11: Unsupervised Learning - Clustering

CSC 4510 – Machine Learning

Dr. Mary-Angela Papalaskari
Department of Computing Sciences
Villanova University

Course website:
www.csc.villanova.edu/~map/4510/

Some of the slides in this presentation are adapted from:
• Prof. Frank Klassner’s ML class at Villanova
• the University of Manchester ML course http://www.cs.manchester.ac.uk/uet/COMP24111/
• The Stanford online ML course http://www.ml-class.org/
Supervised learning

Training set: \{ (x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)}), \ldots, (x^{(m)}, y^{(m)}) \}\n
• The Stanford online ML course http://www.ml-class.org/
Unsupervised learning

Training set: \( \{x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(m)} \} \)

• The Stanford online ML course [http://www.ml-class.org/](http://www.ml-class.org/)
Unsupervised Learning

• Learning “what normally happens”
• No output
• Clustering: Grouping similar instances
• Example applications
  – Customer segmentation
  – Image compression: Color quantization
  – Bioinformatics: Learning motifs
Clustering Algorithms

• K means
• Hierarchical
  – Bottom up or top down
• Probabilistic
  – Expectation Maximization (E-M)
Clustering algorithms

• **Partitioning method:** Construct a partition of n examples into a set of K clusters
  - Given: a set of examples and the number K
  - Find: a partition of K clusters that optimizes the chosen partitioning criterion
    - Globally optimal: exhaustively enumerate all partitions
    - Effective heuristic method: **K-means algorithm**.

http://www.csee.umbc.edu/~nicholas/676/MRSSlides/lecture17-clustering.ppt
K-Means

• Assumes instances are real-valued vectors.
• Clusters based on centroids, center of gravity, or mean of points in a cluster, $c$
• Reassignment of instances to clusters is based on distance to the current cluster centroids.
K-means intuition

- Randomly choose k points as seeds, one per cluster.
- Form initial clusters based on these seeds.
- Iterate, repeatedly reallocating seeds and by re-computing clusters to improve the overall clustering.
- Stop when clustering converges or after a fixed number of iterations.

Based on: www.cs.utexas.edu/~mooney/cs388/slides/TextClustering.ppt
The Stanford online ML course [http://www.ml-class.org/]
The Stanford online ML course [http://www.ml-class.org/]
The Stanford online ML course [http://www.ml-class.org/]
The Stanford online ML course [http://www.ml-class.org/]
The Stanford online ML course [http://www.ml-class.org/]
The Stanford online ML course [http://www.ml-class.org/]
The Stanford online ML course [http://www.ml-class.org/](http://www.ml-class.org/)
• The Stanford online ML course [http://www.ml-class.org/](http://www.ml-class.org/)
The Stanford online ML course [http://www.ml-class.org/](http://www.ml-class.org/)
K-Means Algorithm

• Let $d$ be the distance measure between instances.
• Select $k$ random points $\{s_1, s_2, \ldots, s_k\}$ as seeds.
• Until clustering converges or other stopping criterion:
  – For each instance $x_i$:
    • Assign $x_i$ to the cluster $c_j$ such that $d(x_i, s_j)$ is minimal.
  – (Update the seeds to the centroid of each cluster)
    • For each cluster $c_j$, $s_j = \mu(c_j)$
Distance measures

• Euclidean distance
• Manhattan
• Hamming
Orange schema
Orange schema
Clusters aren’t always separated...

120 Sheer Fashion Pantyhose Sizing Chart

<table>
<thead>
<tr>
<th>Size</th>
<th>16</th>
<th>16.5</th>
<th>17</th>
<th>17.5</th>
<th>18</th>
<th>18.5</th>
<th>19</th>
<th>19.5</th>
<th>20</th>
<th>20.5</th>
<th>21</th>
<th>21.5</th>
<th>22</th>
<th>22.5</th>
<th>23</th>
<th>23.5</th>
<th>24</th>
<th>24.5</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>44</td>
<td>44.5</td>
<td>45</td>
<td>45.5</td>
<td>46</td>
<td>46.5</td>
<td>47</td>
<td>47.5</td>
<td>48</td>
<td>48.5</td>
<td>49</td>
<td>49.5</td>
<td>50</td>
<td>50.5</td>
<td>51</td>
<td>51.5</td>
<td>52</td>
<td>52.5</td>
<td>53</td>
</tr>
</tbody>
</table>

http://store02.prostores.com/selectsocksinc/images/store_version1/Sigvaris%20120%20Pantyhose%20SIZE%20chart.gif
K-means for non-separated clusters

T-shirt sizing

• The Stanford online ML course [http://www.ml-class.org/]
Weaknesses of k-means

• The algorithm is only applicable to numeric data.
• The user needs to specify k.
• The algorithm is sensitive to outliers
  – Outliers are data points that are very far away from other data points.
  – Outliers could be errors in the data recording or some special data points with very different values.
Strengths of k-means

- **Strengths:**
  - Simple: easy to understand and to implement
  - Efficient: Time complexity: $O(tkn)$,
  - where $n$ is the number of data points,
  - $k$ is the number of clusters, and
  - $t$ is the number of iterations.
  - Since both $k$ and $t$ are small, k-means is considered a linear algorithm.

- K-means is the most popular clustering algorithm.