CS57300
PURDUE UNIVERSITY

APRIL 9, 2019

DATA MINING

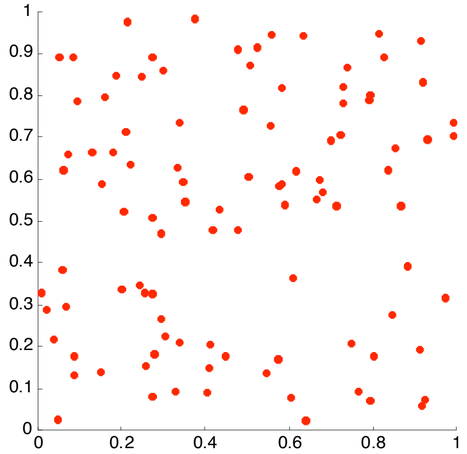
DESCRIPTIVE MODELING: EVALUATION

CLUSTER VALIDITY

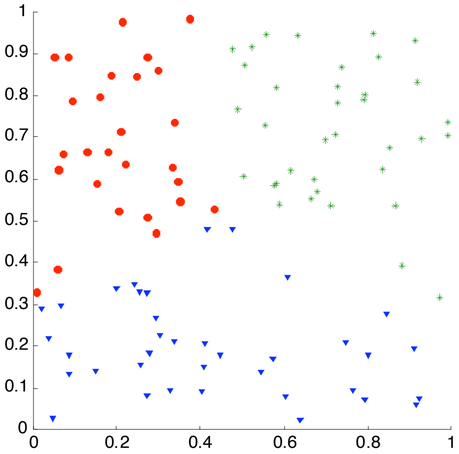
- For prediction tasks there are a variety of external evaluation metrics
  - Accuracy, precision, recall, area under ROC, etc.
- For cluster analysis the external evaluation should evaluate the “goodness” of the resulting clusters
- Why do we want external validation?
  - To avoid finding patterns in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

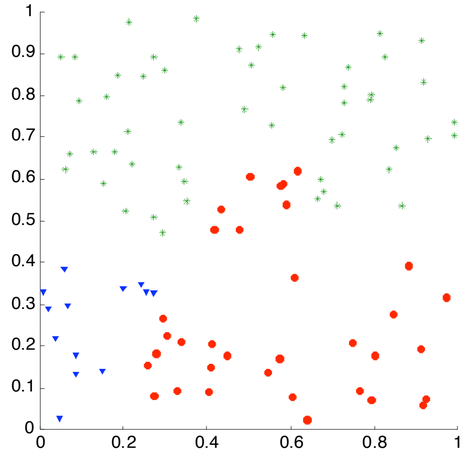
RANDOM DATA: CLUSTERING STILL RETURNS RESULTS

Random Points



K-means





Complete Link

EVALUATION APPROACHES

- Determine the clustering tendency of the data
- Evaluate the quality of clustering results
  - 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

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* points from dataset, generate *p* random points in same space

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

u_i: distance from random point to NN in data

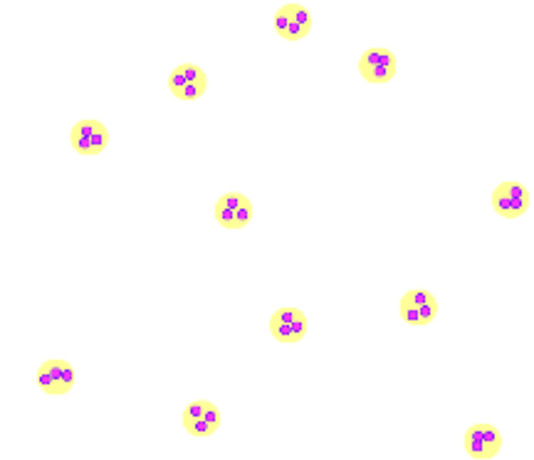
w_i: distance from sample point to NN in data

