Data mining & Machine Learning

CS 373
Purdue University

Dan Goldwasser
dgoldwas@purdue.edu

Today's Lecture

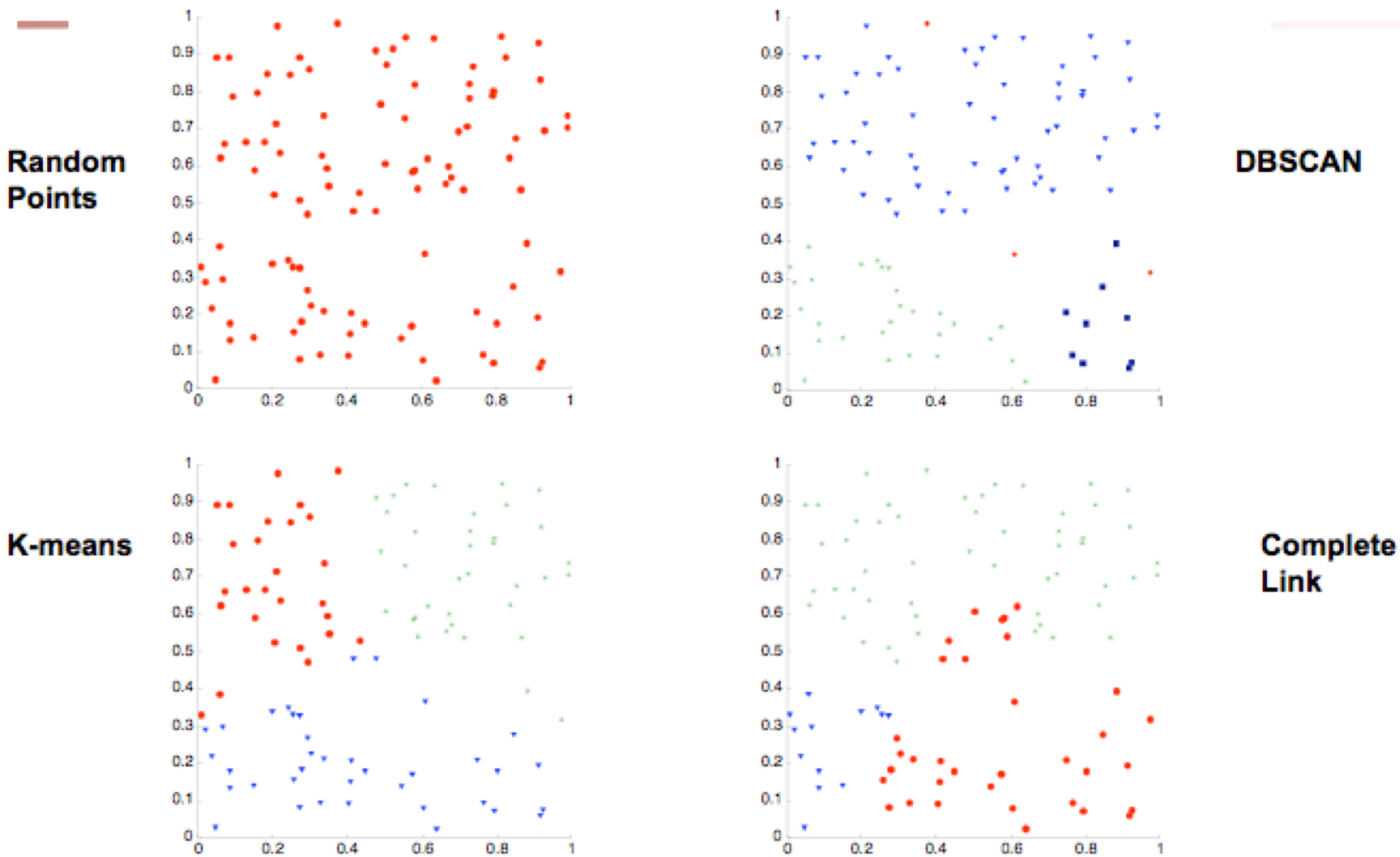
Descriptive modeling:

evaluation

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

Random data



Cluster Evaluation

- **Evaluating the quality of the obtained clusters is very difficult!**
- By definition, unsupervised learning entails a “fuzzy” evaluation criterion
  - Since there is no supervision, there is no clear goal to optimize for
- **Is all hope lost?**
- Our next step would be to find ways to formalize these intuitions

Can you think of reasonable rules-of-thumb to separate good clusters from bad?

