

CS249: ADVANCED DATA MINING

Clustering Evaluation and Practical Issues

Instructor: Yizhou Sun

yzsun@cs.ucla.edu

May 2, 2017

Announcements

- Homework 2 due later today
 - Due May 3rd (11:59pm)


- Course project proposal
 - Due May 8th (11:59pm)

- Homework 3 out
 - Due May 10th (11:59pm)

Learnt Clustering Methods

| | Vector Data | Text Data | Recommender System | Graph & Network |
|------------------------|---|----------------|-------------------------|---------------------------|
| Classification | Decision Tree; Naïve Bayes; Logistic Regression SVM; NN | | | Label Propagation |
| Clustering | K-means; hierarchical clustering; DBSCAN; Mixture Models; kernel k-means | PLSA; LDA | Matrix Factorization | SCAN; Spectral Clustering |
| Prediction | Linear Regression GLM | | Collaborative Filtering | |
| Ranking | | | | PageRank |
| Feature Representation | | Word embedding | | Network embedding |

Evaluation and Other Practical Issues

- Evaluation of Clustering 
- Similarity and Dissimilarity
- Summary

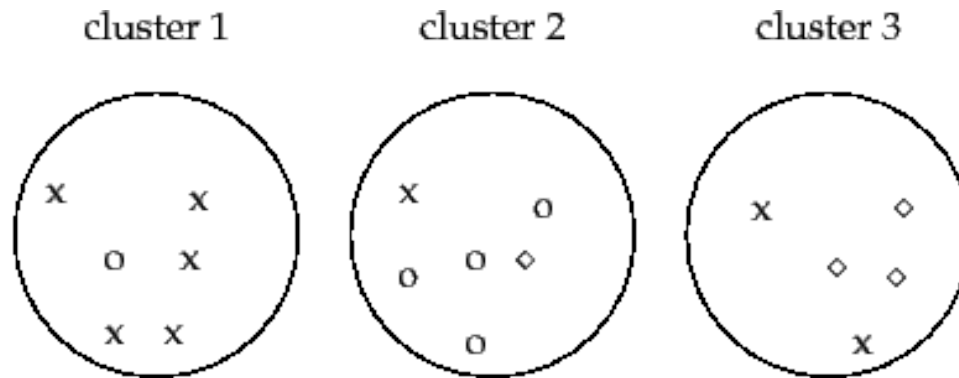
Measuring Clustering Quality

- Two methods: extrinsic vs. intrinsic
- Extrinsic: supervised, i.e., the ground truth is available
 - Compare a clustering against the ground truth using certain clustering quality measure
 - Ex. Purity, precision and recall metrics, normalized mutual information
- Intrinsic: unsupervised, i.e., the ground truth is unavailable
 - Evaluate the goodness of a clustering by considering how well the clusters are separated, and how compact the clusters are
 - Ex. Silhouette coefficient

Purity

- Let $\mathbf{C} = \{c_1, \dots, c_K\}$ be the output clustering result, $\mathbf{\Omega} = \{\omega_1, \dots, \omega_J\}$ be the ground truth clustering result (ground truth class)
 - c_k and w_k are sets of data points
 - $\text{purity}(\mathbf{C}, \mathbf{\Omega}) = \frac{1}{N} \sum_k \max_j |c_k \cap \omega_j|$

Example



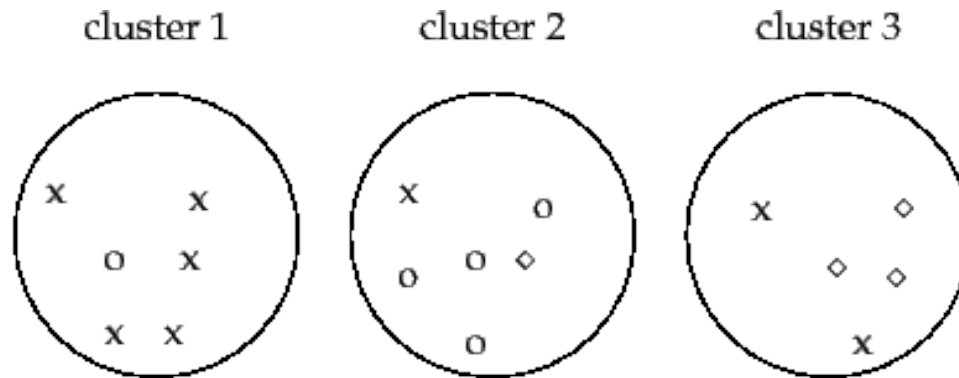
► **Figure 16.1** Purity as an external evaluation criterion for cluster quality. Majority class and number of members of the majority class for the three clusters are: x, 5 (cluster 1); o, 4 (cluster 2); and \diamond , 3 (cluster 3). Purity is $(1/17) \times (5 + 4 + 3) \approx 0.71$.

