Clustering

• Here we assume data is in $\mathbb{R}^n$
  • (some methods can work with distance directly without assuming $\mathbb{R}^n$)

• Task: partition data into groups in some sensible way
  • There is more than one way to define desirable outcomes. For example ...

Clustering Evaluation

• How can we evaluate how good our clustering is?
  - Evaluation by our criterion
  - Evaluation by expert
  - Evaluation by using clustering result for other task.

• Comparing different clustering results (and/or comparing to labels)
  - Evaluation by NMI - defined later on slides

Clustering Evaluation

• Sensitivity to feature scaling and transformations

Comparison of clustering results:

Unnormalized vs. normalized.

Visualization from Carla Brodley’s slides

Clustering

• Basic Definitions and Notation
  
  Partition into $C_1, \ldots, C_k$

  $\mu_j = \frac{1}{|C_j|} \sum_{x \in C_j} x$

  $\mu = \frac{1}{N} \sum_j \sum_{x \in C_j} x$

What does this look like?

Some Clustering Criteria

• (Minimize) Cluster Scatter

  $CS = \sum_j \sum_{x \in C_j} ||x - \mu_j||^2$

• (Maximize) Cluster Distance

  $CD = \sum_j |C_j| \cdot ||\mu_j - \mu||^2$

• (Maximize) Spacing

  $\text{Spacing} = \min_{i,j} \min_{x \in C_i, y \in C_j} ||x - y||^2$

What do these look like?
### Agglomerative Hierarchical Clustering
- **Init**: each data point as single cluster
- **Repeat**:
  - Find two clusters which are "most similar"
  - Replace them with their union
- This requires a distance function over clusters

### Hierarchical Clustering
- **Distance b/w pair of clusters**:
  - $d_{\text{min}}(C_i, C_j) = \min_{x \in C_i, y \in C_j} \| x - y \|^2$
  - $d_{\text{max}}(C_i, C_j) = \max_{x \in C_i, y \in C_j} \| x - y \|^2$
  - $d_{\text{avg}}(C_i, C_j) = \frac{1}{|C_i| \cdot |C_j|} \sum_{x \in C_i} \sum_{y \in C_j} \| x - y \|^2$
- $d_{\text{min}}$ optimizes Spacing and yields MST of data points

### Divisive Hierarchical Clustering
- **Init**: all data points form one cluster
- **Repeat**:
  - Pick a cluster and "the best split" of that cluster
  - Replace cluster with its sub-parts
- Requires quality criterion for split: can use distance function over clusters, or a global criterion for the resulting clustering.

### k-Means Clustering
- **Pick k cluster centers** *(how?)*
- **Repeat**:
  - Associate examples with centers
    - pick nearest center
  - Re-calculate means
    - as average of examples in cluster
- Until convergence

### Soft k-Means Clustering
- **Pick k cluster centers**
- **Repeat**:
  - Associate examples with centers
    - $p_{ij} \sim \text{similarity b/w example } i \text{ and center } j$
  - Re-calculate means
    - as weighted average of examples in cluster
- Until convergence

### k-Means Clustering
- Result sensitive to initialization
- **Can we get around that?**
- Calculation of mean is sensitive to outliers
- **Can we get around that?**
k-Medoids Clustering

- Pick k cluster medoids
- Repeat:
  - Associate examples with medoids pick nearest medoid
  - Re-calculate medoids the example in cluster that has the smallest mean distance to other points in the cluster
- Until convergence

Spectral Clustering

- Can use any distance function
- Or a weighted adjacency matrix of graph induced by examples
- To produce "Laplacian" similarity matrix
- Performs standard clustering on eigen-decomposition of that matrix
- [details beyond scope of course]

How to Choose k?

- Solution 1:
  - Run algorithm with k=2,3,...
  - Evaluate criterion (e.g. CS) for each run
  - Hope to see big drop in criterion until we get "the right k" and moderate drop after that

How to Choose k?

- Solution 2: BIC criterion - add penalty for number of clusters
  - \( \text{BIC} = (\text{min criterion}) + k \log(N) \)
  - Increase k:
    - CS goes down, penalty goes up
    - For some k total starts going up

Comparing Clustering Results

- Sometimes it is useful to check if two clustering results are close or not
- For purpose of evaluating new clustering algorithm: we can compare its results to labels on a labeled dataset
- How? NMI

Mutual Information

Joint entropy: uncertainty/code length for X,Y together
\[ H(X,Y) = \sum_x \sum_y p(x,y) \log \frac{1}{p(x,y)} \]
Condition entropy: additional cost to encode Y given X
\[ H(Y|X) = \sum_x \sum_y p(x,y) \log \frac{1}{p(y|x)} = \sum_x \sum_y p(x,y) \log \frac{p(x)}{p(x,y)} \]
Mutual Information: the average code-length-saving for encoding Y due to knowing X for encoding X due to knowing Y
\[ I(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(y)} + \sum_x \sum_y p(x,y) \log \frac{1}{p(y)} = H(Y) - H(Y|X) = H(X) - H(X|Y) \]
**Comparing Clustering Results**

U, V are two clustering results of R and C clusters respectively

<table>
<thead>
<tr>
<th>U \ V</th>
<th>V_1</th>
<th>V_2</th>
<th>...</th>
<th>V_C</th>
<th>Sums</th>
</tr>
</thead>
<tbody>
<tr>
<td>U_1</td>
<td>n_{11}</td>
<td>n_{12}</td>
<td>...</td>
<td>n_{1C}</td>
<td>a_1</td>
</tr>
<tr>
<td>U_2</td>
<td>n_{21}</td>
<td>n_{22}</td>
<td>...</td>
<td>n_{2C}</td>
<td>a_2</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>U_R</td>
<td>n_{R1}</td>
<td>n_{R2}</td>
<td>...</td>
<td>n_{RC}</td>
<td>a_R</td>
</tr>
<tr>
<td>Sums</td>
<td>b_1</td>
<td>b_2</td>
<td>...</td>
<td>b_C</td>
<td>\sum_j n_{ij} = N</td>
</tr>
</tbody>
</table>

Table 1: The Contingency Table, \( n_{ij} = |U_i \cap V_j| \)

**Comparing Clustering Results**

\[
H(U) = - \sum_{i=1}^{R} \frac{a_i}{N} \log \frac{a_i}{N},
\]

\[
H(U, V) = - \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{n_{ij}}{N} \log \frac{n_{ij}}{N},
\]

\[
I(U, V) = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{n_{ij}}{N} \log \frac{n_{ij}/N}{a_i b_j/N^2}.
\]

**Comparing Clustering Results**

- Mutual Information is sensitive to the number of clusters so that partitions into more clusters will artificially have higher mutual information.
- Normalized Mutual Information corrects for that. Multiple formulations exist. Here we divide by the average entropy:

\[
NMI_{\text{sum}} = \frac{2I(U, V)}{H(U) + H(V)}
\]

**Clustering**

- Data Exploration
- Evaluation by ...
- Several possible criteria
- Hierarchical vs. k-way-partition
- Several algorithms discussed

- Model selection (pick k)
- Comparing different partitions