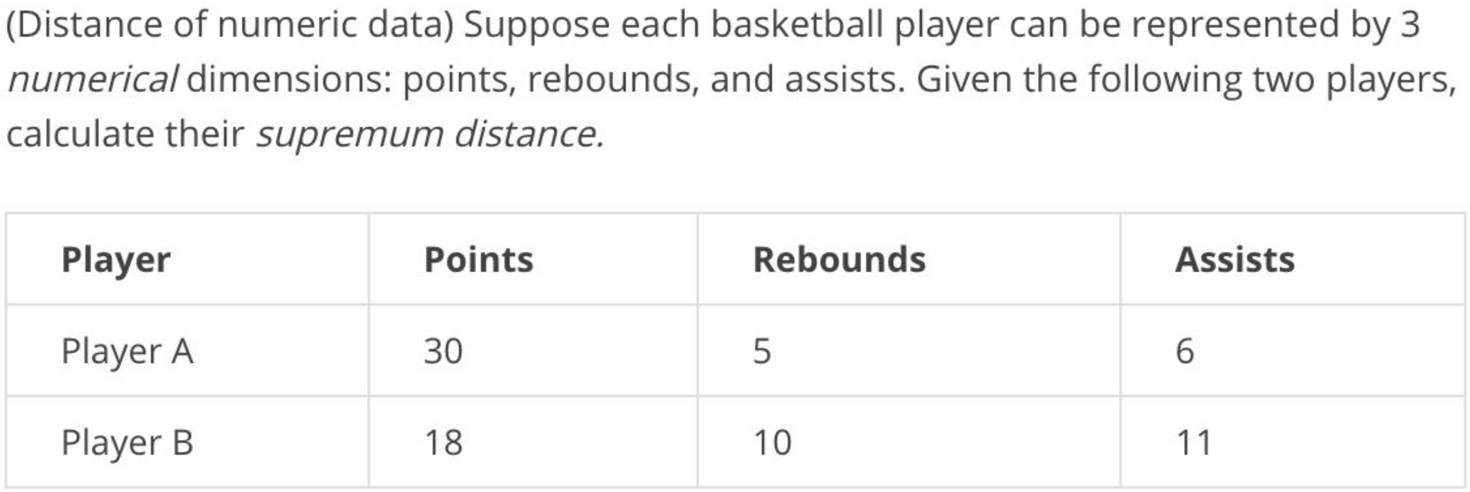
CS412 office hour

Apr 10, 2019

Today's Office Hour

- Exam 2
- QA

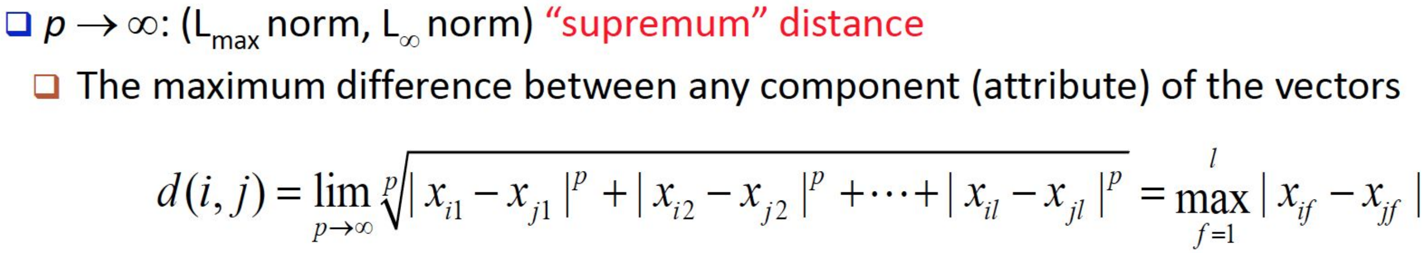
Question 1



(Distance of numeric data) Suppose each basketball player can be represented by 3 *numerical* dimensions: points, rebounds, and assists. Given the following two players, calculate their *supremum distance.*