- Clustering output: cluster 1, cluster 2, and cluster 3
- Ground truth clustering result: x's, \diamond 's, and o's.
- cluster 1 vs. x's, cluster 2 vs. o's, and cluster 3 vs. \diamond 's

Normalized Mutual Information

- $NMI(C, \Omega) = \frac{I(C, \Omega)}{\sqrt{H(C)H(\Omega)}}$
- $I(\Omega, C) = \sum_k \sum_j P(c_k \cap \omega_j) \log \frac{P(c_k \cap \omega_j)}{P(c_k)P(\omega_j)}$
 $= \sum_k \sum_j \frac{|c_k \cap \omega_j|}{N} \log \frac{N|c_k \cap \omega_j|}{|c_k| \cdot |\omega_j|}$
- $H(\Omega) = -\sum_j P(\omega_j) \log P(\omega_j)$
 $= -\sum_j \frac{|\omega_j|}{N} \log \frac{|\omega_j|}{N}$

Example



$NMI=0.36$

| | $ \omega_k \cap c_j $ | | | |
|----------|-----------------------|-----------|-----------|------|
| | Cluster 1 | Cluster 2 | Cluster 3 | sum |
| crosses | 5 | 1 | 2 | 8 |
| circles | 1 | 4 | 0 | 5 |
| diamonds | 0 | 1 | 3 | 4 |
| sum | 6 | 6 | 5 | N=17 |

$|\omega_k|$

$|c_j|$

Precision and Recall


- Random Index (RI) = $(TP+TN)/(TP+FP+FN+TN)$
- F-measure: $2P*R/(P+R)$
 - $P = TP/(TP+FP)$
 - $R = TP/(TP+FN)$
- Consider pairs of data points:
 - hopefully, two data points that are in the same cluster will be clustered into the same cluster (TP), and two data points that are in different clusters will be clustered into different clusters (TN).

| | Same cluster | Different clusters |
|-------------------|--------------|--------------------|
| Same class | TP | FN |
| Different classes | FP | TN |

Example

| Data points | Output clustering | Ground truth clustering (class) |
|-------------|-------------------|---------------------------------|
| a | 1 | 2 |
| b | 1 | 2 |
| c | 2 | 2 |
| d | 2 | 1 |

- # pairs of data points: 6
 - (a, b): same class, same cluster
 - (a, c): same class, different cluster
 - (a, d): different class, different cluster
 - (b, c): same class, different cluster
 - (b, d): different class, different cluster
 - (c, d): different class, same cluster



| |
|--------|
| TP = 1 |
| FP = 1 |
| FN = 2 |
| TN = 2 |

RI = 0.5


P = 1/2, R = 1/3, F = 0.4

Question

- If we flip the ground truth cluster labels (2->1 and 1->2), will the evaluation results be the same?

| Data points | Output clustering | Ground truth clustering (class) |
|-------------|-------------------|---------------------------------|
| a | 1 | 2 |
| b | 1 | 2 |
| c | 2 | 2 |
| d | 2 | 1 |

Evaluation and Other Practical Issues

- Evaluation of Clustering
- Similarity and Dissimilarity 
- Summary

Similarity and Dissimilarity

- **Similarity**
 - Numerical measure of how alike two data objects are
 - Value is higher when objects are more alike
 - Often falls in the range $[0,1]$
- **Dissimilarity** (e.g., distance)
 - Numerical measure of how different two data objects are
 - Lower when objects are more alike
 - Minimum dissimilarity is often 0
 - Upper limit varies
- **Proximity** refers to a similarity or dissimilarity

