Comp 135 Introduction to Machine Learning and Data Mining

Fall 2016

Professor: Roni Khardon

Computer Science
Tufts University

Clustering

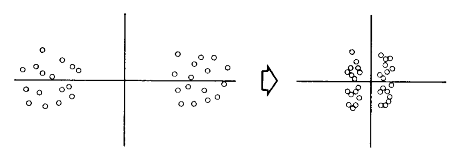
- Here we assume data is in $\mathbf{R}^n$
- (some methods can work with distance directly without assuming $\mathbf{R}^n$)
- Task: partition data into groups in some sensible way
- There is more than one way to define desirable outcomes. For example …

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

Clustering Evaluation

- Sensitivity to feature scaling and transformations



(a) UNNORMALIZED (b) NORMALIZED

Visualization from Carla Brodley's slides

Clustering

- Basic Definitions and Notation

Partition into $C_1, \ldots, C_k$

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

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

Some Clustering Criteria

- (Minimize) Cluster Scatter

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

- (Maximize) Cluster Distance

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

- (Maximize) Spacing

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

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

- Which distance for clusters?

$$d_{min}(C_i, C_j) = \min_{x \in C_i, y \in C_j} \|x - y\|^2$$

$$d_{max}(C_i, C_j) = \max_{x \in C_i, y \in C_j} \|x - y\|^2$$

$$d_{avg}(C_i, C_j) = \frac{1}{|C_i| \cdot |C_j|} \sum_{x \in C_i} \sum_{y \in C_j} \|x - y\|^2$$

- d_{min} optimizes Spacing and yields MST of data points

Divisive Hierarchical Clustering

- Init: all data points from 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
    $p_{i,j}$ ~~ similarity b/w example i 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

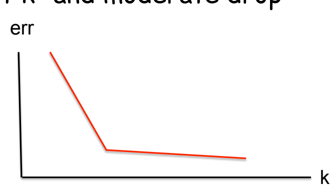
- Pick k cluster medoids
- Repeat:
  - Associate examples with medoids
    pick nearest medoid
  - Re-calculate medoid
    the example in cluster that has the smallest mean distance to other points in the cluster
- Until convergence

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
- BIC = (min criterion) + k log(N)
- = (1/N)CS + 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 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

$$H(X,Y) = \sum_x \sum_y p(x,y) \log \frac{1}{p(x,y)}$$

Conditional 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

$$I(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

$$= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)} + \sum_x \sum_y p(x,y) \log \frac{1}{p(y)}$$

$$= H(Y) - H(Y|X) = H(X) - H(X|Y)$$

Comparing Clustering Results

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

$\mathbf{U}\backslash\mathbf{V}$	V_1	V_2	$\dots$	V_C	Sums
U_1	n_{11}	n_{12}	$\dots$	n_{1C}	a_1
U_2	n_{21}	n_{22}	$\dots$	n_{2C}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
U_R	n_{R1}	n_{R2}	$\dots$	n_{RC}	a_R
Sums	b_1	b_2	$\dots$	b_C	$\sum_{ij} n_{ij} = N$

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

[from Vinh, Epps, Bailey 2010]

Comparing Clustering Results

$$\begin{aligned}
H(\mathbf{U}) &= -\sum_{i=1}^{R} \frac{a_i}{N} \log \frac{a_i}{N}, \\
H(\mathbf{U},\mathbf{V}) &= -\sum_{i=1}^{R} \sum_{j=1}^{C} \frac{n_{ij}}{N} \log \frac{n_{ij}}{N}, \\
H(\mathbf{U}|\mathbf{V}) &= -\sum_{i=1}^{R} \sum_{j=1}^{C} \frac{n_{ij}}{N} \log \frac{n_{ij}/N}{b_j/N}, \\
I(\mathbf{U},\mathbf{V}) &= \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{n_{ij}}{N} \log \frac{n_{ij}/N}{a_i b_j/N^2}.
\end{aligned}$$

[from Vinh, Epps, Bailey 2010]

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:

$$\mathrm{NMI}_{sum} = \frac{2I(\mathbf{U},\mathbf{V})}{H(\mathbf{U})+H(\mathbf{V})}$$

[from Vinh, Epps, Bailey 2010]

Clustering

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

- Model selection (pick k)
- Comparing different partitions