PURDUE UNIVERSITY

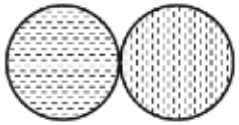
CS37300:
Data Mining and Machine Learning

Evaluating Clustering

Nov 17 2023

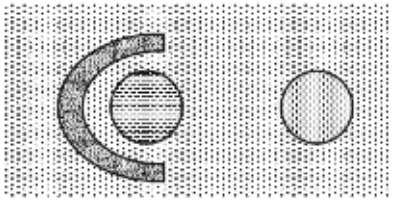
What makes a "good" cluster?

(a) Well-separated clusters. Each point is closer to all of the points in its cluster than to any point in another cluster.

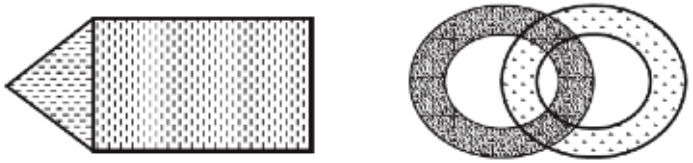


(b) Center-based clusters. Each point is closer to the center of its cluster than to the center of any other cluster.

(c) Contiguity-based clusters. Each point is closer to at least one point in its cluster than to any point in another cluster.



(d) Density-based clusters. Clusters are regions of high density separated by regions of low density.



(e) Conceptual clusters. Points in a cluster share some general property that derives from the entire set of points. (Points in the intersection of the circles belong to both.)

Figure 8.2. Different types of clusters as illustrated by sets of two-dimensional points.