Data Matrix and Dissimilarity Matrix

- Data matrix

- n data points with p dimensions
- Two modes

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

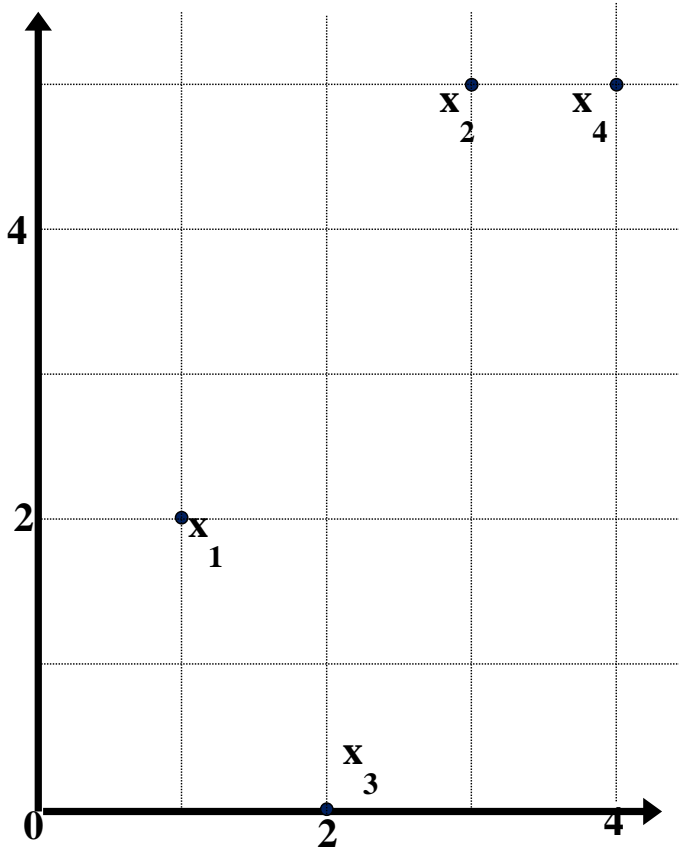
- Dissimilarity matrix

- n data points, but registers only the distance
- A triangular matrix
- Single mode

$$\begin{bmatrix} 0 & & & & & & & & & \\ d(2,1) & 0 & & & & & & & & \\ d(3,1) & d(3,2) & 0 & & & & & & & \\ \vdots & \vdots & \vdots & & & & & & & \\ d(n,1) & d(n,2) & \dots & \dots & & & & & & 0 \end{bmatrix}$$

Example:

Data Matrix and Dissimilarity Matrix



Data Matrix

| point | attribute1 | attribute2 |
|-------|------------|------------|
| $x1$ | 1 | 2 |
| $x2$ | 3 | 5 |
| $x3$ | 2 | 0 |
| $x4$ | 4 | 5 |

Dissimilarity Matrix
(with Euclidean Distance)

| | $x1$ | $x2$ | $x3$ | $x4$ |
|------|------|------|------|------|
| $x1$ | 0 | | | |
| $x2$ | 3.61 | 0 | | |
| $x3$ | 2.24 | 5.1 | 0 | |
| $x4$ | 4.24 | 1 | 5.39 | 0 |

Proximity Measure for Nominal Attributes

- Can take 2 or more states, e.g., red, yellow, blue, green (generalization of a binary attribute)
- Method 1: Simple matching
 - m : # of matches, p : total # of variables

$$d(i, j) = \frac{p - m}{p}$$

- Method 2: Use a large number of binary attributes
 - creating a new binary attribute for each of the M nominal states

Proximity Measure for Binary Attributes

- A contingency table for binary data
- Distance measure for symmetric binary variables:
- Distance measure for asymmetric binary variables:
- Jaccard coefficient (*similarity* measure for *asymmetric* binary variables):

| | | Object j | | sum |
|------------|---|------------|-------|-------|
| | | 1 | 0 | |
| Object i | 1 | q | r | $q+r$ |
| | 0 | s | t | $s+t$ |
| sum | | $q+s$ | $r+t$ | p |

$$d(i, j) = \frac{r + s}{q + r + s + t}$$

$$d(i, j) = \frac{r + s}{q + r + s}$$

$$sim_{Jaccard}(i, j) = \frac{q}{q + r + s}$$

Dissimilarity between Binary Variables

- Example

| Name | Gender | Fever | Cough | Test-1 | Test-2 | Test-3 | Test-4 |
|------|--------|-------|-------|--------|--------|--------|--------|
| Jack | M | Y | N | P | N | N | N |
| Mary | F | Y | N | P | N | P | N |
| Jim | M | Y | P | N | N | N | N |

