A Quick Introduction to Machine Learning (Hierarchical Clustering)

Lecturer: John Guttag

Clustering an Optimization Problem

An objective function and a constraint

Like many optimization problems, computationally nasty

Usually rely on a greedy approximation

- Hierarchical
- K-means

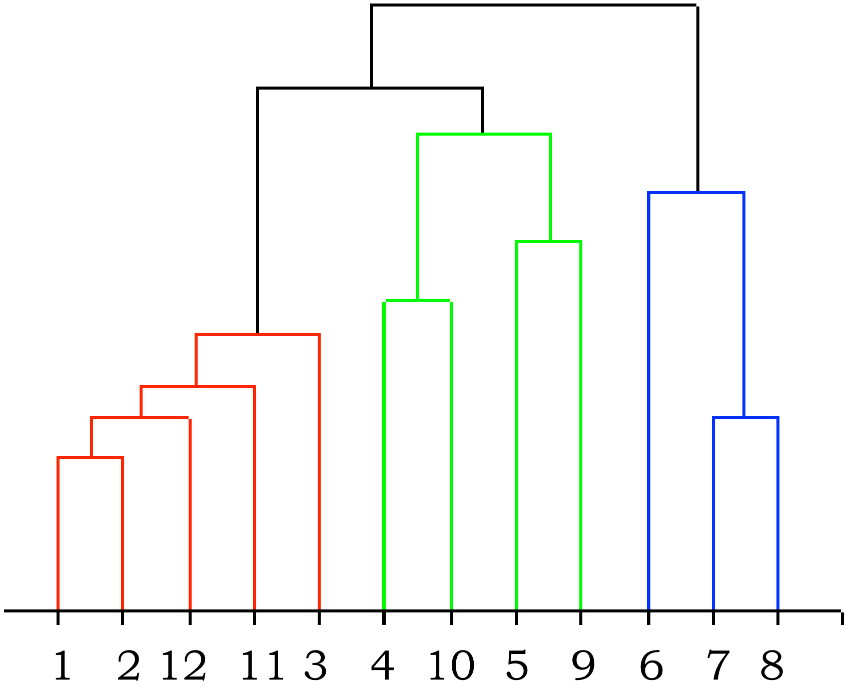
Hierarchical Clustering

1. Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item.

2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster fewer.

3. Continue the process until all items are clustered into a single cluster of size N.

Dendogram (Monthly Temperatures)



Linkage Criteria

In *single-linkage* clustering (also called the *connectedness* or *minimum* method), we consider the distance between one cluster and another cluster to be equal to the shortest distance from any member of one cluster to any member of the other cluster.

Linkage Criteria, continued

In *complete-linkage* clustering (also called the *diameter* or *maximum* method), we consider the distance between one cluster and another cluster to be equal to the greatest distance from any member of one cluster to any member of the other cluster.

Linkage Criteria, continued

In *average-linkage* clustering, we consider the distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any member of the other cluster. A slight variant of this uses the median instead of the mean.

Linkage Criterion

	BOS	**NY**	**CHI**	**DEN**	**SF**	**SEA**
BOS	0	206	963	1949	3095	2979
NY		0	802	1771	2934	2815
CHI			0	966	2142	2013
DEN				0	1235	1307
SF					0	808
SEA						0