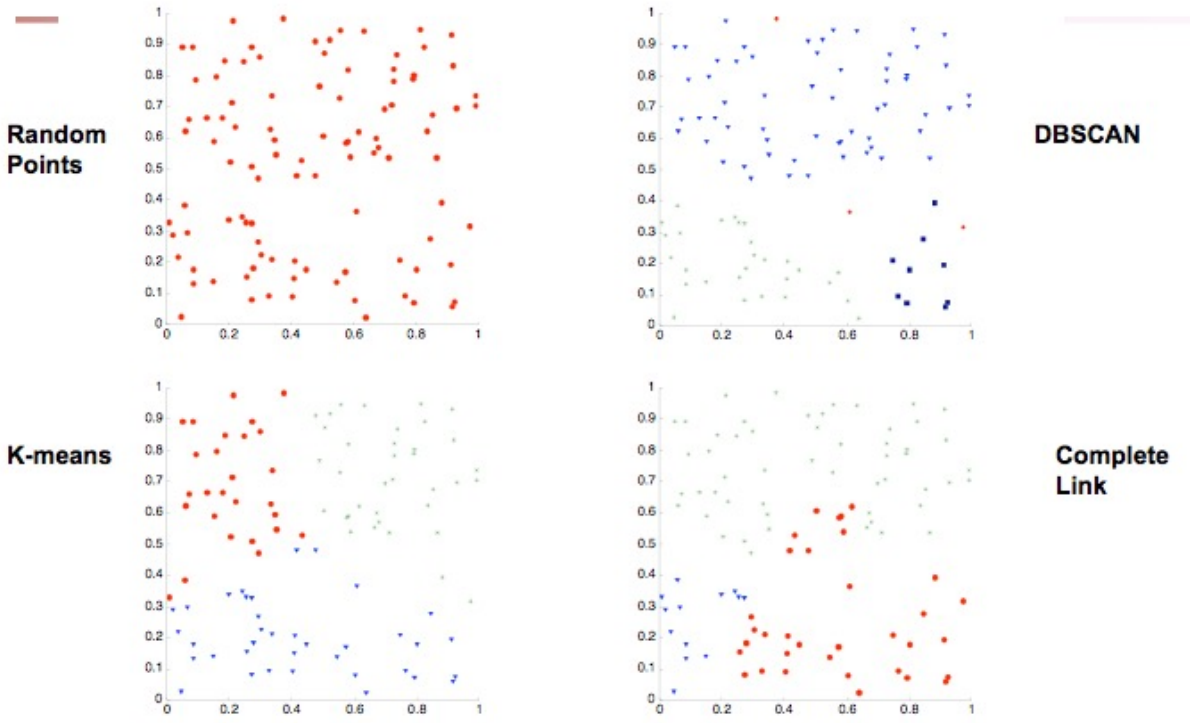
PURDUE UNIVERSITY

Cluster validity

- For **prediction** tasks there are a variety of external evaluation metrics
  - Accuracy, squared loss, area under ROC, etc.
- For **cluster analysis** the external evaluation should evaluate the "goodness" of the resulting clusters
- Why do we want external validation?
  - To avoid finding patterns in noise
  - To compare clustering algorithms
  - To compare two sets of clusters

PURDUE UNIVERSITY

Random data: clustering still returns results



PURDUE UNIVERSITY

Evaluation approaches

- Determine the clustering tendency of the data
- Evaluate the clusters using known class labels
  - Evaluate how well the clusters “fit” the data
- Determine which of two different clustering results is better
- Determine the “correct” number of clusters

PURDUE UNIVERSITY

Clustering tendency

- Evaluate whether a dataset has clusters before clustering
- Most common approach (for low-dimensional Euclidean data)
  - Use a statistical test for spatial randomness
  - Hopkins statistic: sample p=20 points from dataset, generate p=20 random points in same space

w_i: distance from random point to NN in data

u_i: distance from sample point to NN in data

$$H = \frac{\sum_{i=1}^{p} w_i}{\sum_{i=1}^{p} u_i + \sum_{i=1}^{p} w_i}$$

  - Values near 0.5 indicate random data, 1.0 indicates highly clustered, and 0.0 indicates uniformly distributed

PURDUE UNIVERSITY

Types of clustering evaluation measures

- Supervised
  - Measures the extent to which clusters match external class label values
- Unsupervised
  - Measures goodness of fit without class labels

PURDUE
UNIVERSITY®

Descriptive Modeling: Supervised Evaluation

PURDUE UNIVERSITY

Class-label evaluation

- If you have class labels why cluster?
  - Usually labels come from small hand-labeled dataset for evaluation
  - But have remaining large dataset to cluster automatically
  - May want to assess how close clusterings correspond to classes but still allow for more variation in the clusters

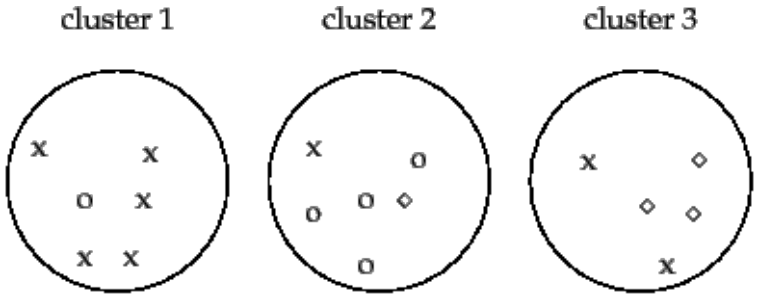
PURDUE UNIVERSITY

Classification-oriented

- **Purity**: another measure of the extent to which cluster/groups (G) contain objects of a particular class (C)
  - The purity of each cluster *i* in G is determined by the number of examples of class *j* *(N_j)* inside each cluster:

$$purity(C, G) = \frac{1}{N} \sum_{i=1}^{K} \max_j N_j \in G_i$$

- High purity is easier to achieve when the number of clusters is large



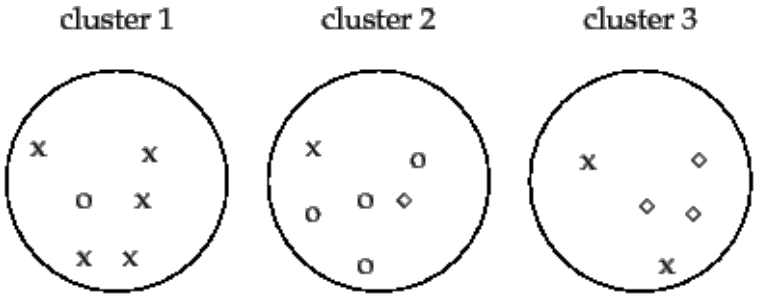
$$= \frac{1}{17}[5 + 4 + 3] = 0.7$$

PURDUE UNIVERSITY

Classification-oriented

- **Entropy**: the degree to which each cluster (G) consists of objects of a single class (C)
  - For each cluster *i* compute the probability of class *j* (within the cluster)

