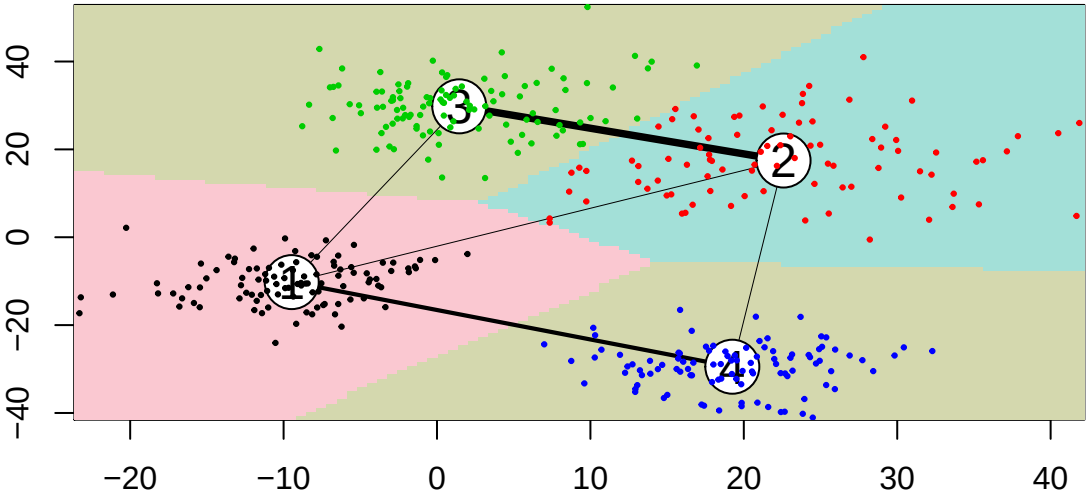
Lecture 9: Unsupervised Learning

Intro to Data Science for Public Policy, Spring 2016

by Jeff Chen & Dan Hammer, Georgetown University McCourt School of Public Policy

Contents

Not all data contain labels, but it does not mean the data do not have patterns. So as long as data is structured, some patterns – weak or strong – are always possible. Unlabeled data holds the potential to be a game cahnger of informing project and policy pursuits. For example, bank may be operating at steady state for decades without considering how segments of its user base may have divergent needs. But these user groups are often times not clearly indicated. An analyst could go through the data and use her intuition to manually identify *clusters* of personas (e.g. incomes over 50k, under 25 years of age, female). But this often times may be challenging as the number of features may be too voluminous to manually analyze and the choice of features may be somewhat arbitrary.



Unsupervised learning can help. It is a branch of machine learning that deals with unlabeled data to identify statistically-occurring patterns. Unsupervised learning can be described by *clustering*, which involves finding which observations or features tend to group together. In sales and recruitment, the task of customer segmentation may dependent on customer data to find distinct customer profiles. In some law firms, data scientists may develop topic modeling algorithms to automatically tag and cluster hundreds of thousands of documents for improved search.

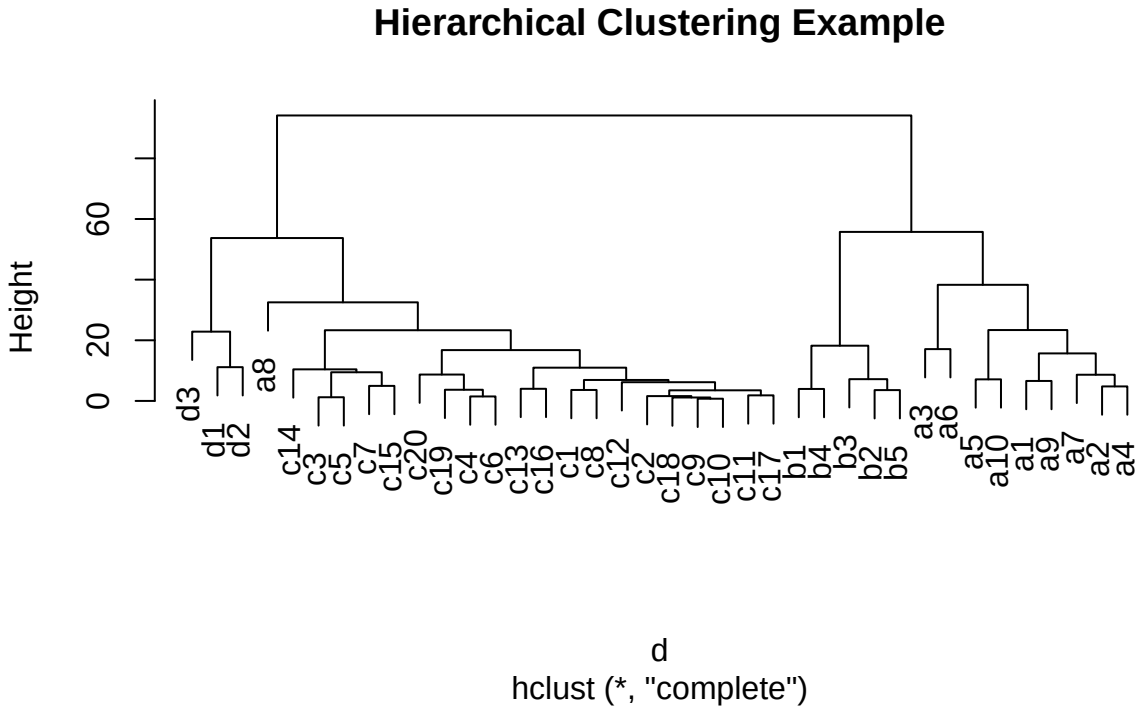
This chapter provides a short survey of types of unsupervised learning and its uses.

Section 1 - An Overview

Whereas classifiers rely on both a labeled target and input features, unsupervised learning relies on unlabeled input features to find regularities in the data. Using various types of optimization techniques, unsupervised

learning includes a wide variety of tasks, including clustering and dimensionality reduction.

Clustering looks for cases where groups of observations have similar values in the feature space. Two commonly used approaches are connectivity-based clustering and centroid models. Connectivity-based approaches rely on point-wise comparisons to find which points are closest, then agglomerating points into large hierarchical groups. An example of connectivity-based learning is *Hierarchical Clustering*, which can be illustrated using *dendrograms*.



- Clustering of records focuses on finding groups of records that are statistically similar
- Dimensionality reduction focuses on finding groups of variables that

Section 2 - Methods

□

K-Means

Technique for finding natural groups

Uses

- customer or persona segmentation, often used for strategic plans and operating plans
- identifying hotspots
- finding natural breaks in data for colour coding

```
Let k > 1, R is the feature space.
Normalize all input features into same range.
Drop or impute all missing values NA
Select a value of k
Initialize by randomly select k centroids in feature space R
  Calculate distance from each record to each of the k centroids
```

```
While each point's assignment changes:
  Assignment Step: Assign each point to the nearest centroid
  Update Step: Re-calculate centroid for each group of points
Stop when assignments no longer change
```

Considerations

- *Stability of Clusters.* A common error is to assume that the groups are stable
- *Missing Values.* K-Means do not handle missing values well as each data point is essentially a coordinate. Thus, often times K-Means are calibrated on complete data sets.
- *Normalization.* All features need to be standardized. Binary or discrete features do not perform well.
- *Stability.* Stability of clusters must be tested

K-Means and World Development Indiactors

Agglomerative Clustering

asd

Principal Components

asd