Simple and Quick: k-means

Description
Failings
Challenges

By David Jackson, 9/13/07
The problem:

- We have data points, we want to find assign data points to k clusters. What should we do?
K-means

- Suppose you have data $X_n$ we want to assign binary indicator variable $r_{nk} \in \{0,1\}$ where $k = 1, \ldots, K$ if $x_j$ is in cluster $l$ then $r_{jk} = \begin{cases} 1 & \text{if } l = k \\ 0 & \text{otherwise} \end{cases}$

- How do we assign $r_{nk}$?

- Objective function: Distortion

Objective function:

$$I = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \| x_n - \mu_k \|^2$$

or

$$I = \sum_{k=1}^{K} \sum_{n: r_{nk,l=k}}^{N} \| x_{nk} - \mu_k \|^2$$

Argument based on Bishop P.R.M.L 2006, section 9.1
Running K-means

- We want to minimize J.
- Step 1. Expectation, assign \( r_{nk} \) to the cluster that increases the distortion by the smallest amount:
  \[
  r_{nk} = \begin{cases} 
  1 & \text{if } k = \arg \min_k \|x_n - \mu_k\|^2 \\
  0 & \text{otherwise}
  \end{cases}
  \]

- Maximization step, take the derivative of J
  \[
  2 \sum_{n=1}^{N} r_{nk} (x_n - \mu_k) = 0
  \]
  Solve for \( u_k \):
  \[
  \mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}}
  \]
K-means: an Example

Gregory Shakhnarovich (used w/out permission)
Why not? (part 1)

We humans have an idea of continuity between data points in the same cluster:

Looking at J, assigning points to concentric circles represents the worst case scenario!

[Images: Ng 2001]
Why not? (part 2)

Consider these points, with $k = 2$
What is K? (a Challenge)

• With most clustering techniques (EM, spectral, amalgamative) the choice of K, the number of clusters, is often somewhat arbitrary but crucial.

• Some modern non-parametric methods/procedures have been created to address this issue.
What is K? (a Heuristic)

- How can we set $k$?

- The relevant statistic: *within-class dissimilarity*

$$W_k = \sum_{c=1}^{k} \sum_{y_i=y_j=c} \| x_i - x_j \|^2.$$ 

- A popular (heuristic) strategy: look for an “elbow” in $W_k$
What is K? (an Answer?)

• We’ll read a paper by Zoubin Ghahramani that proposes a way to “discover” K.

• Any good ideas on this question would be appreciated 1. by me, 2. by the field.