$$entropy(C,G) = \sum_{i=1}^{K} - \sum_{j=1}^{C} p_{ij} log(p_{ij})$$



$$= -\left[\frac{1}{6}log(\frac{1}{6}) + \frac{5}{6}log(\frac{5}{6})\right] - \left[\frac{1}{6}log(\frac{1}{6}) + \frac{1}{6}log(\frac{1}{6}) + \frac{4}{6}log(\frac{4}{6})\right] - \left[\frac{2}{5}log(\frac{2}{5}) + \frac{3}{5}log(\frac{3}{5})\right] = 1.99$$

PURDUE UNIVERSITY

Classification-oriented

- **Normalized mutual information gain:**
  - Measures the amount of information by which our knowledge about the classes (C) increases when we are told what the clusters (G) are

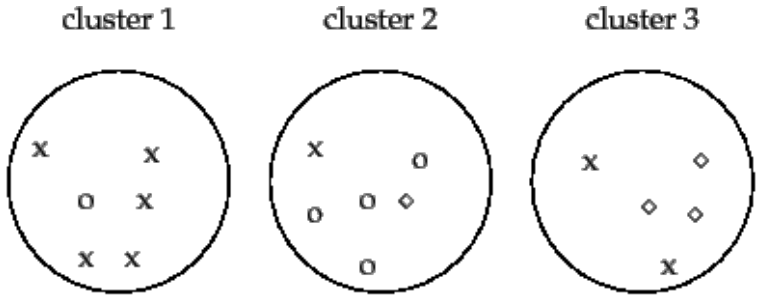
$$NMI(C,G) = \frac{I(C,G)}{H(C)+H(G)}$$

$$= \frac{\sum_c \sum_g p(c,g) log \frac{p(c,g)}{p(c)p(g)}.}{-\sum_c p(c) log\, p(c) - \sum_g p(g) log\, p(g)}$$

  - NMI score is between 0 (min) and 1 (max).
  - Denominator (normalization) adjusts for problem that entropy tends to increase with the number of clusters

PURDUE UNIVERSITY

NMI example



$$H(C) = -\left[\frac{8}{17}log(\frac{8}{17}) + \frac{5}{17}log(\frac{5}{17}) + \frac{4}{17}log(\frac{4}{17})\right] = 1.055$$

$$H(G) = -\left[\frac{6}{17}log(\frac{6}{17}) + \frac{6}{17}log(\frac{6}{17}) + \frac{5}{17}log(\frac{5}{17})\right] = 1.095$$

$$I(C,G) = \left[\frac{5}{6}log(\frac{5}{6}\frac{17}{8}\frac{17}{6}) + \frac{1}{6}log(\frac{1}{6}\frac{17}{5}\frac{17}{6}) + \frac{0}{6}log(\frac{0}{6}\frac{17}{4}\frac{17}{6}) + ...\right]$$

PURDUE UNIVERSITY

Similarity-oriented

- Based on the premise that any pair of objects in the same cluster should have the same class and vice versa
- Construct the “ideal” similarity matrix based on cluster membership
  - Entry i,j is 1 if i and j are in the same cluster, 0 otherwise
- Construct the “ideal” similarity matrix based on class values
  - Entry i,j is 1 if i and j are in the same class, 0 otherwise
- Use measure that compares the two ideal similarity matrices