- Gender is a symmetric attribute
- The remaining attributes are asymmetric binary
- Let the values Y and P be 1, and the value N be 0

$$d(jack, mary) = \frac{0 + 1}{2 + 0 + 1} = 0.33$$

$$d(jack, jim) = \frac{1 + 1}{1 + 1 + 1} = 0.67$$

$$d(jim, mary) = \frac{1 + 2}{1 + 1 + 2} = 0.75$$

Standardizing Numeric Data

- Z-score: $z = \frac{x - \mu}{\sigma}$
 - X: raw score to be standardized, μ : mean of the population, σ : standard deviation
 - the distance between the raw score and the population mean in units of the standard deviation
 - negative when the raw score is below the mean, “+” when above
- An alternative way: Calculate the mean absolute deviation

where

$$s_f = \frac{1}{n} (|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f|)$$
$$m_f = \frac{1}{n} (x_{1f} + x_{2f} + \dots + x_{nf}).$$

- standardized measure (z-score): $z_{if} = \frac{x_{if} - m_f}{s_f}$
- Using mean absolute deviation is more robust than using standard deviation

Distance on Numeric Data: Minkowski Distance

- *Minkowski distance*: A popular distance measure

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \dots + |x_{ip} - x_{jp}|^h}$$

where $i = (x_{i1}, x_{i2}, \dots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \dots, x_{jp})$ are two p -dimensional data objects, and h is the order (the distance so defined is also called L - h norm)

- Properties
 - $d(i, j) > 0$ if $i \neq j$, and $d(i, i) = 0$ (Positive definiteness)
 - $d(i, j) = d(j, i)$ (Symmetry)
 - $d(i, j) \leq d(i, k) + d(k, j)$ (Triangle Inequality)
- A distance that satisfies these properties is a **metric**

Special Cases of Minkowski Distance

- $h = 1$: **Manhattan** (city block, L_1 norm) **distance**
 - E.g., the Hamming distance: the number of bits that are different between two binary vectors

$$d(i, j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \dots + |x_{i_p} - x_{j_p}|$$

- $h = 2$: (L_2 norm) **Euclidean** distance

$$d(i, j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \dots + |x_{i_p} - x_{j_p}|^2)}$$

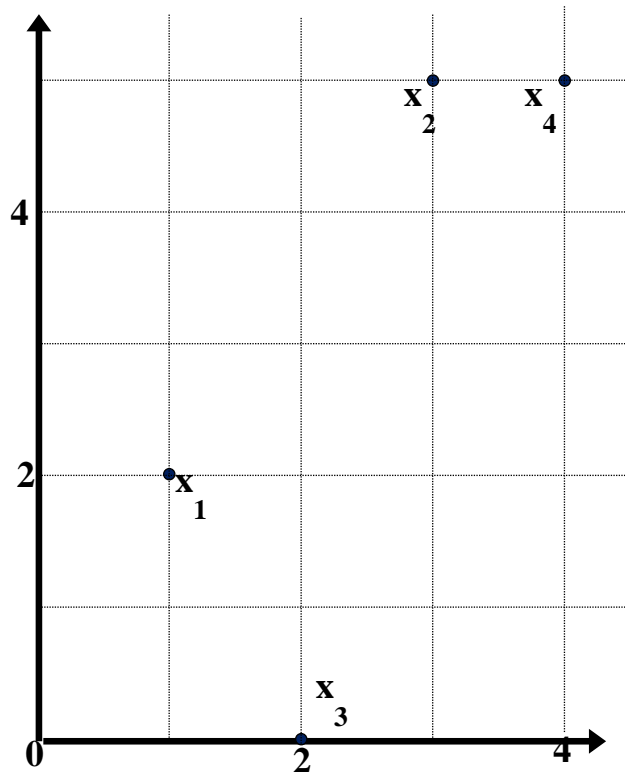
- $h \rightarrow \infty$. **“supremum”** (L_{\max} norm, L_∞ norm) distance.
 - This is the maximum difference between any component (attribute) of the vectors

$$d(i, j) = \lim_{h \rightarrow \infty} \left(\sum_{f=1}^p |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_f |x_{if} - x_{jf}|$$

Example: Minkowski Distance

Dissimilarity Matrices

| point | attribute 1 | attribute 2 |
|-------|-------------|-------------|
| x1 | 1 | 2 |
| x2 | 3 | 5 |
| x3 | 2 | 0 |
| x4 | 4 | 5 |



