Today: High-D Continuous Data, Clustering

Sam Ventura

36-315

Today: Distance Matrices,
Hierarchical Clustering, Dendrograms

Department of Statistics
Carnegie Mellon University

October 24, 2016

Distance = Metric = Distance Metric = Distance Function

Function that defines distance between pairs of observations in a dataset

Properties:

Examples:

Distance Matrices

A **distance matrix** is a data structure that efficiently organizes the pairwise distances between all observations in a dataset.

Pairwise distances are organized into the lower-triangle of a matrix, D

The $(i, j)^{th}$ element of the matrix contains the distance between x_i and x_j:

$D[i, j] = d(x_i, x_j)$

Examples:

Dendrograms / Visualizing High Dimensional Structure

There is no easy way to visualize how far apart observations are in high-dimensional space. One option we do have: **Dendrograms**