PURDUE UNIVERSITY

Measures to compare same-class / same-cluster matrices

- Correlation between two ideal matrices
- Measures of binary similarity between two ideal matrices
  - f_{00} = # pairs of objects having diff class and diff cluster
  - f_{01} = # pairs of objects having diff class and same cluster
  - f_{10} = # pairs of objects having same class and diff cluster
  - f_{11} = # pairs of objects having same class and same cluster

$$Rand = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

$$Jaccard = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

PURDUE
UNIVERSITY®

Descriptive Modeling: Unsupervised Evaluation

PURDUE UNIVERSITY

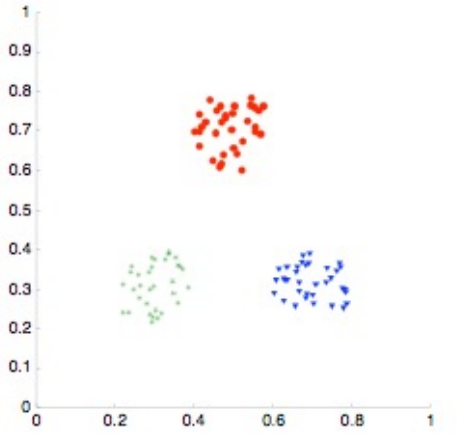
Visual inspection

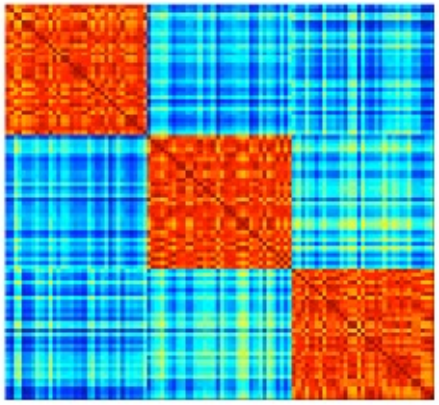
- Order the proximity matrix with respect to cluster labels
- Inspect visually
- Good clusterings exhibit clear block pattern

