CS57300
PURDUE UNIVERSITY

NOVEMBER 17, 2021

DATA MINING

DESCRIPTIVE MODELING: EVALUATION

SUPERVISED EVALUATION

- **Purity**: a measure of the degree to which a cluster/group (G_i) contains objects of c particular class (C_j)
- **Entropy**: the degree to which each cluster (G) consists of objects of a single class
- **Normalized mutual information gain:** Measures the amount of information by w our knowledge about the classes (C) increases when the clusters (G) are identified

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
- Use measure that compares the two ideal similarity matrices

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

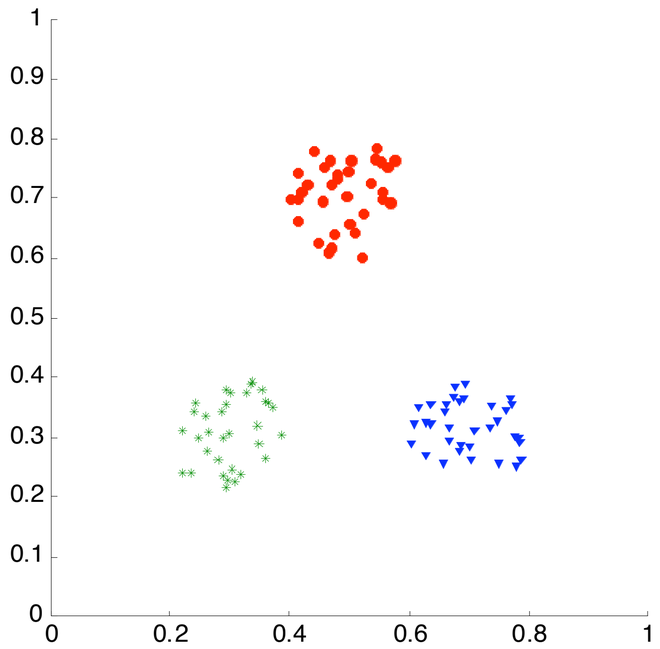
$$Rand = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}} \qquad Jaccard = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