Evaluation approaches

- **Determine the clustering tendency of the data**
  - *Are there good clusters in the data? (regardless of specific ones you find)*
- **Evaluate the clusters using known class labels**
  - Match between clusters and annotated data (meaningful if the labels and clusters should be correlated)
- **Evaluate how well the clusters “fit” the data**
- **Determine which of two different clustering results is better**
- **Determine the “correct” number of clusters**

Evaluation measures

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

Unsupervised

Clustering tendency

- **Evaluate whether a dataset has clusters without clustering**
- Most common approach (for low-dimensional Euclidean data)
  - Use a statistical test for spatial randomness
  - **Hopkins statistic**: *sample 20 points from dataset, generate 20 random points in same space*

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

u_i: distance from random point to NN in data

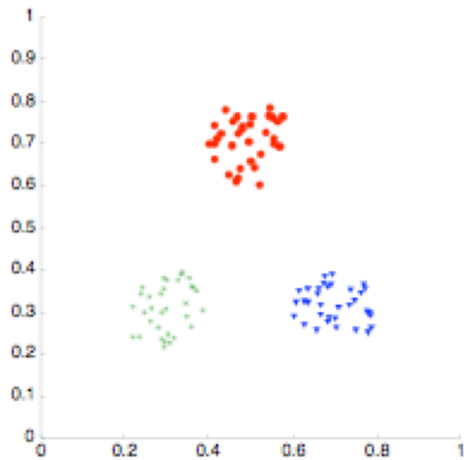
w_i: distance from sample point to NN in data

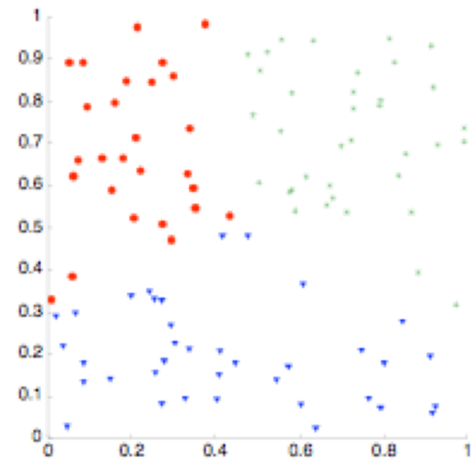
  - Values near 0.5 indicate random data 0 indicates highly clustered, and 1.0 indicates uniformly distributed