HOPKINS STATISTIC

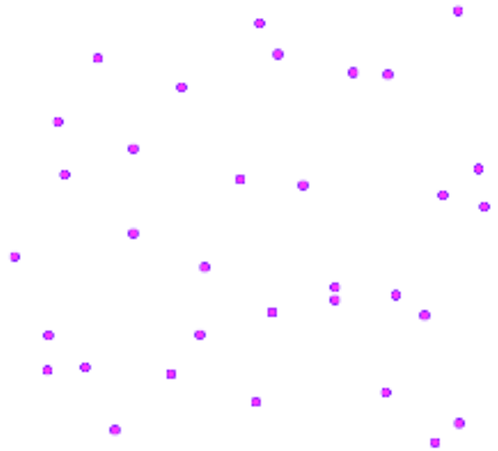
$$H = \frac{\sum_{i=1}^{p} u_i}{\sum_{i=1}^{p} u_i + \sum_{i=1}^{p} w_i}$$

u_i: distance from random point to NN in data

w_i: distance from sample point to NN in data



H close to 1, clustered!



H close to 0.5, random data!

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

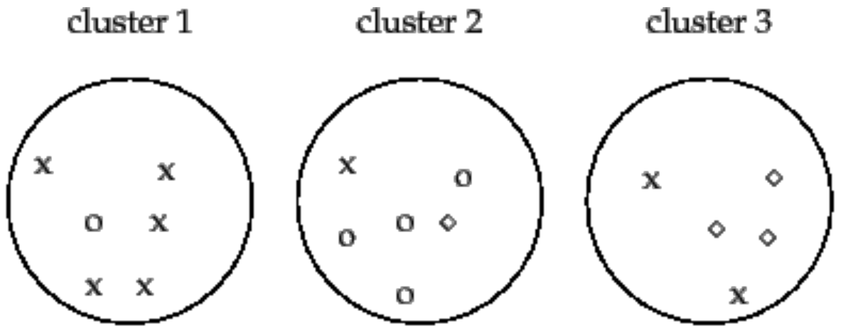
SUPERVISED: EVALUATING CLUSTER QUALITY WITH LABELS

- 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

CLASSIFICATION-ORIENTED

- **Purity**: a measure of the degree to which a cluster/group (G_i) contains objects of one particular class (C_j)
  - A cluster G_i will be labeled as the majority class among all objects in G_i

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



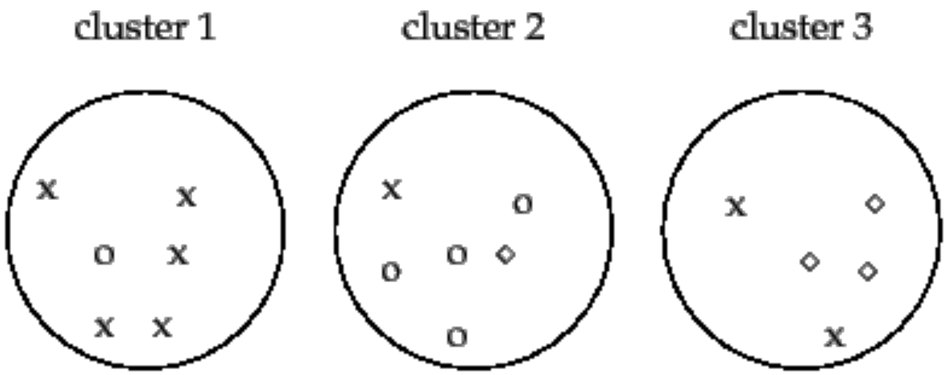
  - High purity is easy to achieve when the number of clusters is large

CLASSIFICATION-ORIENTED

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

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

How does score for this clustering change?



