CS145: INTRODUCTION TO DATA MINING

11: Clustering Evaluation and Practical Issues

Instructor: Yizhou Sun

yzsun@cs.ucla.edu

November 5, 2018

Announcement

- No class next Monday (**Veterans Day**)
- Midterm: 11/14, 12-1:45pm, in-class

Learnt Clustering Methods

	Vector Data	Set Data	Sequence Data	Text Data
Classification	**Logistic Regression; Decision Tree; KNN; SVM; NN**			Naïve Bayes for Text
Clustering	**K-means; hierarchical clustering; DBSCAN; Mixture Models**			PLSA
Prediction	**Linear Regression** GLM*			
Frequent Pattern Mining		Apriori; FP growth	GSP; PrefixSpan	
Similarity Search			DTW	

Evaluation and Other Practical Issues

- Evaluation of Clustering
- Model Selection
- Summary

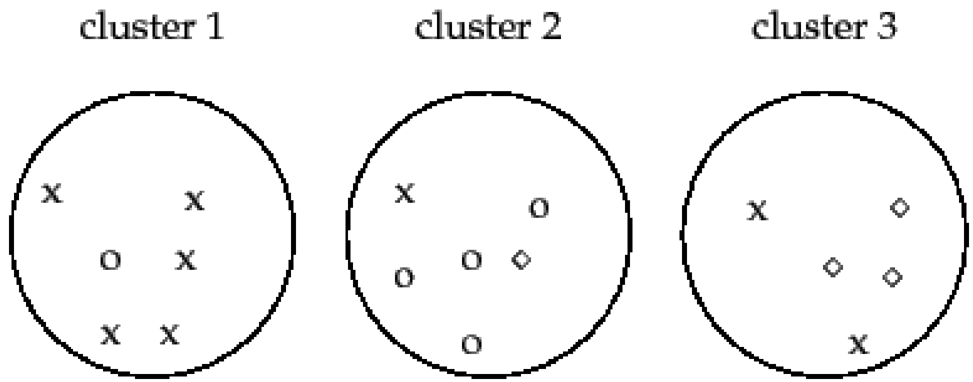
Measuring Clustering Quality

- 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 $\boldsymbol{C} = \{c_1, \ldots, c_K\}$ be the output clustering result, $\boldsymbol{\Omega} = \{\omega_1, \ldots, \omega_J\}$ be the ground truth clustering result (ground truth class)
  - c_k *and* w_k are sets of data points
  - $purity(C, \Omega) = \frac{1}{N} \sum_k \max_j |c_k \cap \omega_j|$
    - *Match each output cluster* c_k *to the best ground truth cluster* ω_j
    - *Examine the overlap of data points between the two matched clusters*
    - *Purity is the proportion of data points that are matched*

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 ⋄, 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: ×'s, ◊'s, and ○'s.

- cluster 1 vs. ×'s, cluster 2 vs. ○'s, and cluster 3 vs. ◊'s

Normalized Mutual Information

- $$NMI(C,\Omega) = \frac{I(C,\Omega)}{\boxed{\sqrt{H(C)H(\Omega)}}}$$

Denominator can be replaced by arithmetic mean, min, or max

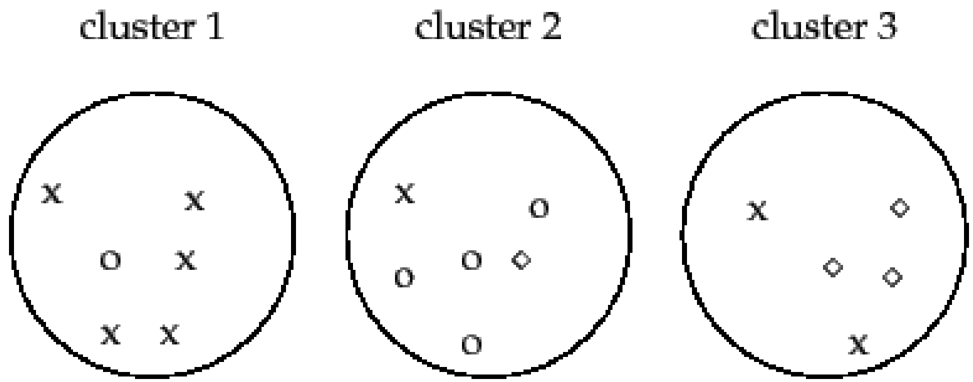
- $$I(\Omega, C) = \sum_k \sum_j P(c_k \cap \omega_j)\, log \frac{P(c_k \cap \omega_j)}{P(c_k)P(\omega_j)}$$

$$= \sum_k \sum_j \frac{|c_k \cap \omega_j|}{N}\, log \frac{N|c_k \cap \omega_j|}{|c_k| \cdot |\omega_j|}$$

- $$H(\Omega) = -\sum_j P(\omega_j) log P(\omega_j)$$

$$= -\sum_j \frac{|\omega_j|}{N}\, log\, \frac{|\omega_j|}{N}$$

Example



NMI=0.36

$|\omega_k \cap c_j|$

$|\omega_k|$

	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

$|c_j|$

Precision and Recall

- Random Index (RI) = (TP+TN)/(TP+FP+FN+TN)
- F-measure: 2Precision*Recall/(Precision+Recall)
  - Precision = TP/(TP+FP)
  - Recall = 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