PURDUE UNIVERSITY

Example 1: good clustering

Proximity matrix reordered
to reflect cluster assignments

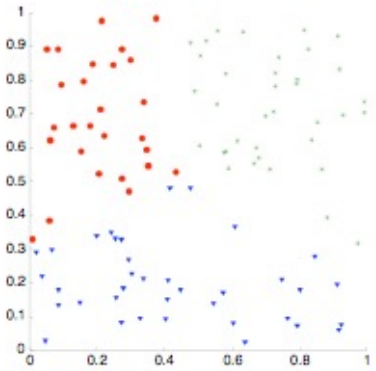


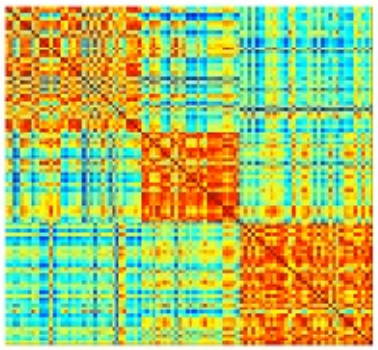


PURDUE UNIVERSITY

Example II: poor clustering

Proximity matrix reordered
to reflect cluster assignments





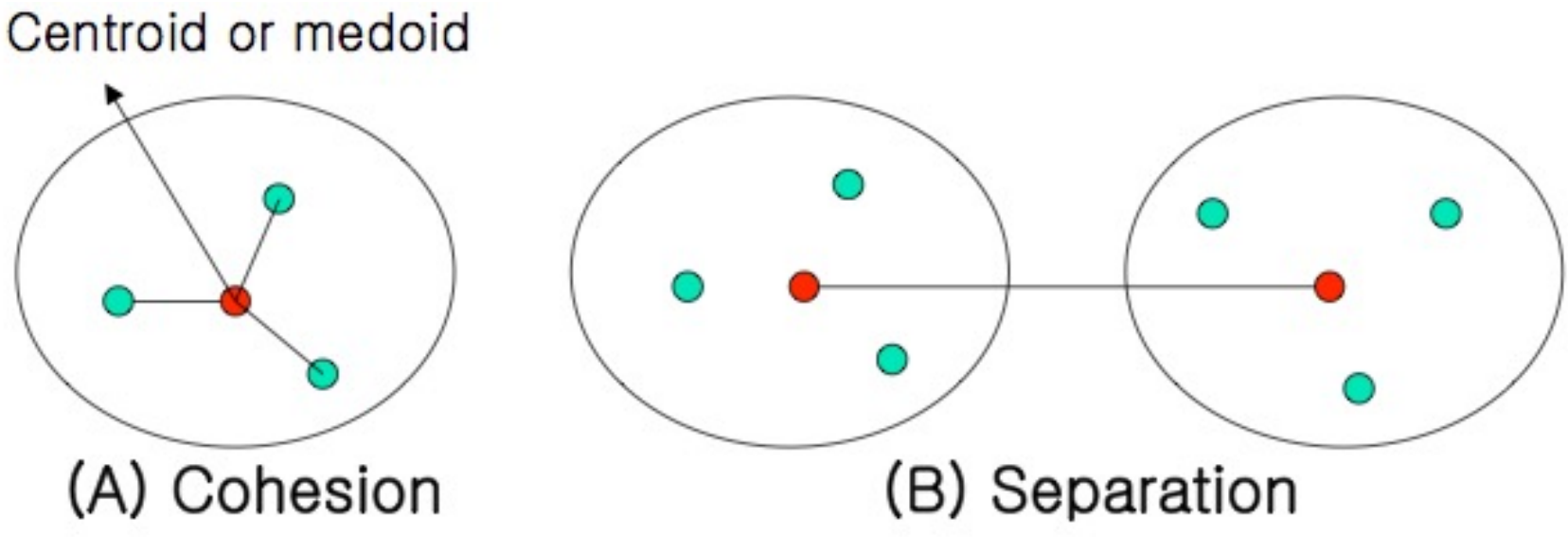
PURDUE UNIVERSITY

Correlation

- Construct the “ideal” similarity matrix based on cluster membership
  - Entry i,j is 1 if i and j are in the same cluster, 0 otherwise
- Compute the correlation between the initial similarity matrix and the “ideal” similarity matrix that corresponds to the cluster results
  - High correlation indicates that points in same cluster are close to each other

PURDUE UNIVERSITY

Cohesion and separation



PURDUE UNIVERSITY

Cohesion

- *Measures how closely related the objects are within each cluster*
- Within cluster sum of squared errors (SSE)
  - For each point, the error is the distance to the centroid
- Within cluster pairwise weighting
  - Sum distance between all pairs of points in same cluster

PURDUE UNIVERSITY

Separation

- *Measures how distinct a cluster is from the other clusters*
- Between cluster SSE (for cluster C)
  - For each cluster C', the error is the distance from the centroid c to the other centroid c'
  - The error is multiplied by the cluster size |C'|
- Between cluster pairwise weighting
  - Sum distance between all pairs of points in different clusters

PURDUE UNIVERSITY

Cohesion and separation

- The sum of the between cluster SSE and within cluster SSE is equal to the total sum of squared error (distance of each point to overall mean)
- Thus minimizing cohesion is equivalent to maximizing separation

