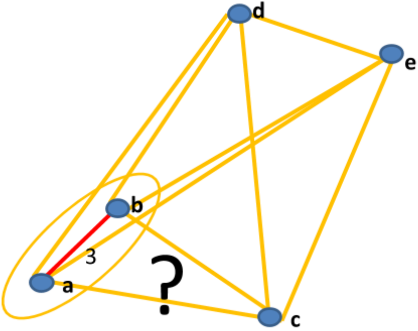
Dendrogram

Step-by-step Execution: Iteration 1 - Fuse

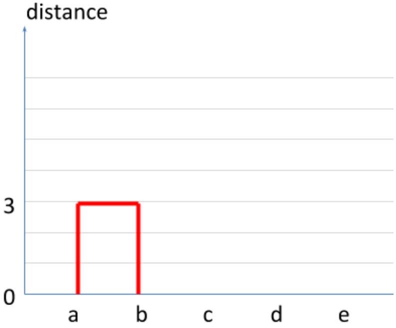


Step-by-step Execution: Iteration 2 - Compute

1 - **Compute** distances between observations
2 – Identify/choose a minimum
3 – Fuse observations



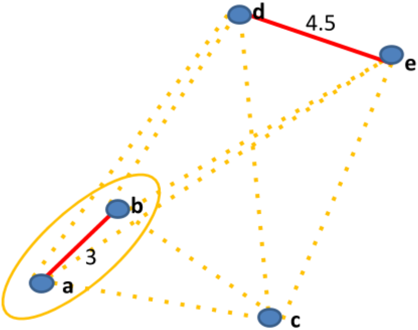
Observations



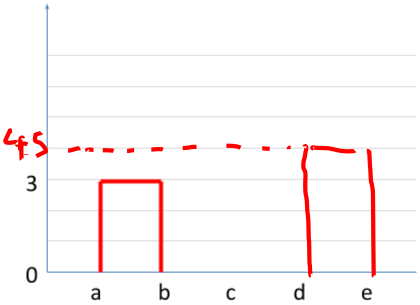
Dendrogram

Step-by-step Execution: Iteration 2 - Identify

1 - Compute distances between observations
2 – **Identify**/choose a minimum
3 – Fuse observations