Player	Points	Rebounds	Assists
Player A	30	5	6
Player B	18	10	11

Solution 1

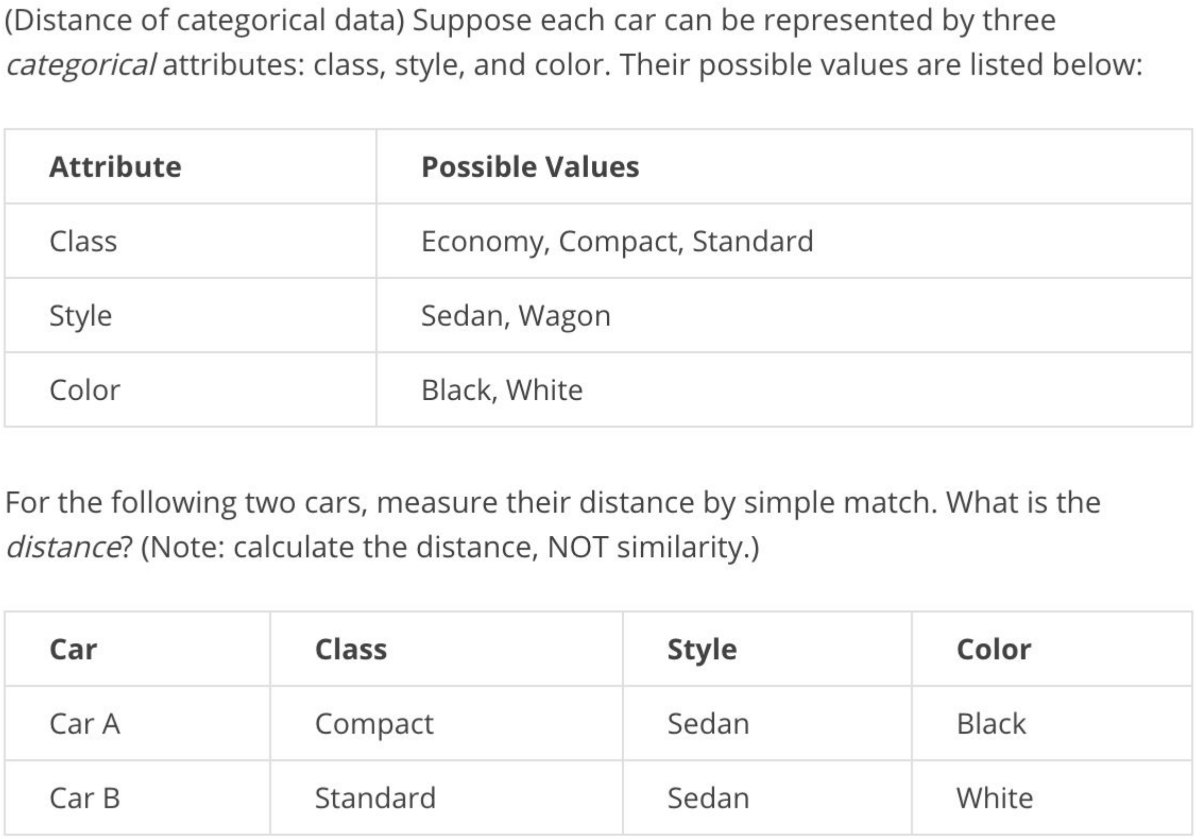


- □ $p \to \infty$: (L_{max} norm, L_{∞} norm) “supremum” distance
  - □ The maximum difference between any component (attribute) of the vectors

$$d(i,j) = \lim_{p \to \infty} \sqrt[p]{|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{il} - x_{jl}|^p} = \max_{f=1}^{l} |x_{if} - x_{jf}|$$

Supremum distance = max(|30-18|, |5-10|, |6-11|) = 12

Question 2



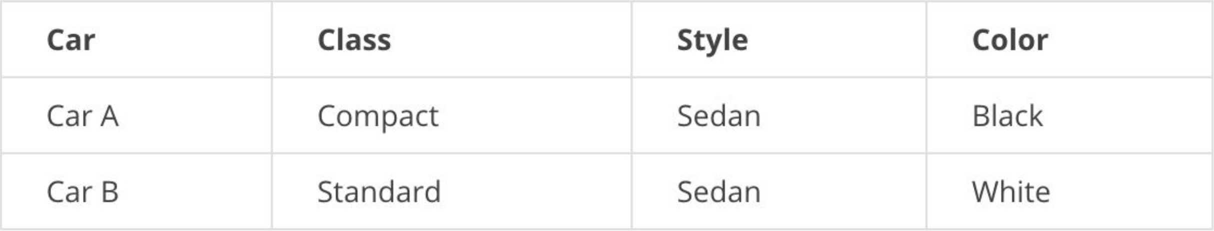
(Distance of categorical data) Suppose each car can be represented by three *categorical* attributes: class, style, and color. Their possible values are listed below:

Attribute	Possible Values
Class	Economy, Compact, Standard
Style	Sedan, Wagon
Color	Black, White

For the following two cars, measure their distance by simple match. What is the *distance*? (Note: calculate the distance, NOT similarity.)

Car	Class	Style	Color
Car A	Compact	Sedan	Black
Car B	Standard	Sedan	White

Solution 2



Car	Class	Style	Color
Car A	Compact	Sedan	Black
Car B	Standard	Sedan	White

mismatch match mismatch

Simple match distance = #mismatch / #attributes = 2 / 3

Question 3

Suppose random variables X_1 and X_2 represent the price (in USD) and the lifespan (in months), respectively, of a smartphone. Given that the covariance of X_1 and X_2 is σ_{12}=480, the standard deviation of X_1 is σ_1=100*sqrt(2) and the standard deviation of X_2 is 3*sqrt(2), where sqrt() is the square root function, which of the following is correct?

1. The Pearson's correlation coefficient ρ_{12} is 0.8 of -0.8.
2. The Pearson's correlation coefficient ρ_{12} is 0.8.
3. X_1 and X_2 must be negatively correlated.
4. X_1 is correlated with X_2 but X_2 is not necessarily correlated to X_1.

Solution 3

$\rho_{12} = \sigma_{12} / \sigma_1\sigma_2 = 480 / (100 * \mathrm{sqrt}(2) * 3 * \mathrm{sqrt}(2)) = 480 / 600 = 0.8 > 0$

Therefore X_1 and X_2 are positively correlated.

Question 4

At a certain iteration in the *k-means* algorithm, there is a cluster consisting of (only) 3 points A=(3, 1), B=(3, 2) and C=(2, 2). What will be the centroid (mean point) of this cluster?

Solution 4

At a certain iteration in the *k-means* algorithm, there is a cluster consisting of (only) 3 points A=(3, 1), B=(3, 2) and C=(2, 2). What will be the centroid (mean point) of this cluster?

Centroid = (mean(3, 3, 2), mean(1, 2, 2)) = (8/3, 5/3)

Question 5

Which one of the following statements is *CORRECT*?

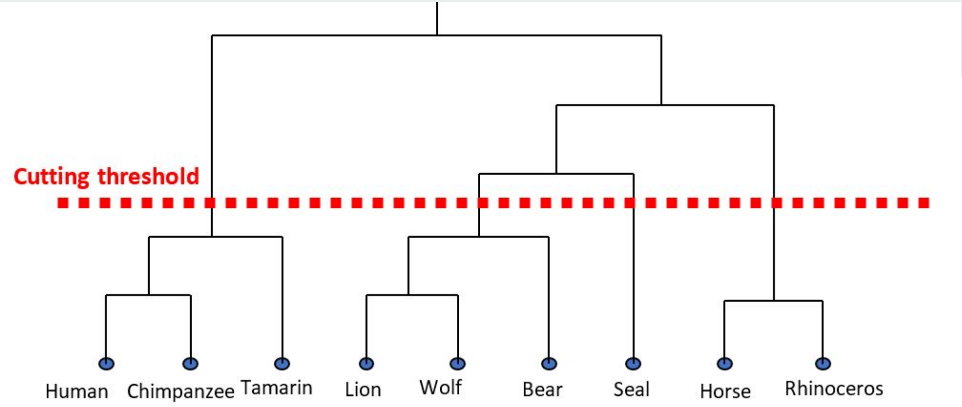
1. The centroids in the k-means algorithm may not be any observed data points.
2. The centroids in the k-medoids algorithm may not be any observed data points.
3. The k-means algorithm always gives the same clustering results, even with different initializations.
4. The k-median algorithm is less robust to outliers compared to the k-means algorithm.