Precision= ½, Recall= 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
- Model Selection
- Summary

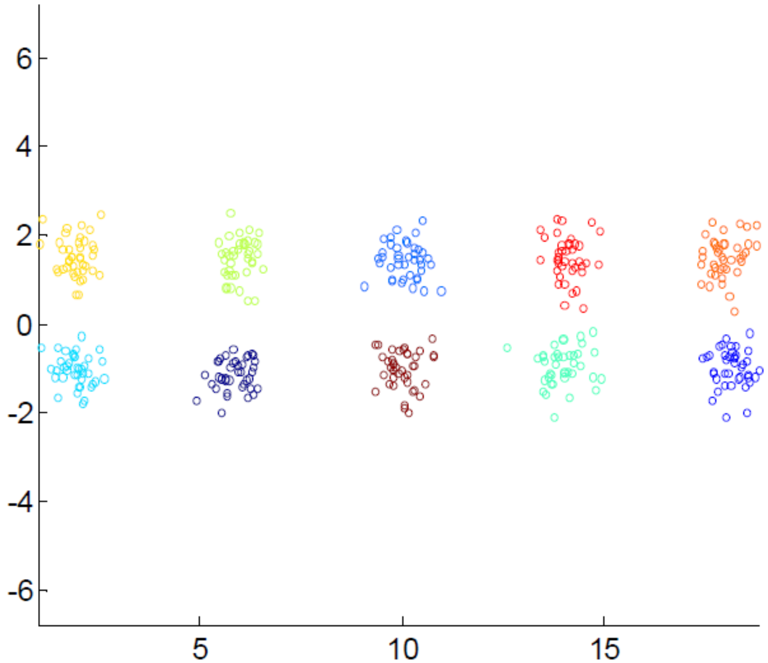
Selecting K in K-means and GMM

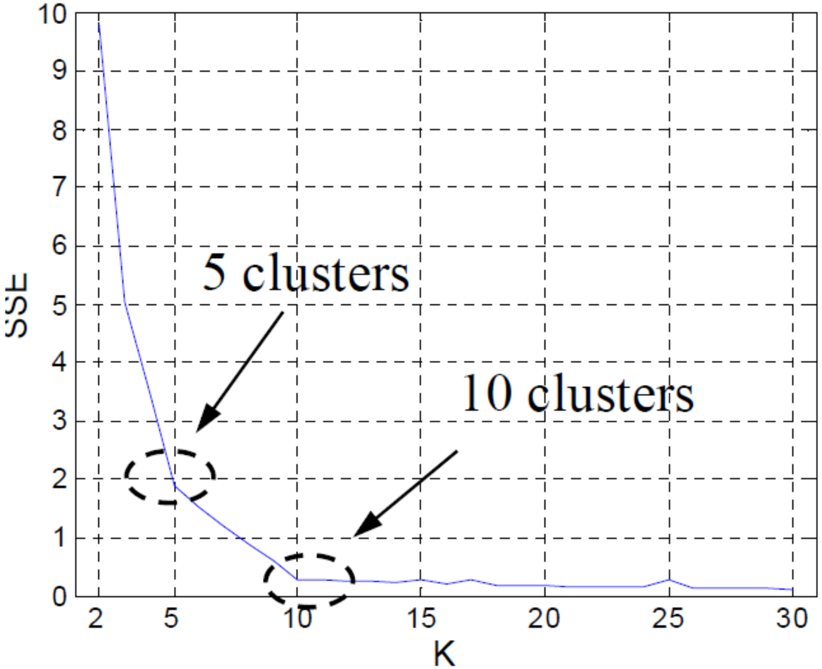
- Selecting K is a model selection problem

- Methods
  - Heuristics-based methods
  - Penalty method
  - Cross-validation

Heuristic Approaches

- For K-means, plot sum of squared error for different k
  - Bigger k always leads to smaller cost
  - Knee points suggest good candidates for k





Penalty Method: BIC

- For model-based clustering, e.g., GMM, choose k that can maximizes BIC

$$\underbrace{2l_{\mathcal{M}}(x,\theta)}_{\text{Loglikelihood of the resulting Gaussian Mixture Model}} - m_{\mathcal{M}} \log(n) \equiv \text{BIC}$$

\# of parameters to be estimated in *M*

  - Larger k increases the likelihood, but also increases the penalty term: a trade-off!

Cross-Validation Likelihood

- The likelihood of the training data will increase when increasing k
- Compute the likelihood on unseen data
  - For each possible k
  - Partition the data into training and test
  - Learn the GMM related parameters on training dataset and compute the log-likelihood on test dataset
  - Repeat this multiple times to get an average value
  - Select k that maximizes the average log-likelihood on test dataset

Evaluation and Other Practical Issues

- Evaluation of Clustering
- Model Selection
- Summary

Summary

- Evaluation of Clustering
  - Purity, NMI, F-measure
- Model selection
  - How to select k for k-means and GMM