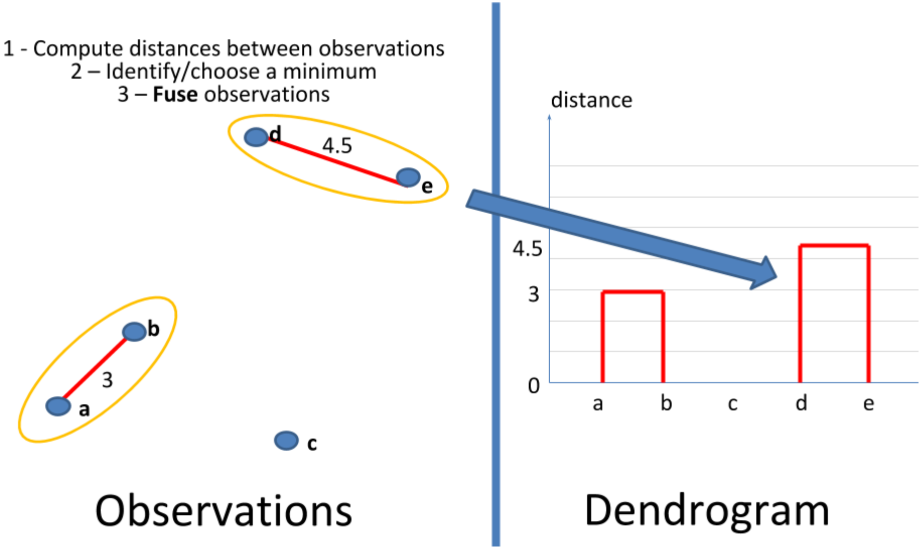


Observations

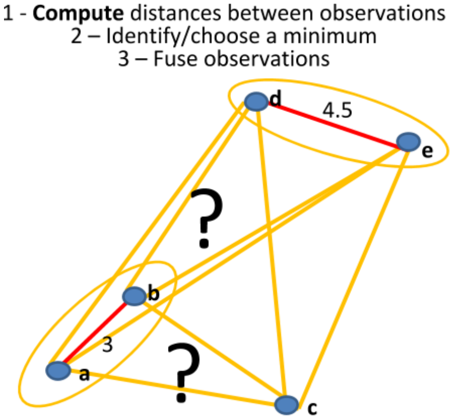


Dendrogram

Step-by-step Execution: Iteration 2 - Fuse



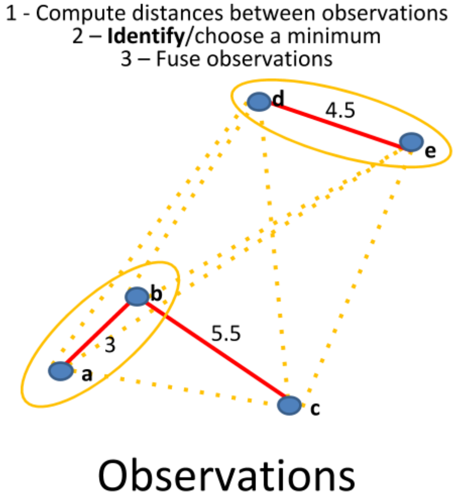
Step-by-step Execution: Iteration 3 - Compute

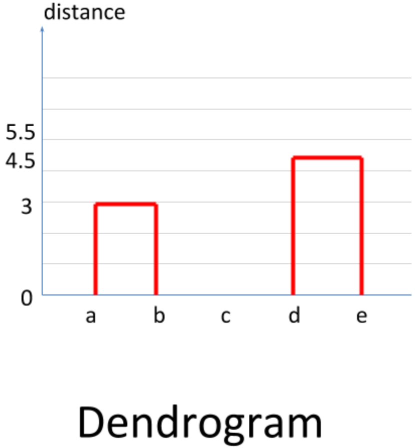


Observations

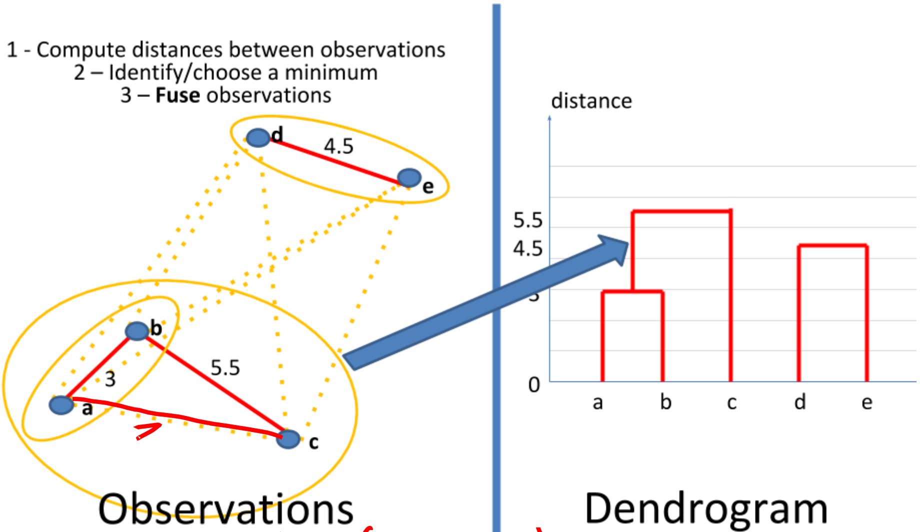
Dendrogram

Step-by-step Execution: Iteration 3 - Identify





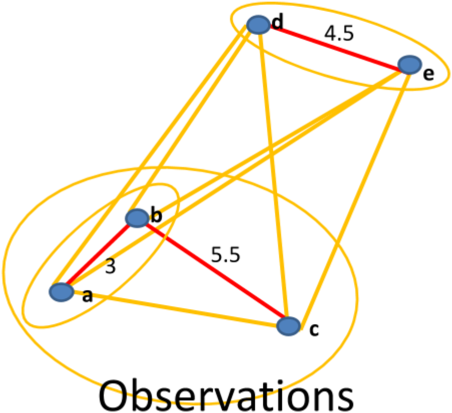
Step-by-step Execution: Iteration 3 - Fuse