Solution 5

Which one of the following statements is *CORRECT*?

1. **The centroids in the k-means algorithm may not be any observed data points.**
2. The centroids in the k-medoids algorithm may not be any observed data points.
3. The k-means algorithm always gives the same clustering results, even with different initializations.
4. The k-median algorithm is less robust to outliers compared to the k-means algorithm.

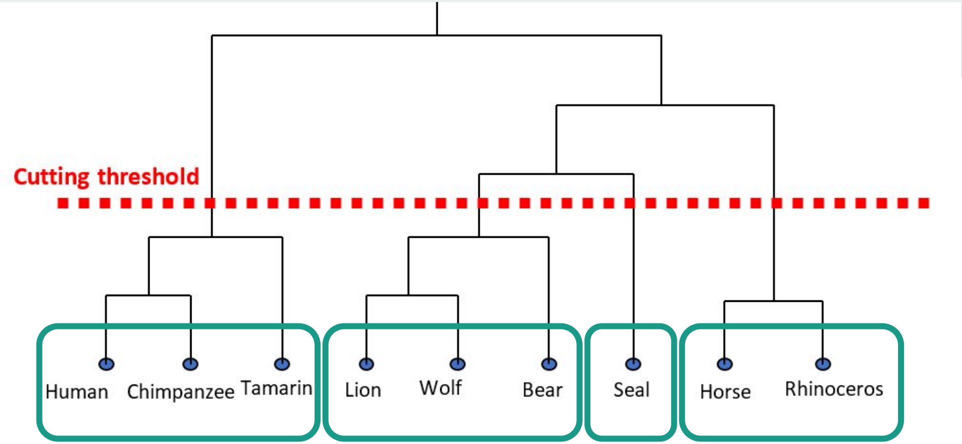
Question 6



Given the dendrogram and the cutting threshold, which one of the following statements is *CORRECT*?

1. Lion and seal are in the same cluster.
2. Rhinoceros and lion are in the same cluster.
3. Human and chimpanzee are in different clusters.
4. Bear and seal are in different clusters.

Solution 6



Given the dendrogram and the cutting threshold, which one of the following statements is *CORRECT*?

1. Lion and seal are in the same cluster.
2. Rhinoceros and lion are in the same cluster.
3. Human and chimpanzee are in different clusters.
4. **Bear and seal are in different clusters.**

