Lecture 23: Clustering Evaluation

April 29, 2010
Today

• Cluster Evaluation
  – Internal
    • We don’t know anything about the desired labels
  – External
    • We have some information about the labels
Internal Evaluation

• Clusters have been identified.
• How successful a partitioning of the data set was constructed through clustering?
• Internal measures have the quality that they can be directly optimized.
Intercluster variability

• How far is a point from its cluster centroid

\[ J(x|\theta) = \frac{1}{N} \sum_{i} ||x_i - c_i||^2 \]

• Intuition: every point assigned to a cluster should be closer to the center of that cluster than any other cluster

• K-means optimizes this measure.
Model Likelihood

\[ p(x|\theta) \]

• Intuition: the model that fits the data best represents the best clustering

• Requires a probabilistic model.

• Can be included in AIC and BIC measures to limit the number of parameters.

• GMM style: \[ p(x|\theta) = \sum_k \pi_k p(x|\mu_k, \Sigma_k) \]
Point similarity vs. Cluster similarity

\[ \frac{1}{2} \sum_{ij} \text{sim}(x_i, x_j)(c_i - c_j)^2 \]

- Intuition: two points that are similar should be in the same cluster
- Spectral Clustering optimizes this function.
Internal Cluster Measures

• Which cluster measure is best?
  – Centroid distance
  – Model likelihood
  – Point distance

• It depends on the data and the task.
External Cluster Evaluation

• If you have a little bit of labeled data, unsupervised (clustering) techniques can be evaluated using this knowledge.

• Assume for a subset of the data points, you have class labels.

• How can we evaluate the success of the clustering?
External Cluster Evaluation

• Can’t we use Accuracy?
  – or “Why is this hard?”
• The number of clusters may not equal the number of classes.
• It may be difficult to assign a class to a cluster.
Some principles.

• Homogeneity
  – Each cluster should include members of as few classes as possible

• Completeness
  – Each class should be represented in as few clusters as possible.
Some approaches

• Purity

\[
Purity = \sum_{r=1}^{k} \frac{1}{n} \max_{i} (n_{r,i})
\]

points of class \(i\) in cluster \(r\)

• F-measure
  
  – Cluster definitions of Precision and Recall
  
  – Combined using harmonic mean as in traditional f-measure

\[
R(c_i, k_j) = \frac{n_{ij}}{|c_i|}
\]

\[
P(c_i, k_j) = \frac{n_{ij}}{|k_j|}
\]
The problem of matching

F-measure: 0.6

F-measure: 0.6
The problem of matching

F-measure: 0.5

F-measure: 0.5
V-Measure

• Conditional Entropy based measure to explicitly calculate homogeneity and completeness.

\[ V_\beta = \frac{(1 + \beta) \times h \times c}{(\beta \times h) + c} \]

\[ h = \begin{cases} 
1 & \text{if } H(C, K) = 0 \\
1 - \frac{H(C|K)}{H(C)} & \text{else}
\end{cases} \]

\[ c = \begin{cases} 
1 & \text{if } H(K, C) = 0 \\
1 - \frac{H(K|C)}{H(K)} & \text{else}
\end{cases} \]
Contingency Matrix

- Want to know how much the introduction of clusters is improving the information about the class distribution.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>3</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>
Entropy

- Entropy calculates the amount of “information” in a distribution.
- Wide distributions have a lot of information
- Narrow distributions have very little
- Based on Shannon’s limit of the number of bits required to transmit a distribution
- Calculation of entropy:

\[ H(x) = - \sum_{i} p(x_i) \log_2 p(x_i) \]
Example Calculation of Entropy

$x = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$

$H(x) = -\frac{1}{4} \log \frac{1}{4} - \frac{1}{4} \log \frac{1}{4} - \frac{1}{4} \log \frac{1}{4} - \frac{1}{4} \log \frac{1}{4}$

$H(x) = -\frac{1}{4}(-2) - \frac{1}{4}(-2) - \frac{1}{4}(-2) - \frac{1}{4}(-2)$

$= 2$
Example Calculation of Entropy

\[ x = \begin{bmatrix} \frac{4}{10} & \frac{4}{10} & \frac{1}{10} & \frac{1}{10} \end{bmatrix} \]

\[
H(x) = -\frac{4}{10} \log \frac{4}{10} - \frac{4}{10} \log \frac{4}{10} - \frac{1}{10} \log \frac{1}{10} - \frac{1}{10} \log \frac{1}{10} \\
= -\frac{4}{10}(-1.32) - \frac{4}{10}(-1.32) - \frac{1}{10}(-3.32) - \frac{1}{10}(-3.32) \\
= 1.72
\]
Pair Based Measures

• Statistics over every pair of items.
  – SS – same cluster, same class
  – SD – same cluster, different class
  – DS – different cluster, same class
  – DD – different cluster different class

• These can be arranged in a contingency matrix similar to when accuracy is constructed.
Pair Based Measures

- **Rand:**
  \[ Rand = \frac{SS + DD}{SS + SD + DS + DD} \]

- **Jaccard**
  \[ Jaccard = \frac{SS}{SS + SD + DS} \]

- **Folkes-Mallow**
  \[ FM = \sqrt{\frac{SS}{SS + SD} \cdot \frac{SS}{SS + DS}} \]
B-cubed

- Similar to pair based counting systems, B-cubed calculates an element by element precision and recall.

\[
\text{Precision} = \frac{\sum_i \text{Precision}(e)}{n} \quad \text{Recall} = \frac{\sum_i \text{Recall}(e)}{n}
\]
External Evaluation Measures

• There are many choices.

• Some should almost certainly never be used
  – Purity
  – F-measure

• Others can be used based on task or other preferences
  – V-Measure
  – VI
  – B-Cubed
Next Time

• Project Presentations
  – The schedule has 15 minutes per presentation.
  – This includes transition to the next speaker, and questions.
  – Prepare for 10 minutes.

• Course Evaluations
Thank you