Big Data and Security

Jeffrey Borowitz, PhD

Lecturer

Sam Nunn School of International Affairs

Clustering

Clustering

- Sometimes, we are interested in finding groups of similar data points
  - This is particularly useful in things like disease diagnosis
  - But also useful in fraud detection
- Clustering is a way to do this
  - Find groups of data that seem “similar”

Clustering: The Process

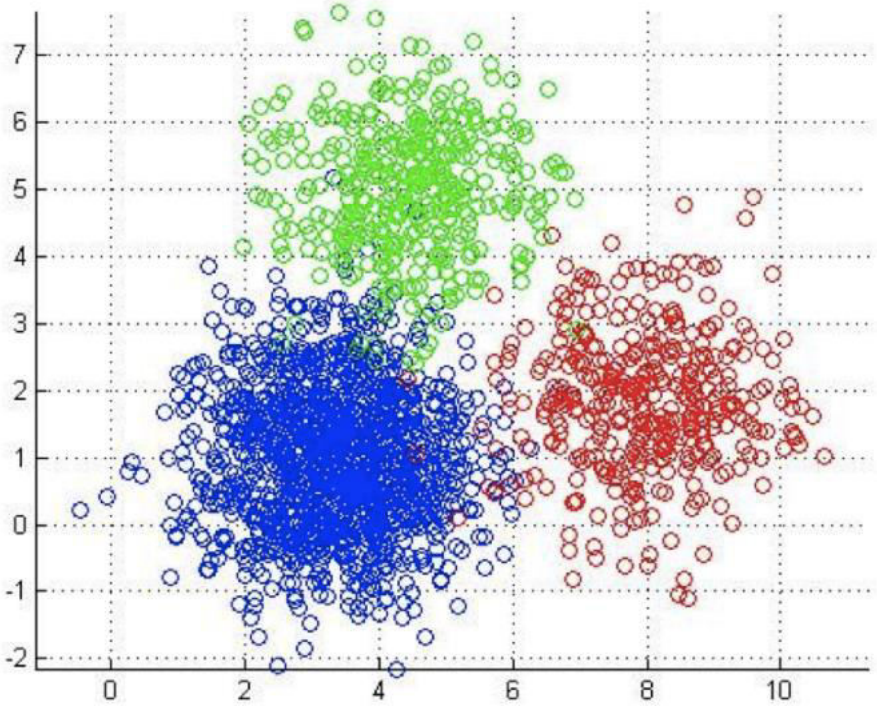
First, you define “similarity”

- A typical definition is “Euclidean” distance
- But many other measures are possible:
  - “Manhattan” distance based on “blocks”

Algorithm:

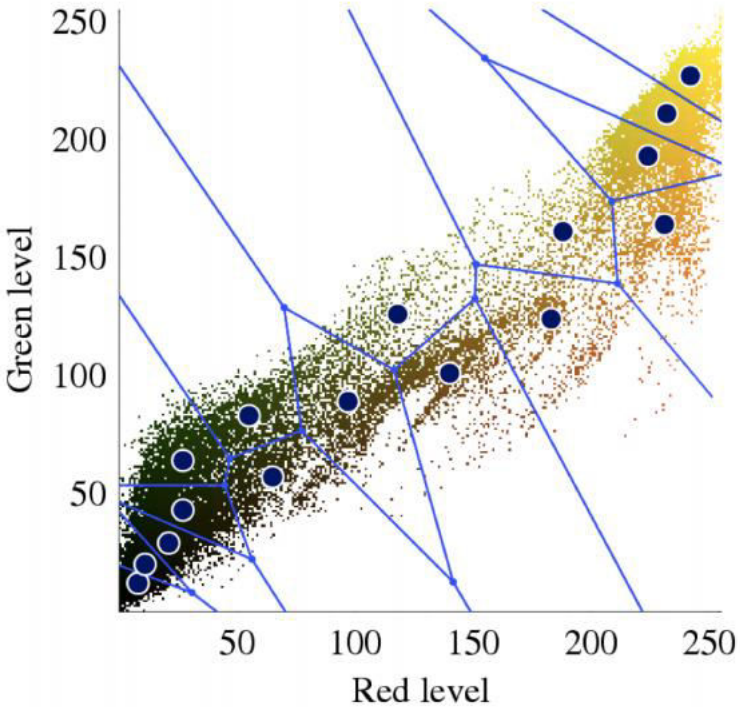
1. Start with a set of k different cluster centers.
2. Assign each element to the nearest cluster center, using your similarity measure
3. Move each cluster center to the average of points that are in cluster
4. Repeat this until cluster centers don’t move anymore.

Clustering Picture



Georgia Tech

Clustering with Centroids Drawn



Georgia Tech

Problems with Clustering

- You get a sense that this might not always converge to the same thing, depending on where your clusters start
- In general, computing the “best” set of *k* clusters is a very hard problem
- The algorithm above is approximate
  - For practical purposes, there are only approximate algorithms for this problem
- The curse of dimensionality
  - In more dimensions, there is more noise in the inputs to clustering - differences along various dimensions
  - So the difference between “close” and “far” elements is harder to pick up.
  - The normal way to deal with this is to reduce dimensions, by e.g. principal components

Dimension Reduction Deals with Curse of Dimensionality

- Sometimes there are too many dimensions of data to estimate
- You want to do a series or sieve based estimator
- You want to do something on text
  - Is the presence of a particular word a variable?
  - What about the number of each word?
  - Or combinations of words?

Georgia Tech

Principal Component Analysis Deals with Curse

- We have lots of dimensions
- Many are likely correlated
  - Each of the components of X might or might not be correlated with each other component
  - The case of predicting student outcomes:
    - Family income, parent education, income in neighborhood, government policies all matter
    - But many of these are correlated with each other: highly educated parents have high income, live in good neighborhoods
- Let's define an unobserved factor of "socioeconomic background", as whatever variable we can use to explain the most of these characteristics

Georgia Tech

PCA: Details

- Each principal component is a set of weightings on each variable
  - You can redefine your variables as your old set of variables times each weight
  - If you take all the principal components, there are the same number as there were variables
- In general you get a bunch of components, ordered by importance.
  - You never know what each component means, but sometimes you can guess from signs: SES would have positive weights
  - The first component explains the most variation, the second the second most, etc.
- PCA is nonparametric in the sense that it allows you to look at lots of data and take the “most relevant” part without knowing what that part is beforehand

Topic Models

- Topic models apply to text documents
- They take the stance that each document is “about” a set of topics, in some proportions
    - So a news article is partly about foreign affairs, partly about war, partly about politics, etc.
- Topics are basically estimated like clustering

Topic Modeling

- Our document corpus is turned into a term-document matrix
  - The rows are documents
  - The columns are words
- “Topics” represent a relative weighting on different terms
  - e.g. 25% “politics”, 50% “war”, 25% “leader” for a topic about international security.
- Every document is expressed as a set of weights on all the different topics
  - So a document is 25% about international security, 40% domestic politics, 35% international affairs

Georgia Tech

Topics in NYT articles

- Source: Blei (2012)
- You can tell what broad classes of topics there are

Lesson Summary

- More non-parametric methods
  - Clustering is a way to group data points that are “similar”
  - One of the major issues with clustering is high dimensionality
  - Principle Component Analysis helps mitigate the issue of dimensionality by allowing one to look at the “most relevant part”
- Topic models are used on text documents with “topics” that have relative weights on separate terms