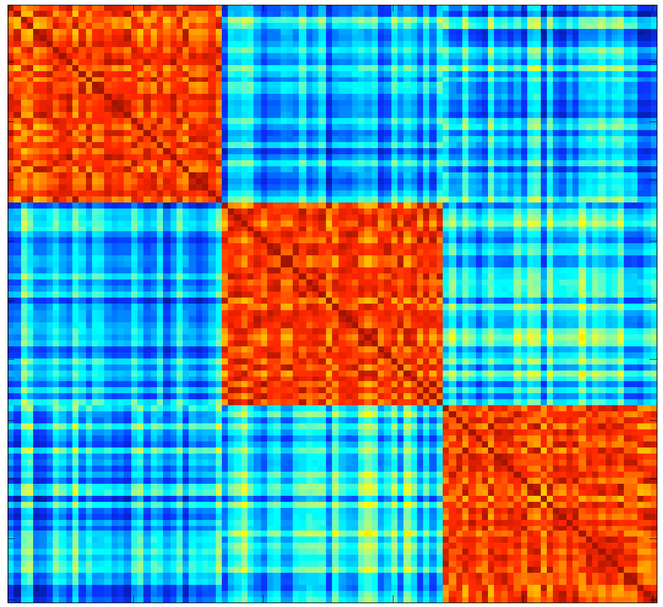
UNSUPERVISED: VISUAL INSPECTION

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

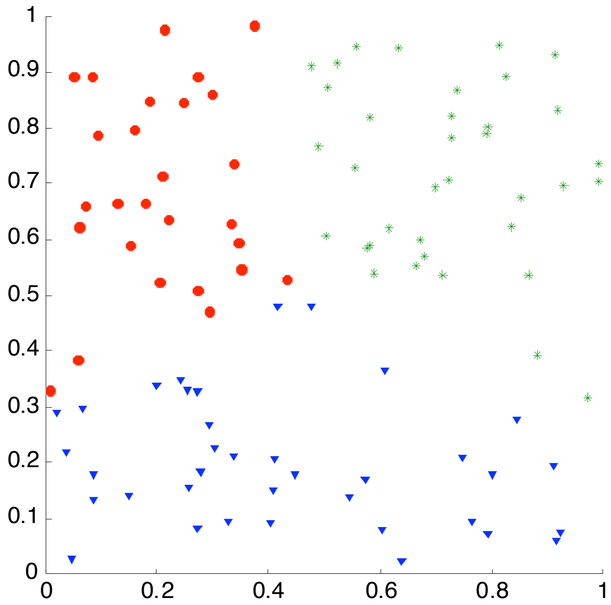
EXAMPLE 1: GOOD CLUSTERING

Proximity matrix reordered
to reflect cluster assignments

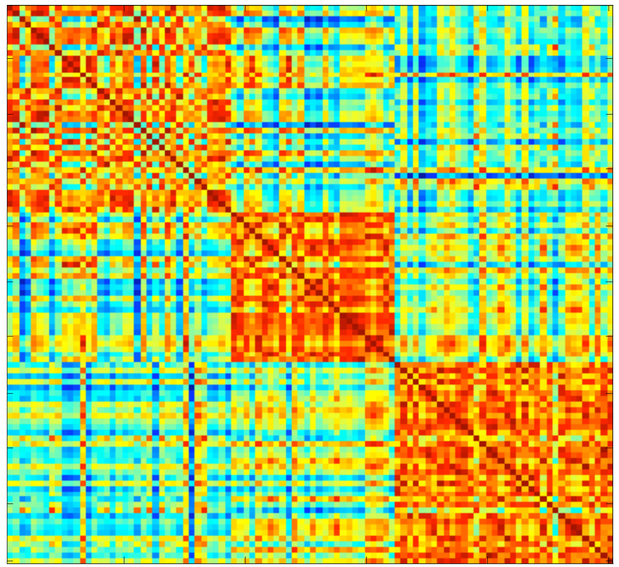




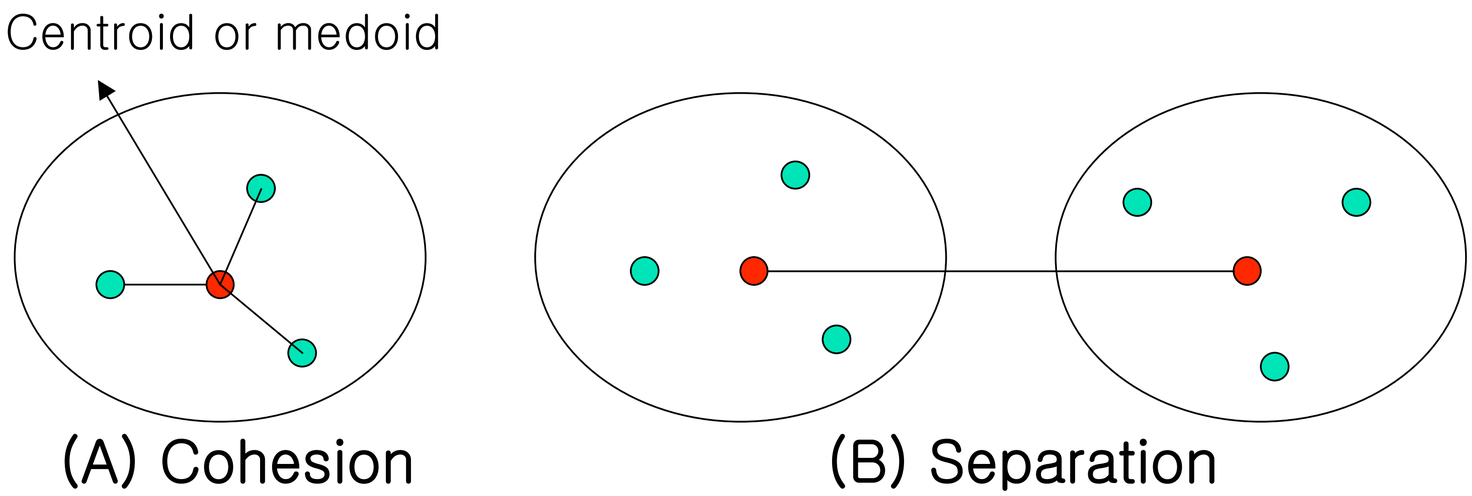
EXAMPLE II: POOR CLUSTERING



Proximity matrix reordered
to reflect cluster assignments



COHESION AND SEPARATION



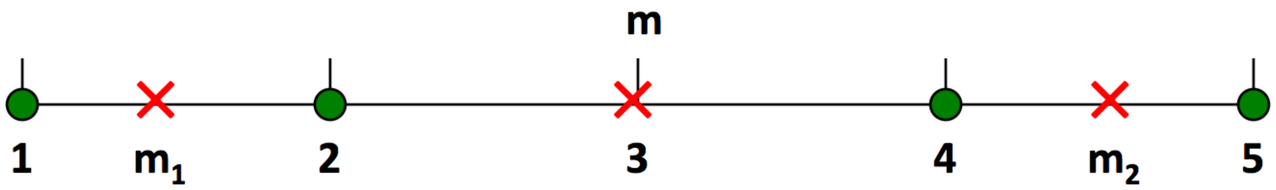
COHESION AND SEPARATION

- Cohesion: Measures how closely related the objects are within each cluster
- Separation: Measures how distinct a cluster is from the other clusters

COHESION AND SEPARATION: EXAMPLE

- **Cohesion: Within cluster sum of squared errors (WSS)**
  - For each point, the error is the distance to the centroid $WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2$
- **Separation: Between cluster sum of squared errors (BSS)**
  - For each cluster C′, the error is the distance from its centroid c′ to the centroid of the entire dataset
  - The error is multiplied by the cluster size |C′| $BSS = \sum_i |C_i| (m - m_i)^2$

COHESION AND SEPARATION: EXAMPLE



K=1 :

$$WSS = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$
$$BSS = 4 \times (3-3)^2 = 0$$
$$Total = 10 + 0 = 10$$

K=2 :

$$WSS = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$
$$BSS = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$$
$$Total = 1 + 9 = 10$$

K=4:

$$WSS = (1-1)^2 + (2-2)^2 + (4-4)^2 + (5-5)^2 = 0$$
