Clustering: (Spherical) K-Means, Document Clustering

Rebecca Nugent
Department of Statistics, Carnegie Mellon University
http://www.stat.cmu.edu/∼rnugent/PCMI2016

PCMI Undergraduate Summer School 2016

July 15, 2016

What did we think about last time?

- Clustering Goal: identify distinct groups in a data set and assign a group label to each observation; observations are partitioned such that observations in one subset are more similar to each other than to observations in different subsets
- Algorithmic vs Statistical Approaches
- Hierarchical Clustering
  - Dendrogram stores solutions for 1 to n clusters
  - User chooses distance/dissimilarity, linkage type, threshold
- K-means: partitioning obs into spherical clusters; algorithm iterates between choosing optimal assignments and optimal centers

Now we'll

- revisit K-Means
- motivate spherical k-means
- work with document clustering

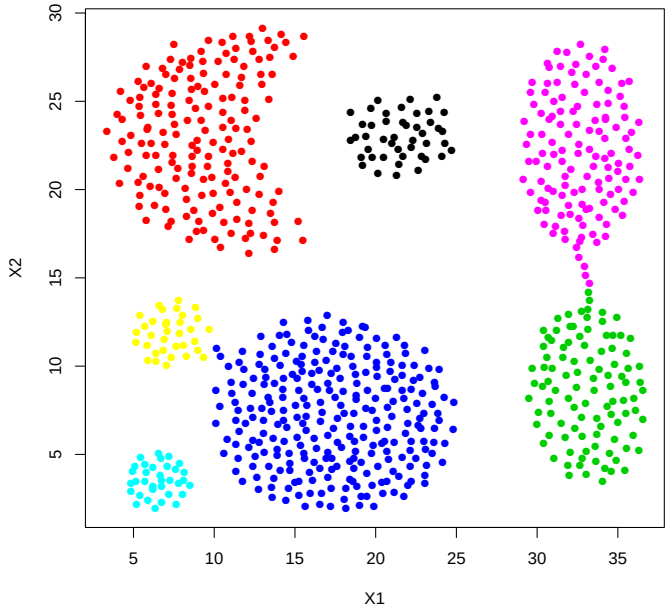
Back to K-means

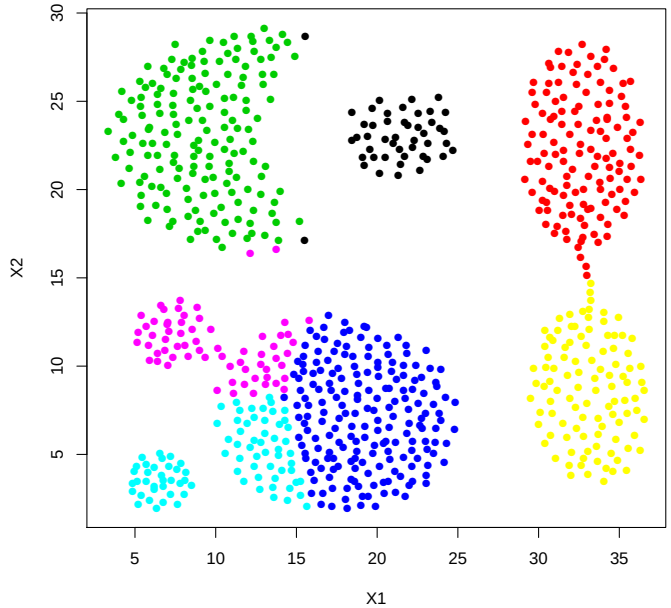
- Measure cluster "quality" : *within-cluster squared-error criterion*

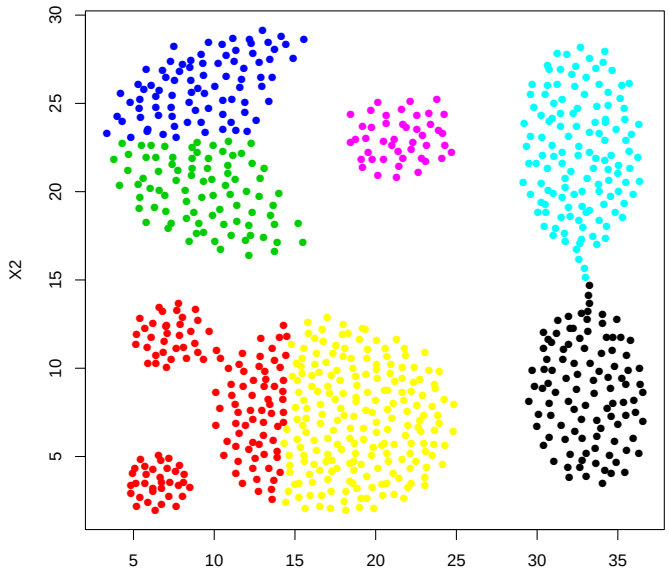
$$\sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - \bar{x}_k)^2$$

