UNSUPERVISED LEARNING 2011

LECTURE :K-MEANS

Rita Osadchy

Some slides are due to Eric Xing, Olga Veksler
What is clustering?

- **Input:**
  - Training samples \( \{x_1, \ldots, x_m\} \in \mathbb{R}^n \)
  - No labels \( y_i \) are given

- **Goal:** group input samples into classes of similar objects – cohesive “clusters.”
  - high intra-class similarity
  - low inter-class similarity
  - It is the commonest form of unsupervised learning
First (?) Application of Clustering

- John Snow, a London physician plotted the location of cholera deaths on a map during an outbreak in the 1850s.
- The locations indicated that cases were clustered around certain intersections where there were polluted wells -- thus exposing both the problem and the solution.

From: Nina Mishra HP Labs
Application of Clustering

- Astronomy
  - SkyCat: Clustered $2 \times 10^9$ sky objects into stars, galaxies, quasars, etc based on radiation emitted in different spectrum bands.

From: Nina Mishra HP Labs
Applications of Clustering

- Image segmentation
  - Find interesting “objects” in images to focus attention at

From: Image Segmentation by Nested Cuts, O. Veksler, CVPR2000
Applications of Clustering

- Image Database Organization
  - for efficient search
Applications of Clustering

- Data Mining
  - Technology watch
    - Derwent Database, contains all patents filed in the last 10 years worldwide
    - Searching by keywords leads to thousands of documents
    - Find clusters in the database and find if there are any emerging technologies and what competition is up to

- Marketing
  - Customer database
  - Find clusters of customers and tailor marketing schemes to them
Applications of Clustering

- gene expression profile clustering
  - similar expressions, expect similar function

<table>
<thead>
<tr>
<th>Gene</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>U18675 4CL</td>
<td>-0.151</td>
<td>-0.207</td>
<td>0.126</td>
</tr>
<tr>
<td>M84697 a-TUB</td>
<td>0.188</td>
<td>0.030</td>
<td>0.111</td>
</tr>
<tr>
<td>M95595 ACC2</td>
<td>0.000</td>
<td>0.041</td>
<td>0.000</td>
</tr>
<tr>
<td>X66719 ACO1</td>
<td>0.058</td>
<td>0.155</td>
<td>0.082</td>
</tr>
<tr>
<td>U41998 ACT</td>
<td>0.096</td>
<td>-0.019</td>
<td>0.137</td>
</tr>
<tr>
<td>AF057044 ACX1</td>
<td>0.268</td>
<td>0.403</td>
<td>0.679</td>
</tr>
<tr>
<td>AF057043 ACX2</td>
<td>0.415</td>
<td>0.000</td>
<td>-0.053</td>
</tr>
<tr>
<td>U40856 AIG1</td>
<td>0.096</td>
<td>-0.106</td>
<td>-0.027</td>
</tr>
<tr>
<td>U40857 AIG2</td>
<td>0.311</td>
<td>0.140</td>
<td>0.257</td>
</tr>
<tr>
<td>AF123253 AIM1</td>
<td>-0.040</td>
<td>-0.202</td>
<td>-0.040</td>
</tr>
<tr>
<td>X92510 AOS</td>
<td>0.473</td>
<td>0.560</td>
<td>0.914</td>
</tr>
</tbody>
</table>

Applications of Clustering

- Profiling Web Users
  - Use web access logs to generate a feature vector for each user
  - Cluster users based on their feature vectors
  - Identify common goals for users
    - Shopping
    - Job Seekers
    - Product Seekers
    - Tutorials Seekers
  - Can use clustering results to improving web content and design
The k-means clustering algorithm

1. Initialize cluster centroids $\mu_1, \ldots, \mu_k \in \mathbb{R}^n$ randomly.
2. Repeat until convergence: {
   For every $i$, set
   $$c_i = \arg\min_j \|x_i - \mu_j\|^2$$
   For each $j$, set
   $$\mu_i = \frac{\sum_{i=1}^{m} 1\{c_i = j\} x_i}{\sum_{i=1}^{m} 1\{c_i = j\}}$$
}
K-means, comments.

- $k$ – the number of clusters
  a parameter of the algorithm.
- $\mu_i$  cluster centroids
  represent our current guesses for the positions of the centers of the clusters

Initialization: pick $k$ random training samples.
  Other initialization methods are also possible.
K-means, intuition

- The inner-loop of the algorithm repeatedly carries out two steps:
  
  (i) “Assigning” each training example $x_i$ to the closest cluster centroid $\mu_j$.

  (ii) Moving each cluster centroid $\mu_j$ to the mean of the points assigned to it.
K-means, example
Coordinate Descent

- Minimize a multivariate function $F(x)$ by minimizing it along one direction at a time.
  - Choose search directions from the coordinate directions.
  - Minimizes the $F(x)$ along one coordinate direction at a time, iterating through the list of search directions cyclically.

- Given $x^k$, the $i$th coordinate of $x^{k+1}$ is given by
  \[
  x_{i}^{k+1} = \arg \min_{y \in \mathbb{R}} f(x_{1}^{k+1}, ..., x_{i-1}^{k+1}, y, x_{i+1}^{k}, ..., x_{n}^{k}),
  \]
  - Thus one begins with an initial guess $x^0$ for a local minimum of $F$, and get a sequence $x^0, x^1, x^2, ...$ iteratively.
  - By doing line search in each iteration, we automatically have
    \[
    F(x^0) \geq F(x^1) \geq F(x^2), ...
    \]
  - It can be shown that this sequence has similar convergence properties as steepest descent.
Coordinate Descent Example

\[ 5x^2 - 6xy + 5y^2 - 0.0259 = 0 \]
K-means, convergence

- Define objective function:

\[ J(c, \mu) = \sum_{i=1}^{m} ||x_i - \mu_{c_i}||^2 \]

- k-means is exactly coordinate descent on \( J \).

Inner-loop of k-means repeatedly
- minimizes \( J \) with respect to \( c \) while holding \( \mu \) fixed
- minimizes \( J \) with respect to \( \mu \) while holding \( c \) fixed.

Thus \( J \) must monotonically decrease \( \Rightarrow \) value of \( J \) must converge.