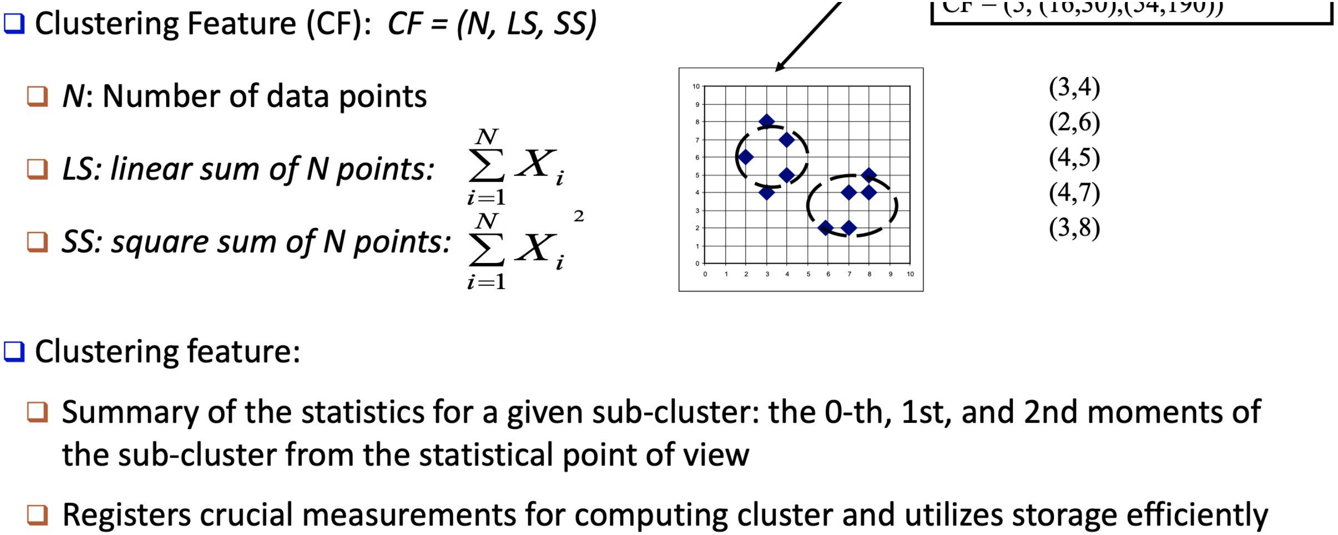
Question 7

(BIRCH) In the BIRCH algorithm, for each micro-cluster, which of the following is *NOT* directly stored as the clustering features (CF) in the in-memory CF tree or cannot be directly calculated from the clustering features?

- Number of data points
- Centroid
- Linear sum
- **All data points**

Solution 7

- The whole idea of BIRCH is efficiency. At each level, for each node, it stores only statistics



CF = (5, (16,30),(54,190))

❑ Clustering Feature (CF): *CF = (N, LS, SS)*

- ❑ *N*: Number of data points
- ❑ *LS: linear sum of N points:* $\sum_{i=1}^{N} X_i$
- ❑ *SS: square sum of N points:* $\sum_{i=1}^{N} X_i^2$

(3,4)
(2,6)
(4,5)
(4,7)
(3,8)

❑ Clustering feature:

- ❑ Summary of the statistics for a given sub-cluster: the 0-th, 1st, and 2nd moments of the sub-cluster from the statistical point of view
- ❑ Registers crucial measurements for computing cluster and utilizes storage efficiently

Question 8

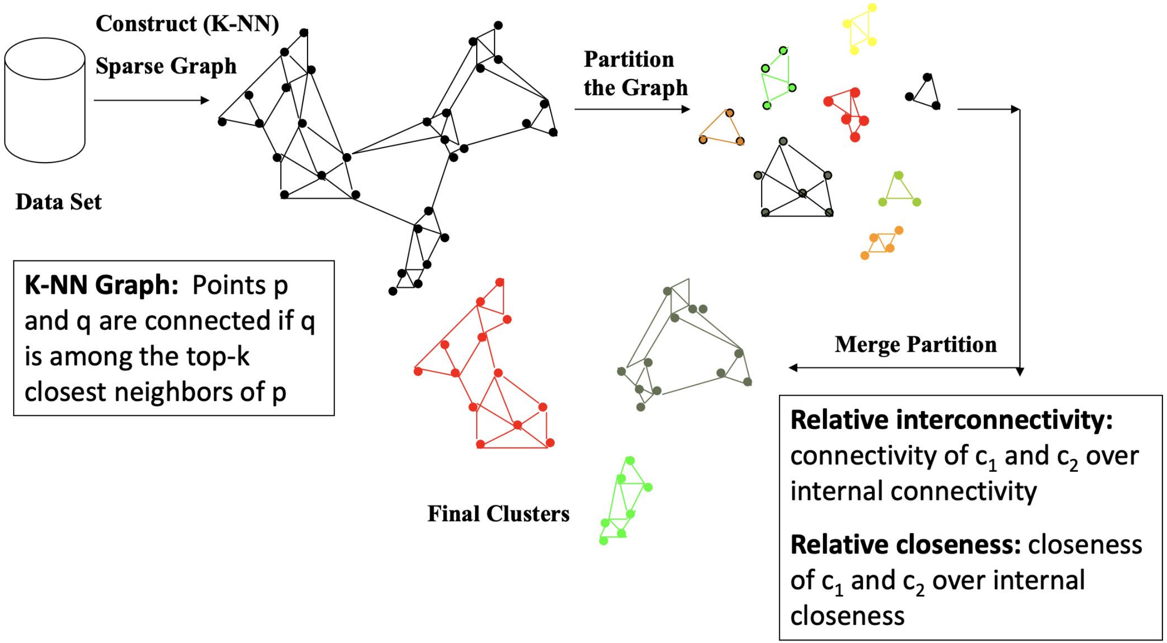
(CHAMELEON) Which one of the following statements is *WRONG* about the CHAMELEON algorithm?