Manhattan (L_1)

| L | x1 | x2 | x3 | x4 |
|----|----|----|----|----|
| x1 | 0 | | | |
| x2 | 5 | 0 | | |
| x3 | 3 | 6 | 0 | |
| x4 | 6 | 1 | 7 | 0 |

Euclidean (L_2)

| L2 | x1 | x2 | x3 | x4 |
|----|------|-----|------|----|
| x1 | 0 | | | |
| x2 | 3.61 | 0 | | |
| x3 | 2.24 | 5.1 | 0 | |
| x4 | 4.24 | 1 | 5.39 | 0 |

Supremum

| L_∞ | x1 | x2 | x3 | x4 |
|------------|----|----|----|----|
| x1 | 0 | | | |
| x2 | 3 | 0 | | |
| x3 | 2 | 5 | 0 | |
| x4 | 3 | 1 | 5 | 0 |

Ordinal Variables

- Order is important, e.g., rank
- Can be treated like interval-scaled
 - replace x_{if} by their rank $r_{if} \in \{1, \dots, M_f\}$
 - map the range of each variable onto $[0, 1]$ by replacing i -th object in the f -th variable by

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

- compute the dissimilarity using methods for interval-scaled variables

Attributes of Mixed Type

- A database may contain all attribute types
 - Nominal, symmetric binary, asymmetric binary, numeric, ordinal
- One may use a weighted formula to combine their effects

$$d(i, j) = \frac{\sum_{f=1}^p \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p \delta_{ij}^{(f)}}$$

- f is binary or nominal:
 $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, or $d_{ij}^{(f)} = 1$ otherwise
- f is numeric: use the normalized distance
- f is ordinal
 - Compute ranks r_{if} and
 - Treat z_{if} as interval-scaled

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

**Clustering algorithm:
K-prototypes**

Cosine Similarity

- A **document** can be represented by thousands of attributes, each recording the *frequency* of a particular word (such as keywords) or phrase in the document.

| <i>Document</i> | <i>teamcoach</i> | <i>hockey</i> | <i>baseball</i> | <i>soccer</i> | <i>penalty</i> | <i>score</i> | <i>win</i> | <i>loss</i> | <i>season</i> |
|-----------------|------------------|---------------|-----------------|---------------|----------------|--------------|------------|-------------|---------------|
| Document1 | 5 | 0 | 3 | 0 | 2 | 0 | 2 | 0 | 0 |
| Document2 | 3 | 0 | 2 | 0 | 1 | 1 | 1 | 0 | 1 |
| Document3 | 0 | 7 | 0 | 2 | 1 | 0 | 3 | 0 | 0 |
| Document4 | 0 | 1 | 0 | 0 | 1 | 2 | 0 | 3 | 0 |

- Other vector objects: gene features in micro-arrays, ...
- Applications: information retrieval, biologic taxonomy, gene feature mapping, ...
- Cosine measure: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / (||d_1|| ||d_2||),$$

where \bullet indicates vector dot product, $||d||$: the length of vector d

**Clustering algorithm:
Spherical k-means**

Example: Cosine Similarity

- $\cos(d_1, d_2) = (d_1 \bullet d_2) / (||d_1|| ||d_2||)$,
where \bullet indicates vector dot product, $||d||$: the length of vector d
- Ex: Find the **similarity** between documents 1 and 2.

$$d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$$

$$d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$$


$$d_1 \bullet d_2 = 5 \cdot 3 + 0 \cdot 0 + 3 \cdot 2 + 0 \cdot 0 + 2 \cdot 1 + 0 \cdot 1 + 0 \cdot 1 + 2 \cdot 1 + 0 \cdot 0 + 0 \cdot 1 = 25$$

$$||d_1|| = (5^2 + 0^2 + 3^2 + 0^2 + 2^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2)^{0.5} = (42)^{0.5} = 6.481$$

$$||d_2|| = (3^2 + 0^2 + 2^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2)^{0.5} = (17)^{0.5} = 4.12$$

$$\cos(d_1, d_2) = 0.94$$

Evaluation and Other Practical Issues

- Evaluation of Clustering
- Similarity and Dissimilarity
- Summary 

Summary

- Evaluation of Clustering
 - Purity, NMI, RI, F-measure
- Similarity and Dissimilarity
 - Nominal attributes
 - Numerical attributes
 - Combine attributes
 - High dimensional feature vector