CLASSIFICATION-ORIENTED

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

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

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

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

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

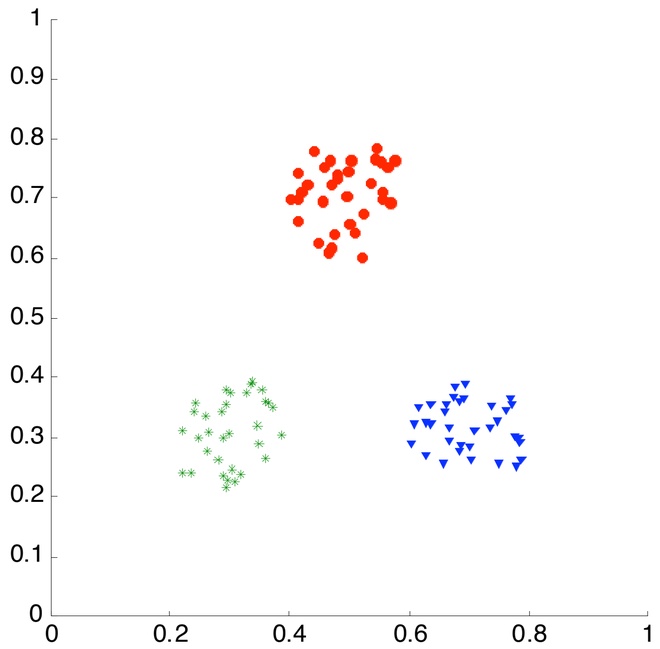
$$Rand = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}} \qquad Jaccard = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