PURDUE UNIVERSITY

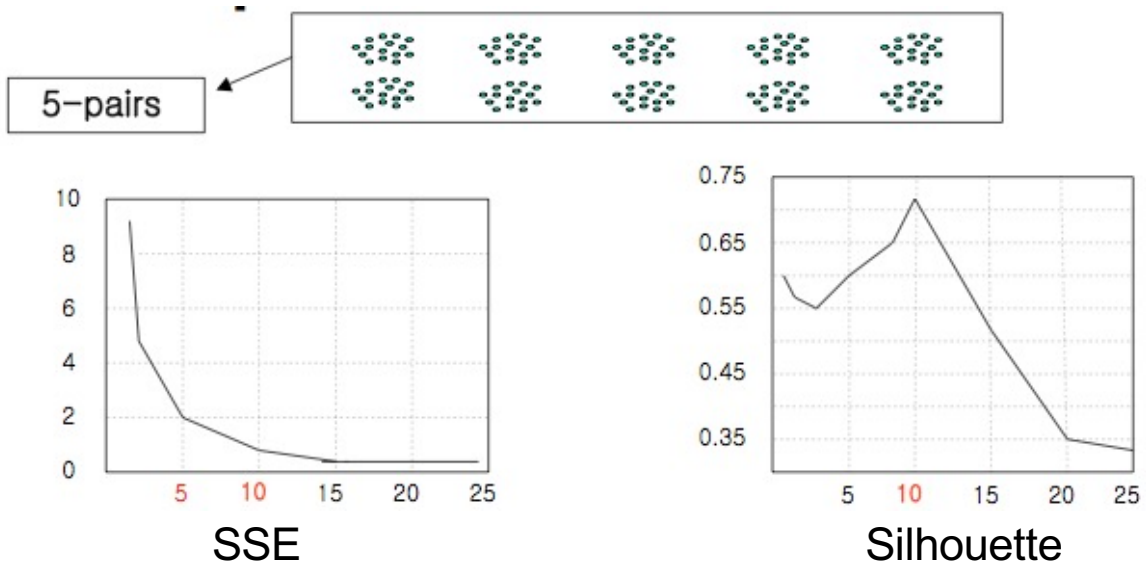
Silhouette coefficient

- Combines both cohesion and separation
- For an individual point i:
  - A = average distance of i to points in same cluster
  - B = average distance of i to points in other clusters
  - S = (B-A) / max(A,B)
- Can calculate average S for a cluster or clustering
  - Closer to 1 is better

PURDUE UNIVERSITY

Determining k (revisited)

- Approach: evaluate over a range of k, look for peak, dip, or knee in evaluation measure



SSE

Silhouette

PURDUE UNIVERSITY

Assessing significance

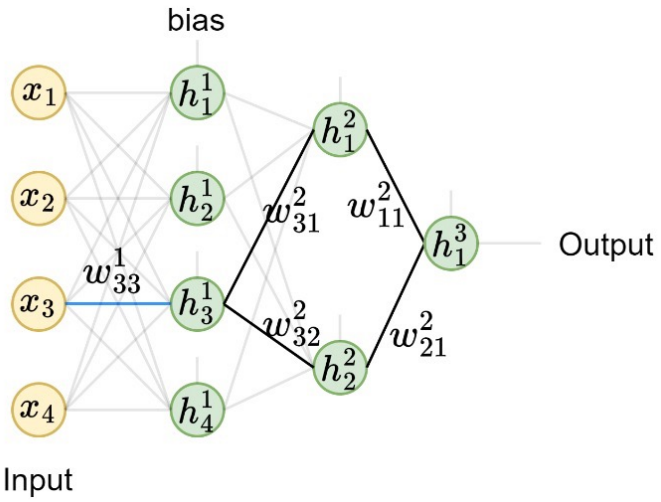
- How do we know a score is “good”?
- How do we know that a difference between two algorithms is significant?
- This is the problem we also have for predictive models
- Need a sampling distribution to compare to
  - Can generate random data in same space and compute empirical sampling distribution
  - Can partition data to get multiple folds for evaluation

PURDUE UNIVERSITY

Take-Home Quiz
(Due 9:00am on Nov 20 on Gradescope)

- Given the neural network on the right, calculate the partial derivative of the loss function w.r.t. w_{33}^1 (i.e., $\frac{\partial L}{\partial w_{33}^1}$) via backpropagation. This neural network uses *tanh* as the activation function, and the loss function is given below:

$$L = -\frac{1}{N}\sum_{i=1}^{N} y_i log(h_1^3(x_i)) + (1-y_i)log(1-h_1^3(x_i))$$



- Your solution should only involve outputs of neurons (i.e., h_i^j), weights (i.e., w_{ab}^j), and x_3
- Hints:
  - $\frac{\partial \tanh(x)}{\partial x} = 1 - \tanh(x)^2$
  - Look at the example in the backpropagation slides

The definition of four hidden units involved in this backpropagation

$$h_3^1(x) = tanh((w_{\{0,1,2,3,4\}3}^1)^\top (1, x_1, x_2, x_3, x_4))$$

$$h_1^2(x) = tanh((w_{\{0,1,2,3,4\}1}^2)^\top (1, h_1^1(x), h_2^1(x), h_3^1(x), h_4^1(x)))$$

$$h_2^2(x) = tanh((w_{\{0,1,2,3,4\}2}^2)^\top (1, h_1^1(x), h_2^1(x), h_3^1(x), h_4^1(x)))$$

$$h_1^3(x) = tanh((w_{\{0,1,2\}1}^3)^\top (1, h_1^2(x), h_2^2(x)))$$