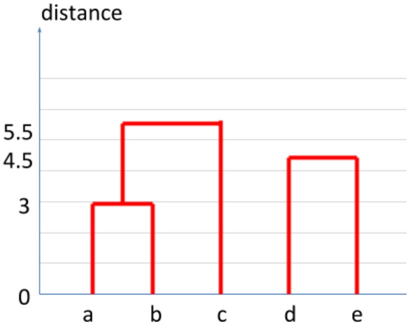
$$D(ab,c) = \frac{1}{2\times 1}\left(\underset{5.5}{D(a,c)} + \underset{7}{D(b,c)}\right) = \dots$$

Step-by-step Execution: Iteration 4 - Compute

1 - **Compute** distances between observations
2 – Identify/choose a minimum
3 – Fuse observations



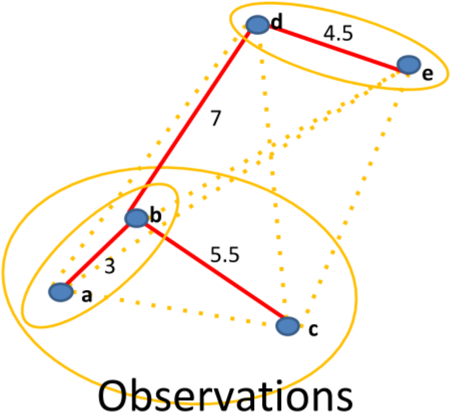
Observations



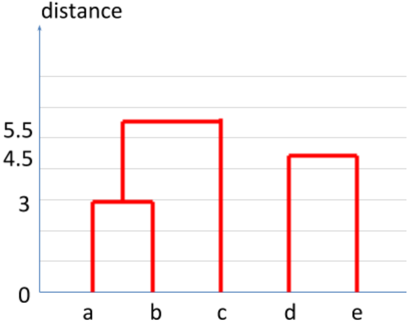
Dendrogram

Step-by-step Execution: Iteration 4 - Identify

1 - Compute distances between observations
2 – **Identify**/choose a minimum
3 – Fuse observations

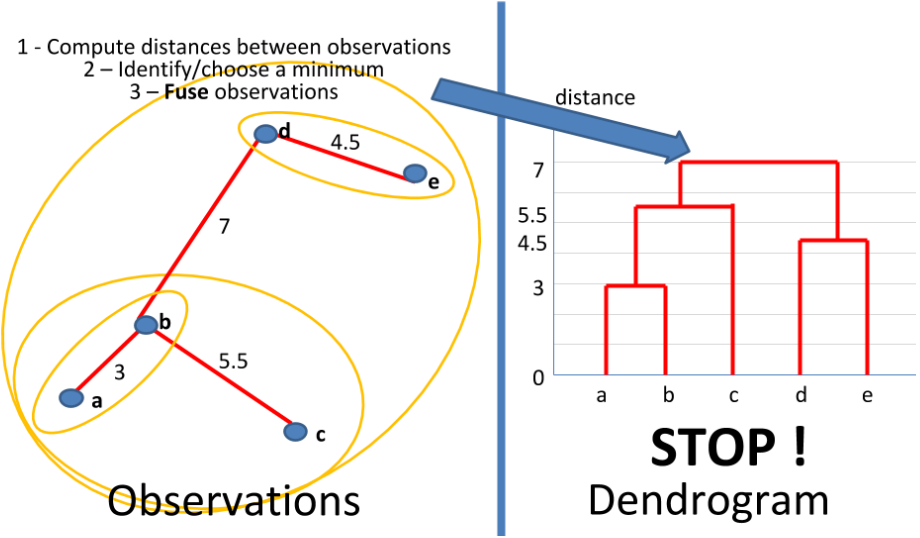


Observations

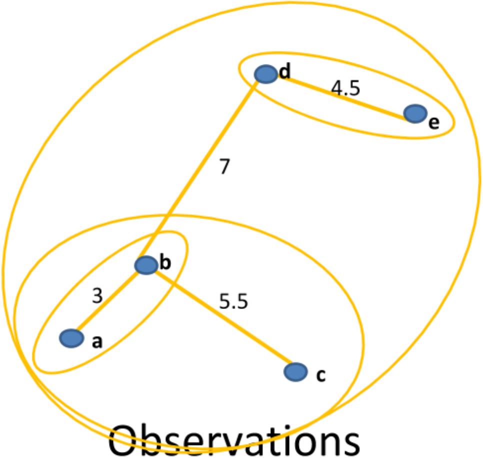


Dendrogram

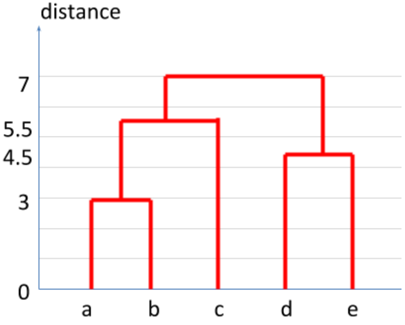
Step-by-step Execution: Iteration 4 - Fuse