BOS NY CHI DEN SF SEA

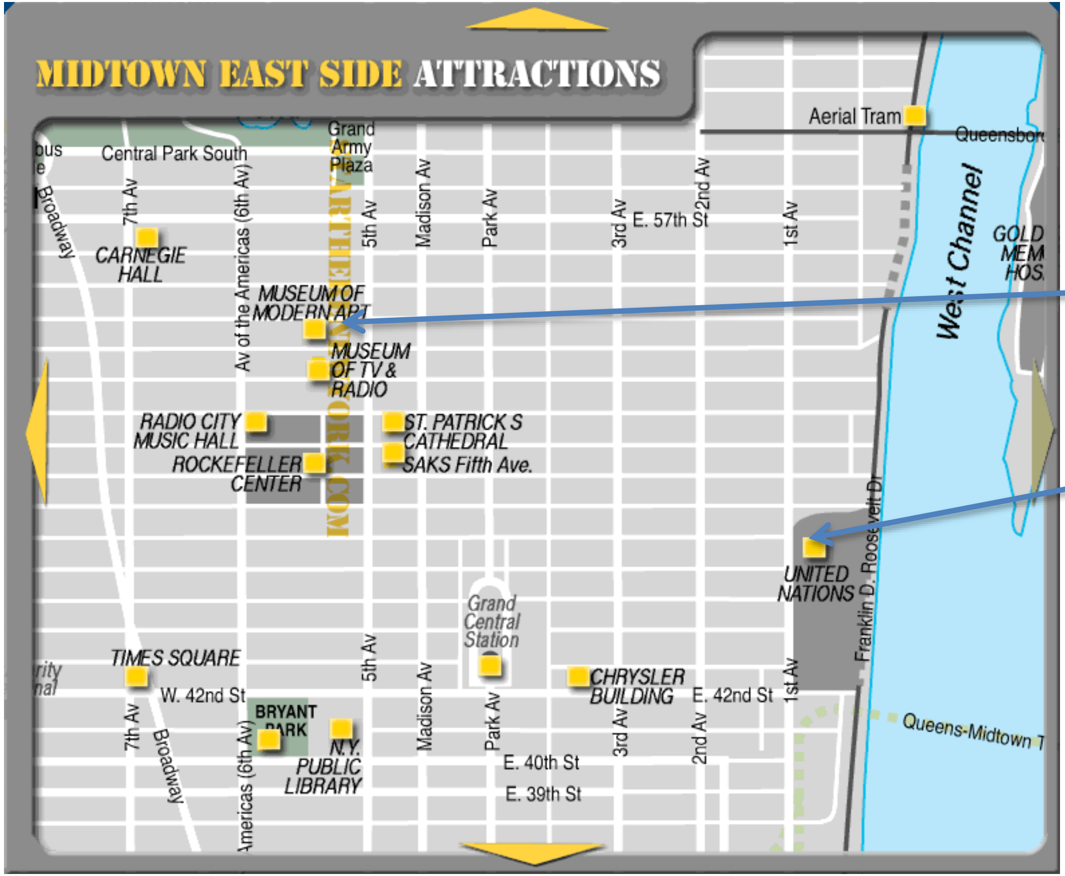
{BOS, NY} CHI DEN SF SEA

{BOS, NY, CHI} DEN SF SEA

{BOS, NY, CHI} DEN {SF, SEA}

{BOS, NY, CHI, DEN} {SF, SEA, *DEN*}

{BOS, NY, CHI, DEN, SF, SEA}



55th and 5th

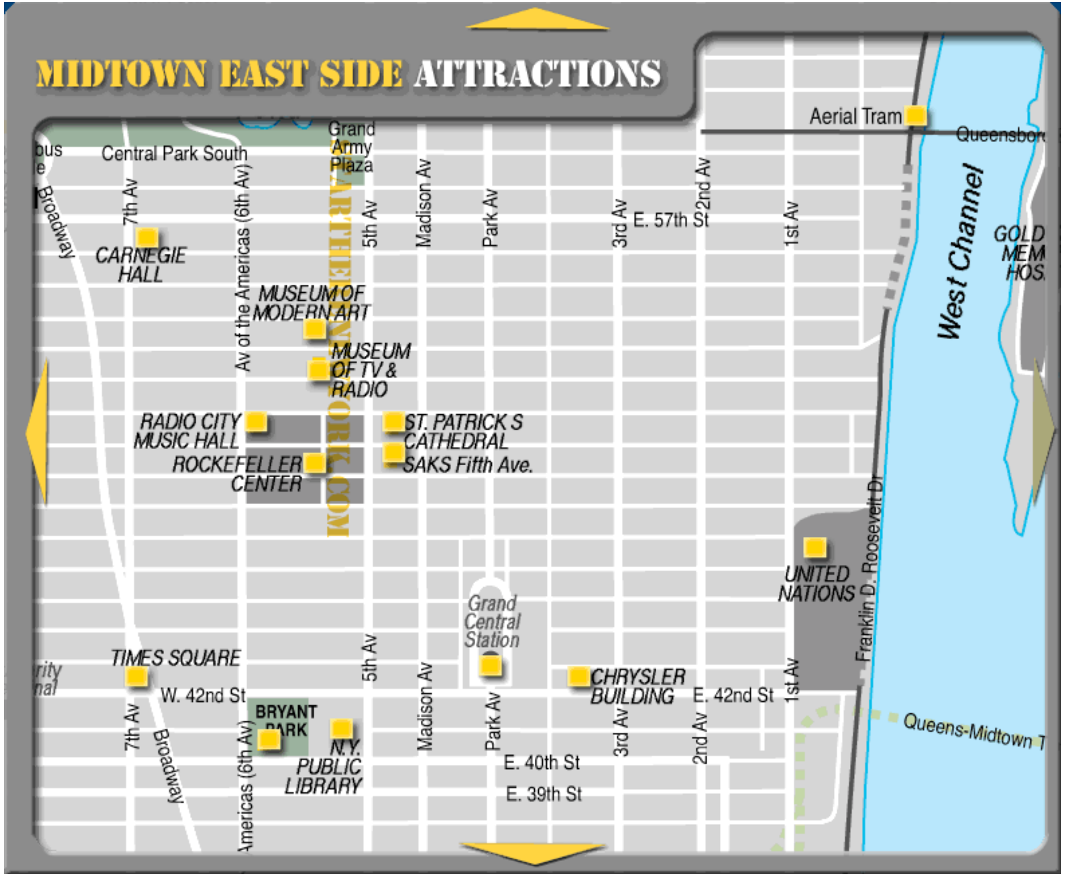
46th and 1st

Distance(<55,5>, <46,1>)

Minkowski Metric

$$dist(X1, X2, p) = \left(\sum_{k=1}^{len} abs(X1_k - X2_k)^p\right)^{1/p}$$

p = 1: Manhattan Distance
P = 2: Euclidean Distance


MIDTOWN EAST SIDE ATTRACTIONS
Central Park South
Grand Army Plaza
Broadway
7th Av
Av of the Americas (6th Av)
5th Av
Madison Av
Park Av
3rd Av
2nd Av
1st Av
E. 57th St
Aerial Tram
Queensboro
West Channel
CARNEGIE HALL
MUSEUM OF MODERN ART
MUSEUM OF TV & RADIO
RADIO CITY MUSIC HALL
ROCKEFELLER CENTER
ST. PATRICK S CATHEDRAL
SAKS Fifth Ave.
UNITED NATIONS
Franklin D. Roosevelt Dr
Grand Central Station
TIMES SQUARE
W. 42nd St
E. 42nd St
CHRYSLER BUILDING
BRYANT PARK
N.Y. PUBLIC LIBRARY
E. 40th St
E. 39th St
Queens-Midtown T
ARTHEEYORK.COM