$$BSS = 1 \times (1-3)^2 + 1 \times (2-3)^2 + 1 \times (4-3)^2 + 1 \times (5-3)^2 = 10$$
$$Total = 0 + 10 = 10$$

- WSS + BSS is a constant (squared distance of each point to centroid of the entire dataset)
- Minimize WSS is maximize BSS

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

HOW TO CHOOSE K?

- Choose k to maximize likelihood/minimize WSS?
- As K increases, likelihood is increasing and WSS is decreasing
- Thus more complex models will always improve likelihood / decrease WSS
- How to compare models with different complexities?

MODEL SELECTION SCORING FUNCTIONS

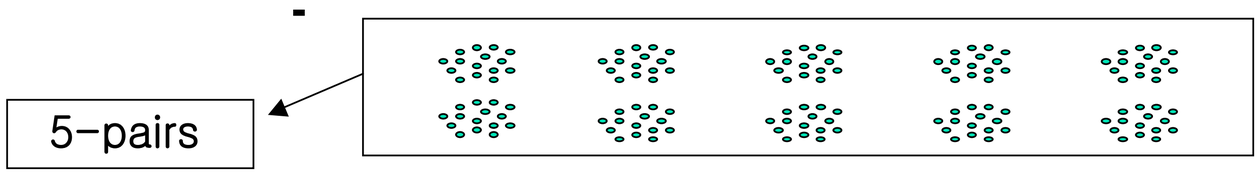
- **Goal 1**: *Describe* data as precisely as possible
- **Goal 2**: *Generalize* to new data
  - Goodness of fit is part of the evaluation, but since the data is not the entire population, we want to learn a model that will generalize to other new data instances
- Thus, want to strike a balance between between how well the model fits and the data and the simplicity of the model

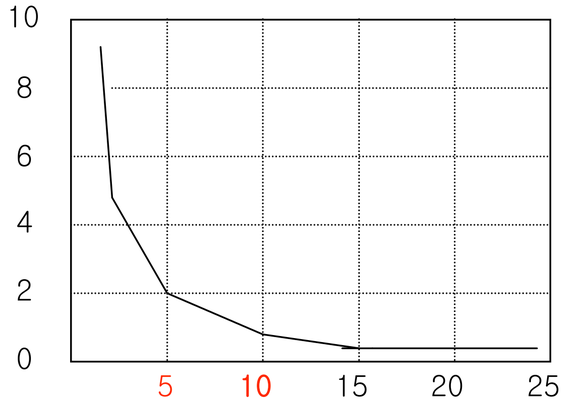
PENALIZED SCORE FUNCTIONS

- Penalized score functions include a term that reflects how well the model fits the data and another (penalty) term to value the simplicity of the model
- Score(θ,M) = error(M) + penalty(M)
  - Penalty may depend on the number of parameters in the model (p) and the number of data points (n)
  - Error is generally based on likelihood of the data given the model (L)
- AIC (Akaike information criterion): $\text{Score}_{AIC} = -2 \log L + 2p$
- BIC (Bayesian information criterion): $\text{Score}_{BIC} = -2 \log L + p \log n$