HAC: Final Dendrogram



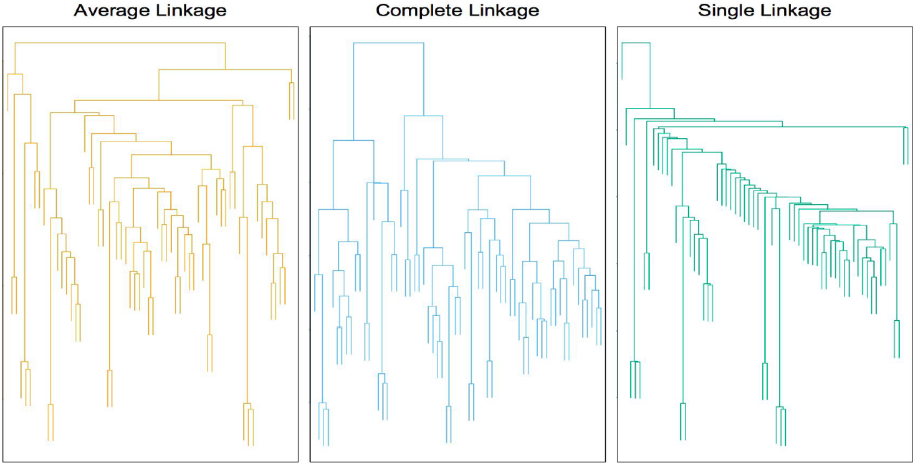
Observations



Dendrogram

Overview

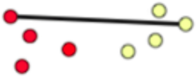
Distance betweeen clusters?



Complete Linkage

Maximal intercluster dissimilarity

- Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the **largest** of these dissimilarities



$$D(A, B) = \max_{a \in A, b \in B} D(a, b)$$

- Pro: Balanced clusters
- Cons: More sensitive to outliers; may violate "closeness"
  - Forces "spherical" clusters with consistent "diameter"

$\rightarrow$ More commonly used

Single Linkage

Minimal intercluster dissimilarity

- Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the smallest of these dissimilarities

$$D(A, B) = \min_{a \in A, b \in B} D(a, b)$$

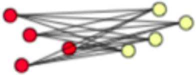
- Pro: Less senstive to outliers; handles irregular shapes fairly naturally
- Con: Extended, trailing clusters
  - Can fuse single observations one-at-a-time, producing long chains $a \rightarrow b \rightarrow ... \rightarrow z$

$\rightarrow$ Less commonly used

Average Linkage

Mean intercluster dissimilarity

- Compute all pairwise dissimilarities between the observations in cluster A and the the observations in cluster B, and record the **average** of these dissimilarities



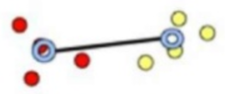
$$D(A, B) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} D(a, b)$$

- Compromise between complate and single linkages
- Pro: Balanced clusters
  - Less affected by outliers

$\rightarrow$ More commonly used

Centroid Linkage

Dissimilarity between the centroid for cluster A and the centroid for cluster B



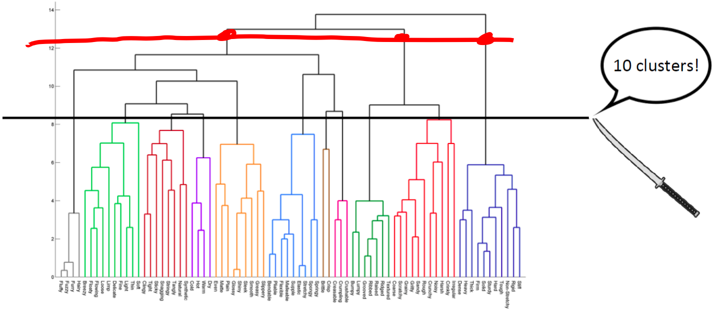
$$D(A,B) = D\Big(\frac{1}{|A|}\sum_{a\in A}\vec{a}, \frac{1}{|B|}\sum_{b\in B}\vec{b}\Big)$$

