Document Clustering

CISC489/689-010, Lecture #17

Monday, April 20th

Ben Carterette

Classification Review

- Items (documents, web pages, emails) are represented with features
- Some items are assigned a class from a fixed set
- Classification goal: use known class assignments to "learn" a general function f(x) for classifying new instances
- Naïve Bayes classifier:

$$f(x) = \arg\max_j P(C_j|x) = \arg\max_j \prod_{i=1}^{n} P(t_i|C_j)P(C_j)$$

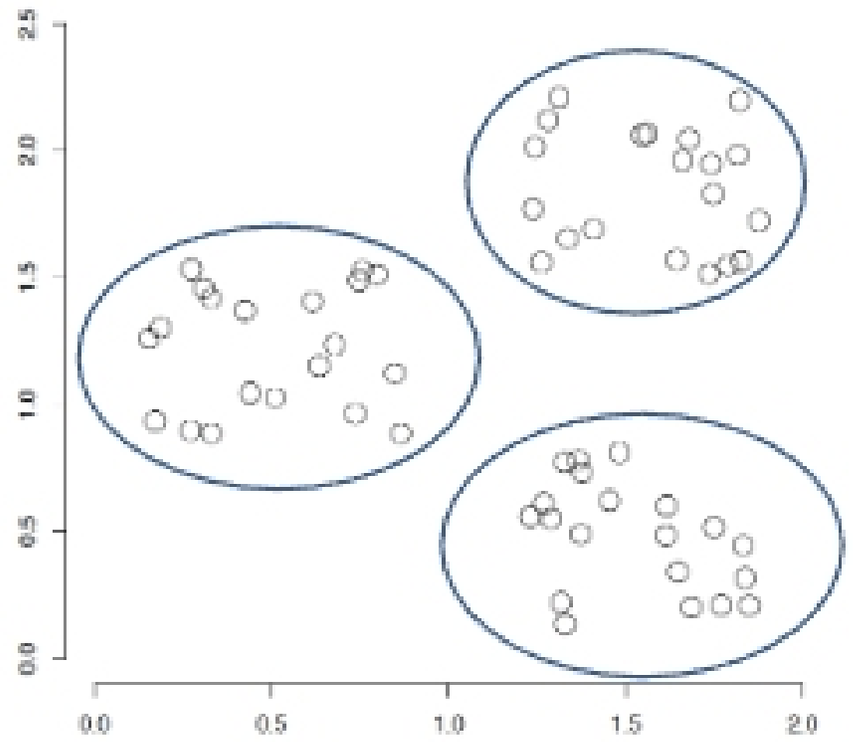
Clustering

- A set of algorithms that attempt to find latent (hidden) structure in a set of items
- Goal is to identify groups (clusters) of similar items
  - Two items in the same group should be similar to one another
  - An item in one group should be dissimilar to an item in another group

Clustering Example

- Suppose I gave you the shape, color, vitamin C content, and price of various fruits and asked you to cluster them
  - What criteria would you use?
  - How would you define similarity?
- Clustering is very sensitive to how items are represented and how similarity is defined!

Clustering in Two Dimensions



How would you cluster these points?

Classification vs Clustering

- Classification is *supervised*
  - You are given a fixed set of classes
  - You are given class labels for certain instances
  - This is data you can use to learn the classification function
- Clustering is *unsupervised*
  - You are not given any information about how documents should be grouped
  - You don't even know how many groups there should be
  - There is no training data to learn from
- One way to think of it: learning vs discovery