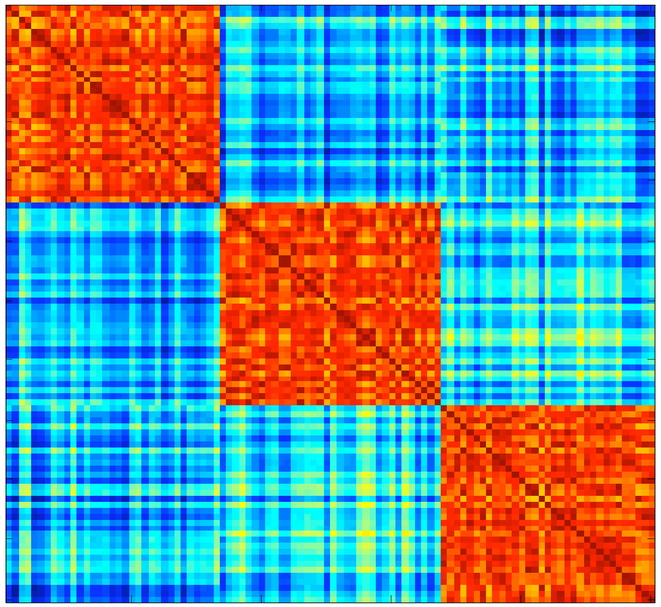
UNSUPERVISED: VISUAL INSPECTION

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

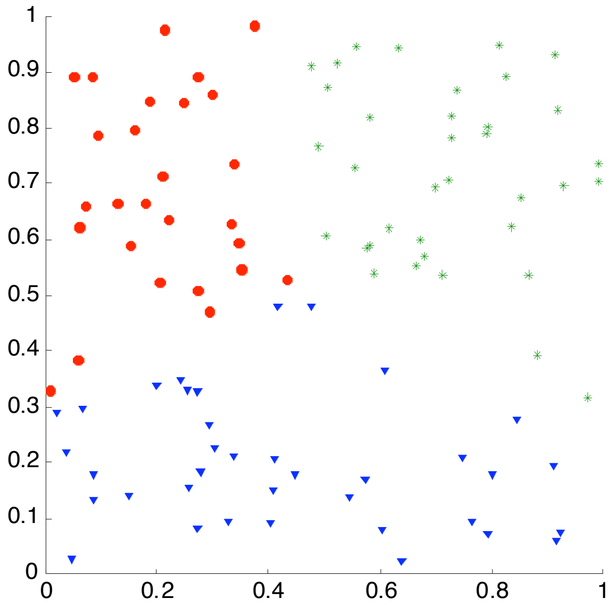
EXAMPLE 1: GOOD CLUSTERING

Proximity matrix reordered
to reflect cluster assignments

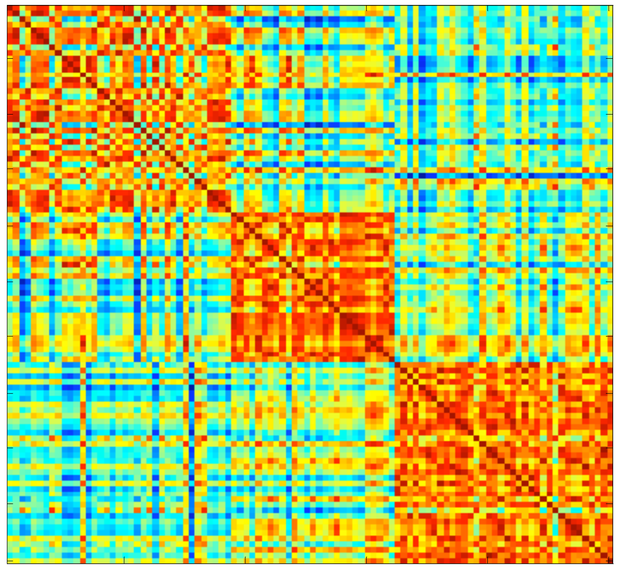




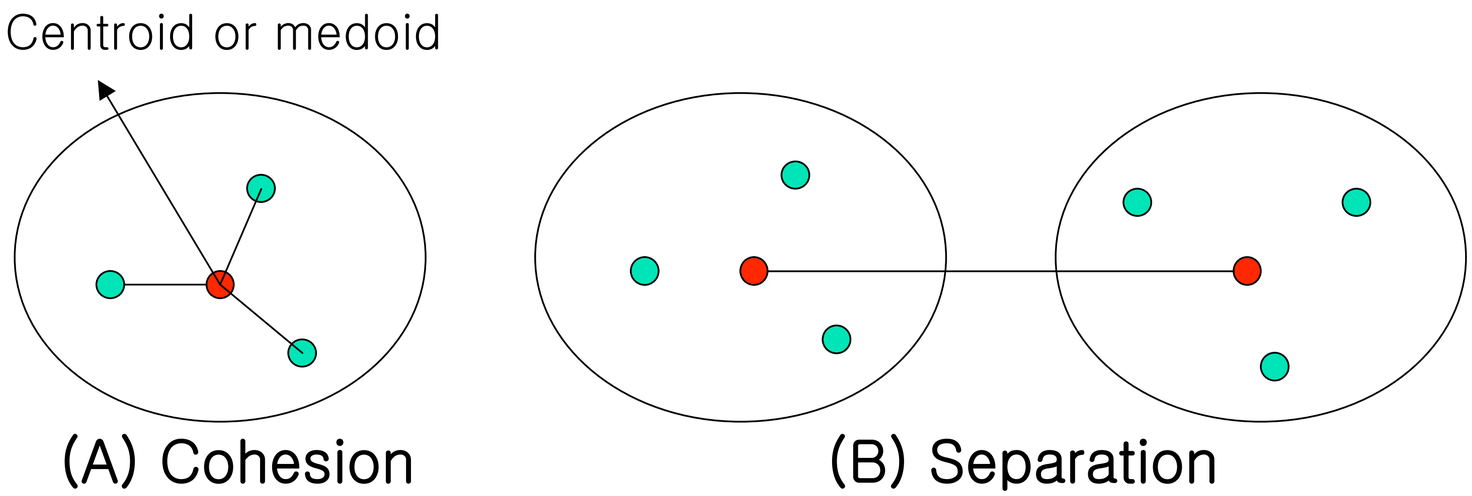
EXAMPLE II: POOR CLUSTERING



Proximity matrix reordered
to reflect cluster assignments



COHESION AND SEPARATION



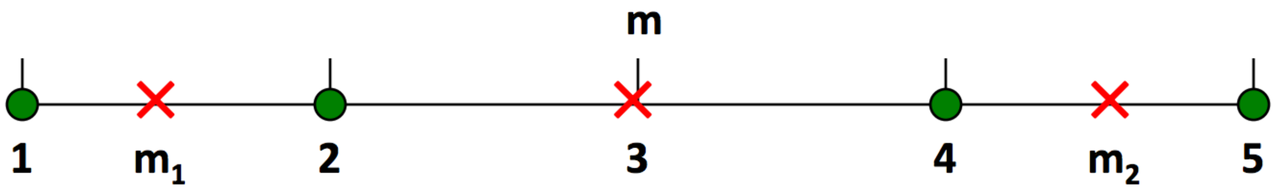
COHESION AND SEPARATION

- 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

COHESION AND SEPARATION: EXAMPLE

- **Cohesion: Within cluster pairwise weight**
  - Sum of distance between all pairs of points in same cluster
- **Separation: Between cluster pairwise weight**