- Centroid linkage can result in undesirable inversions

$\rightarrow$ Not as commonly used, though popular in Genomics

Overview

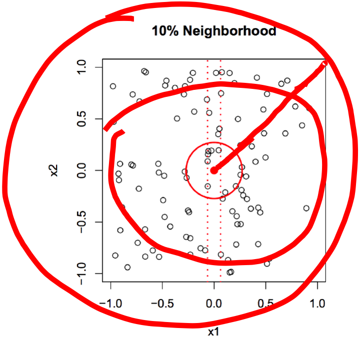
Choosing k

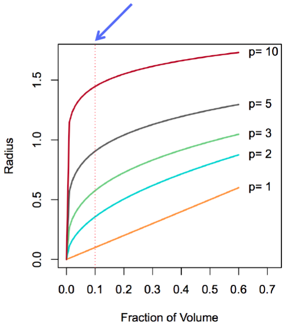


- In constrast to k-Means, we don't have to choose k from the start
- Depending on where precisely we cut, we have anywhere from 1 to n clusters

The curse of dimensionality

Just as nearest neighbors breaks down in high dimensional space...
Distance based clustering breaks down in high dimensional space...





The curse of dimensionality

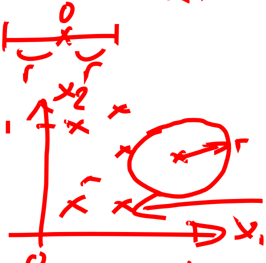
1D

$\rho = 10\%$

$2r$

$$\frac{2r}{1} = \rho \qquad r = \frac{\rho}{2} = 0.05$$

2D



point in dataset

dataset

$$\frac{\pi r^2}{1} = \rho \qquad r = \sqrt{\frac{\rho}{\pi}}$$

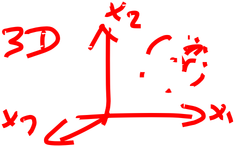
$$= \sqrt{\frac{0.1}{3.14}}$$

$$= 0.178$$

$k = 5$

$n = 50$ → 10% density

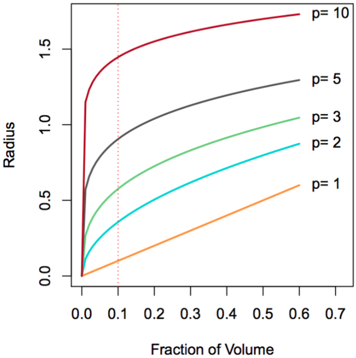
3D



$$\frac{\frac{4}{3}\pi r^3}{1} = \rho \qquad r = \left(\frac{\rho}{\frac{4}{3}\pi}\right)^{1/3} = .287$$

The curse of dimensionality

Can you work out some of the points on the plot?



Dimension	Volume of a ball of radius R	Radius of a ball of volume V
0	1	All balls have volume 1
1	$2R$	$V/2$
2	πR^2	$\frac{V^{1/2}}{\sqrt{\pi}}$
3	$\frac{4}{3}\pi R^3$	$\left(\frac{3V}{4\pi}\right)^{1/3}$
4	$\frac{\pi^2}{2}R^4$	$\frac{(2V)^{1/4}}{\sqrt{\pi}}$
5	$\frac{8\pi^2}{15}R^5$	$\left(\frac{15V}{8\pi^2}\right)^{1/5}$
6	$\frac{\pi^3}{6}R^6$	$\frac{(6V)^{1/6}}{\sqrt{\pi}}$
7	$\frac{16\pi^3}{105}R^7$	$\left(\frac{105V}{16\pi^3}\right)^{1/7}$
8	$\frac{\pi^4}{24}R^8$	$\frac{(24V)^{1/8}}{\sqrt{\pi}}$
9	$\frac{32\pi^4}{945}R^9$	$\left(\frac{945V}{32\pi^4}\right)^{1/9}$
10	$\frac{\pi^5}{120}R^{10}$	$\frac{(120V)^{1/10}}{\sqrt{\pi}}$