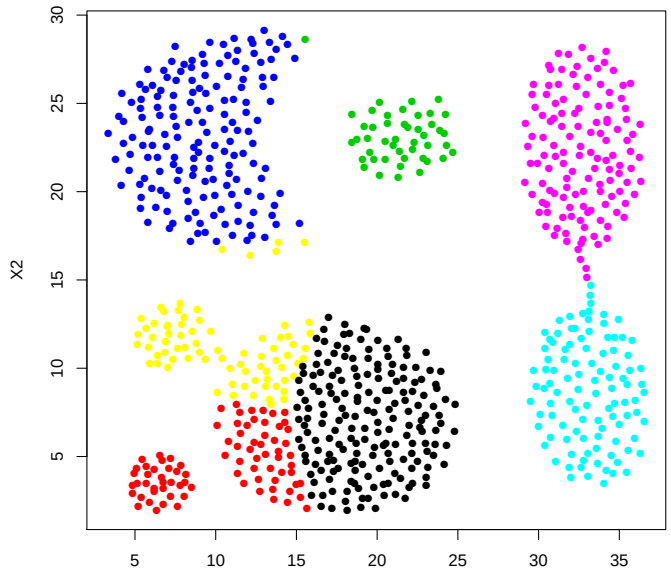
- Set the number of clusters, K, in advance.
- Iterative alternating between estimating the centers $\bar{x}_k$ and assigning the observations
- First few steps correspond to large drops in the criterion; later steps correspond to negligible drops.
- Use K randomly chosen observations as the starting centers (but don't have to; can choose specific centers)
- In theory, converge on global optimum (D Pollard); in practice, solutions can vary given set of starting centers

K-Means Variation



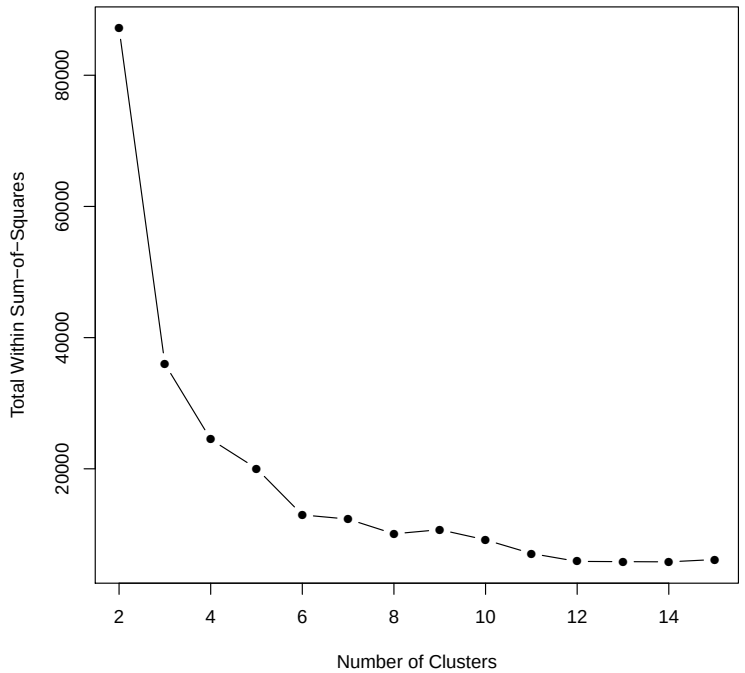






K-means

We use an *elbow graph* to determine a "useful" K.



Caveat: Criterion should always go down with K; why sometimes do we see an increase in our elbow graph?
What else could we do to find a stable solution?

Spherical K-Means

In spherical K-means, we make the following changes:

- all vectors are normalized
- we use cosine dissimilarity

$$d(x, \bar{x}_k) = 1 - \cos(x, \bar{x}_k) = 1 - \frac{< x, \bar{x}_k >}{\|x\| \|\bar{x}_k\|}$$

Why?

Sometimes we want to think about the angles between our observations; we might also want to normalize our observations for fairer comparisons/analyses

Note: Spherical k-means is the same as projecting our observations to the unit sphere and then using Euclidean distance

Document Clustering

Might want to cluster emails, blog posts, letters, articles, tweets.

What kind of data do we have?
What makes two documents similar?

Common to describe each document by the words (or phrases) that it contains. One measure is the Term Frequency - Inverse Document Frequency (TF-IDF):

$$TFIDF_{td} = \frac{n_{td}}{n_d} \cdot \log \frac{D}{D_t}$$

Interested in removing influence of terms/words that appear everywhere; want to identify important "tag words" that will indicate the cluster

How would the length of the document influence this transformation?

Document-Term Matrix

Store these TF-IDF values in a matrix
(rows = no. of documents; cols = no. of terms)
Cluster the TF-IDF observations using spherical k-means

What kind of dimensionality problems are we going to have?
What can we do about it?

- Punctuation
- Stripping
- Stop Words
- Numbers
- Anything else?

Let's look at some examples