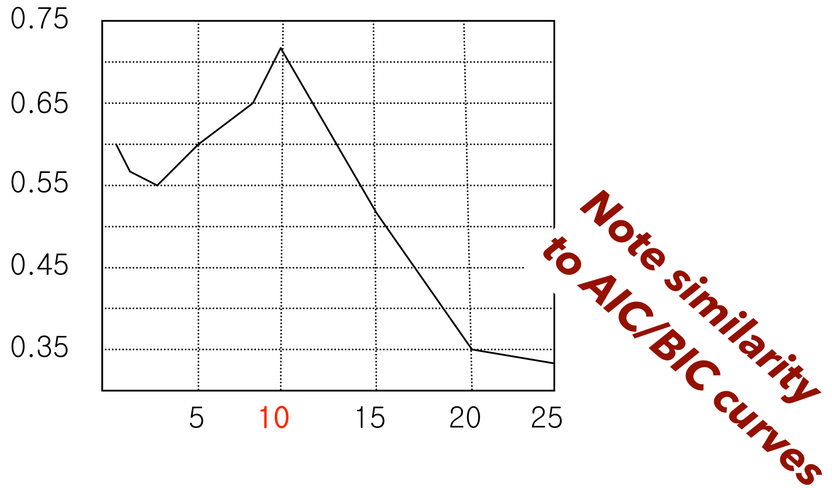
DETERMINING K

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





WSS



Silhouette

PATTERN MINING

DATA MINING COMPONENTS

- Task specification: **Pattern discovery**
- Knowledge representation
- Learning technique
- Evaluation

PATTERN DISCOVERY

- Models describe entire dataset (or large part of it)
- Pattern characterizes local aspects of data
- Pattern: predicate/statement that returns “true” for the instances in the data where the pattern occurs and “false” otherwise

PATTERN IN TABULAR DATA

- Primitive pattern: subset of all possible observations over variables $X_1,\ldots,X_p$
  - If X_k is categorical then $X_k=c$ is a primitive pattern
  - If X_k is ordinal then $X_k \leq c$ is a primitive pattern
- Start from primitive patterns and combine using logical connectives such as AND and OR
  - age<40 AND income<100,000
  - chips=1 AND (beer=1 OR soda=1)

PATTERN DISCOVERY TASK

- Find all “interesting” patterns in the data
  - Find a pattern that is frequently true
  - Find associative property between patterns

EXAMPLES

- Supermarket transaction database
  - 10% of the customers buy wine and cheese
- Telecommunications alarms database
  - If alarms A and B occur within 30 seconds of each other then alarm C occurs within 60 seconds with p=0.5
- Web log dataset
  - If a person visits the CNN website, there is a 60% chance the person will visit the ABC News website in the same month

DATA MINING COMPONENTS

- Task specification
- **Knowledge representation**
- Learning technique
- Evaluation

RULE

- A rule is an expression of the form $\theta \rightarrow \varphi$
- A statement about the co-occurrence of events/patterns
- **Support** (aka frequency)
  - $s(\theta \rightarrow \varphi) = fr(\theta \wedge \varphi) / N$
  - Proportion of N items with antecedent θ and consequent φ
- **Confidence** (aka accuracy)
  - $c(\theta \rightarrow \varphi) = p(\varphi \mid \theta) = fr(\theta \wedge \varphi) / fr(\theta)$
  - Proportion of items which have antecedent θ that also have consequent φ

ASSOCIATION RULES

- Find all rules of the form $\theta \rightarrow \varphi$ that satisfy the following constraints:
  - Support of the rule is greater than threshold *s*
  - Confidence of the rule is greater than threshold *c*

ASSOCIATION RULE EXAMPLE

- Support threshold: 30%, confidence threshold: 70%
- Flour –> Eggs
- Eggs –> Milk
- Milk –> Eggs
- Flour –> Milk
- Eggs, Flour –> Milk
- Flour, Milk –> Eggs

Transaction ID	beer	eggs	flour	milk
1	0	1	1	1
2	1	1	0	0
3	0	1	0	1
4	0	1	1	1
5	0	0	0	1