Correlation

- **Construct an "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

Example





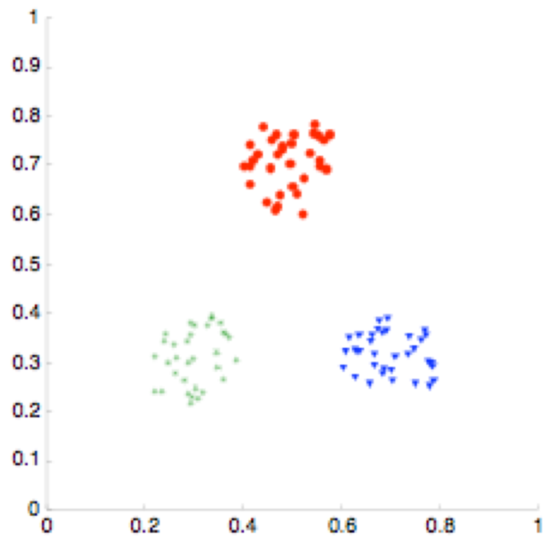
Corr = -0.9235

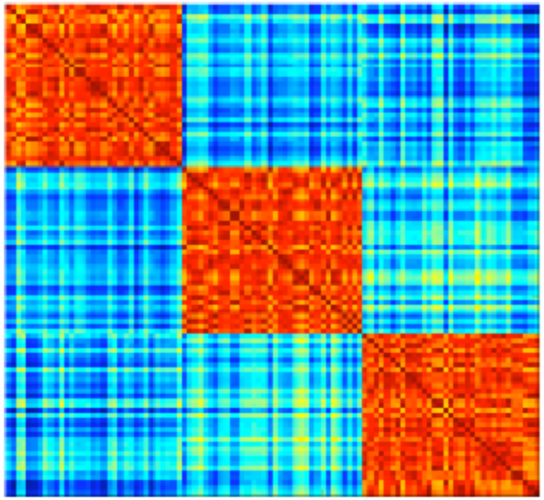
Corr = -0.5810

Visual inspection

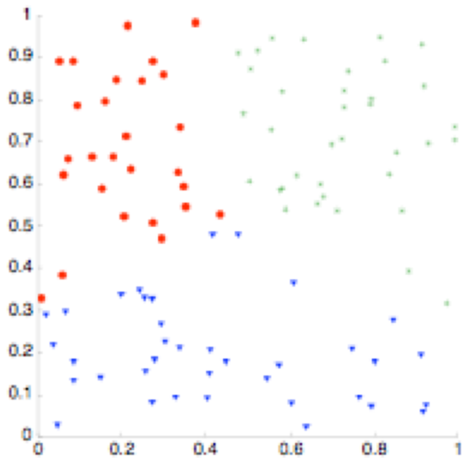
- Order the proximity matrix with respect to cluster labels
- Inspect visually
- **Good clustering exhibit clear block pattern**

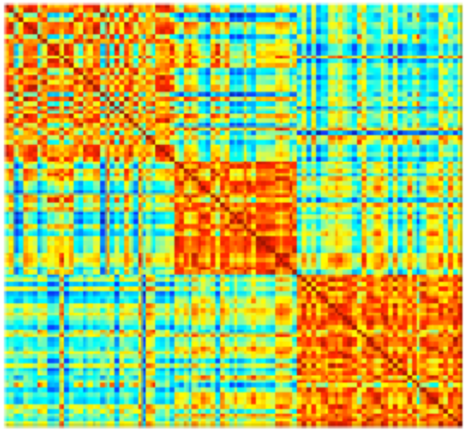
Example 1





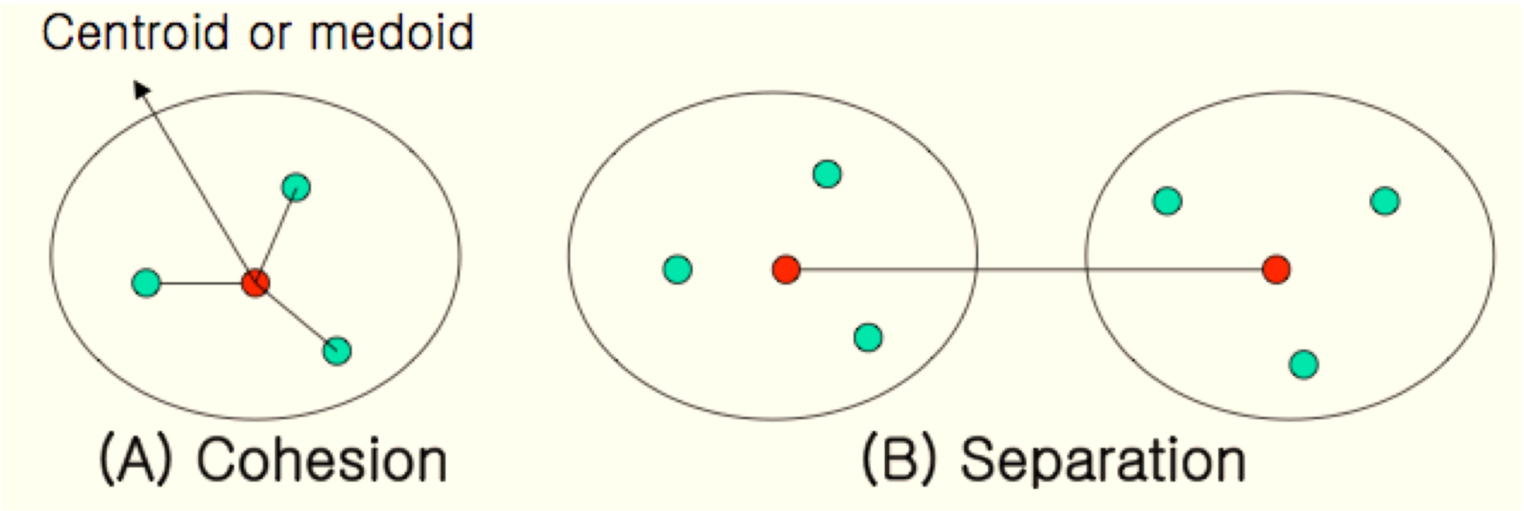
Example II





Cohesion and separation

Cluster Separation: Measure how distinct or well- separated a cluster is from other clusters