  - Sum of 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

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

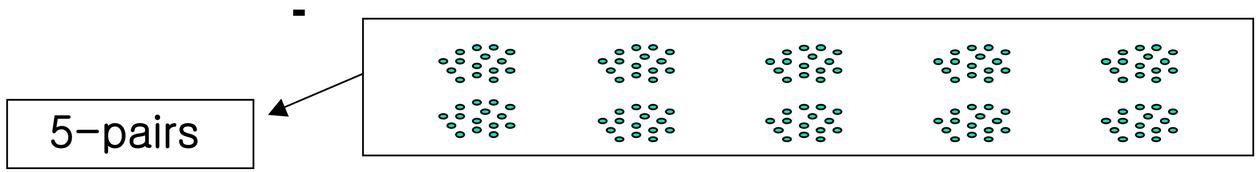
- **Goal 1**: *Describe* data as precisely as possible
- **Goal 2**: *Generalize* to new data
  - Goodness of fit is part of the evaluation, but since the data is not the entire population, we want to learn a model that will generalize to other new data instances
- Thus, want to strike a balance between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

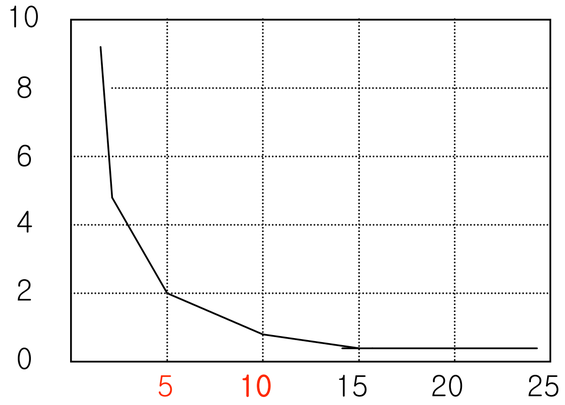
PENALIZED SCORE FUNCTIONS

- 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): $\text{Score}_{BIC} = -2 \log L + p \log n$

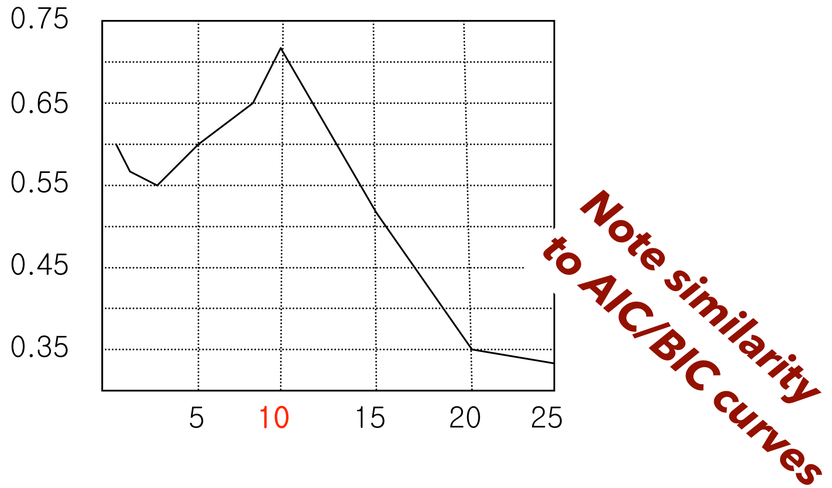
DETERMINING K

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





WSS



Silhouette