Cluster Cohesion: Measures how closely related are objects in a cluster

Cohesion

- **Measures how closely related the objects are within each cluster**

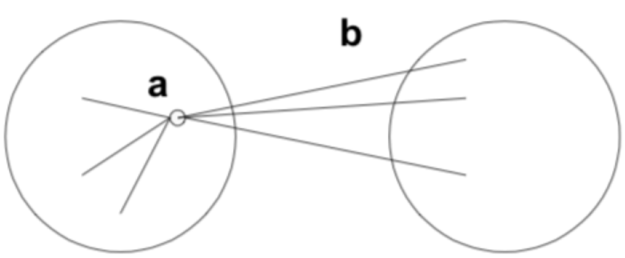
- Within cluster sum of squared errors (SSE)
  - *For each point, the error is the distance to the centroid*
- Within cluster pairwise weighting
  - *Sum distance between all pairs of points in same cluster*

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*

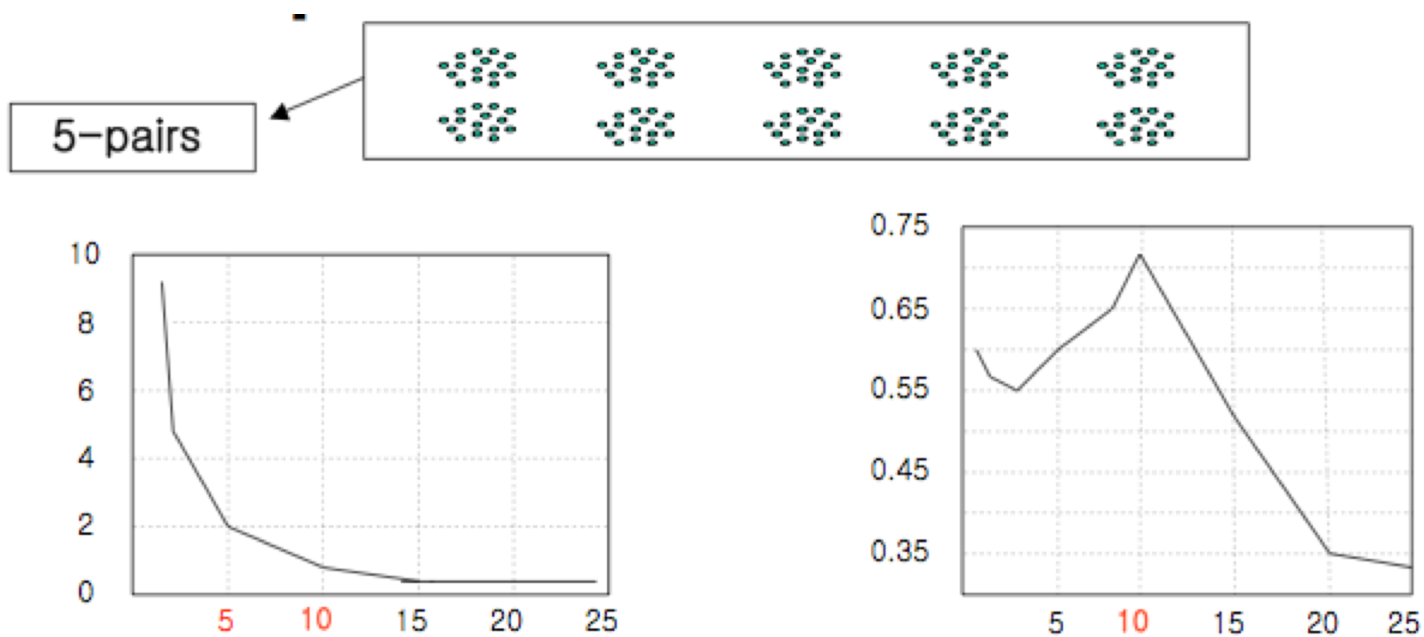
Silhouette coefficient

- **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



Determining k

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



SSE

Silhouette

Supervised

Class-label evaluation

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

Classification-oriented

- **Purity**: *a measure of the extent to which a cluster contains objects of a single class*
  - The purity of cluster i is $p_i = max\ p_{ij}$

$p_{ij} = m_{ij} / m_i$

$$purity = \sum_{i=1}^{K} \frac{N_i}{N} p_i$$

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

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

$$e_i = -\sum_{j=1}^{C} p_{ij} log\ p_{ij}$$

Classification-oriented

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

Mutual information between C and G

Entropy of G

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

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

classes

clusters

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

Classification-oriented

- **Precision**
    - The fraction of a cluster that consists of objects of a specified class
- **Recall**
    - The extent to which a cluster contains all objects of a specified class
- **Accuracy**
    - Why is it hard to measure the accuracy of a clustering if you know class labels?

Similarity-oriented

- Based on 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
- Compare the two ideal similarity matrices

Approaches

- 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_{1} = # pairs of objects having same class and same cluster

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