- **In CHAMELEON, two clusters are merged if the absolute interconnectivity between the two clusters is above the predefined threshold.**
- CHAMELEON is capable of generating quality clusters at clustering complex objects.
- CHAMELEON is a graph-based approach that is applicable to cluster two-dimensional spatial data.
- CHAMELEON is a hierarchical clustering algorithm.

Solution 8

- CHAMELEON: A graph partitioning approach (G. Karypis, E. H. Han, and V. Kumar, 1999)
- Measures the similarity based on a dynamic model
  - Two clusters are merged only if the *interconnectivity* and *closeness (proximity)* between two clusters are high *relative to* the internal interconnectivity of the clusters and closeness of items within the clusters
- A graph-based, two-phase algorithm
  1. Use a graph-partitioning algorithm: Cluster objects into a large number of relatively small sub-clusters
  2. Use an agglomerative hierarchical clustering algorithm: Find the genuine clusters by repeatedly combining these sub-clusters

Solution 8



K-NN Graph: Points p and q are connected if q is among the top-k closest neighbors of p

Relative interconnectivity: connectivity of c_1 and c_2 over internal connectivity

Relative closeness: closeness of c_1 and c_2 over internal closeness

Question 9

(DBSCAN) For the DBSCAN algorithm with given Eps and MinPts, which one of the following statements is *NOT* correct?

- If a point p is directly density-reachable from a point q, then p is also density-reachable from q
- If a point p is density-reachable from a point q, then p is also density-connected to q
- **If a point p is density-connected to a point q, then q is also density-reachable from p**
- If a point p is density-reachable from a point q, q might **not** be density-reachable from p.

Solution 9

(DBSCAN) For the DBSCAN algorithm with given Eps and MinPts, which one of the following statements is *NOT* correct?

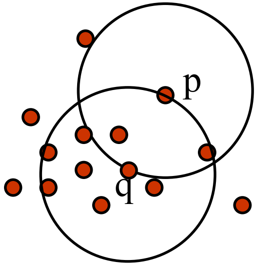
- If a point p is directly density-reachable from a point q, then p is also density-reachable from q
- If a point p is density-reachable from a point q, then p is also density-connected to q
- **If a point p is density-connected to a point q, then q is also density-reachable from p**
  - **Counter example: p and q are border points**
- If a point p is density-reachable from a point q, q might **not** be density-reachable from p.

DBSCAN

- Directly density-reachable

MinPts = 5

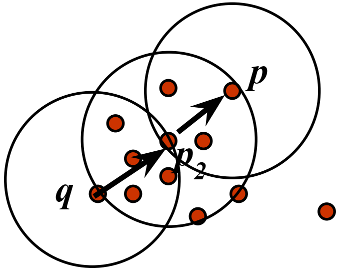
Eps = 1 cm



p is directly density-reachable from q
q is NOT directly density-reachable from p
(asymmetric)

DBSCAN

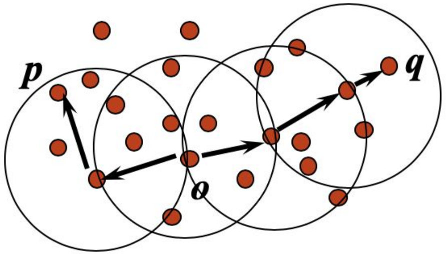
- Density-reachable



p is density-reachable from q
q is NOT density reachable from p
(asymmetric)

DBSCAN

- Density-connected



p is density-connected to q
q is density-connected to p
(symmetric)

Question 10

(OPTICS) Which one of the following statements is *CORRECT* about OPTICS?

- **OPTICS stores clustering order using information of both core distance and reachability distance.**
- OPTICS is an extension of DBSCAN that can automatically find clusters; however, it cannot be used for interactive cluster analysis.
- Since DBSCAN cannot find hierarchically nested clustering structures, as an extension of DBSCAN, neither can OPTICS.
- OPTICS is sensitive to parameter setting.

Question 11

(Comparison) Which one of the following algorithms *CANNOT* generate clusters of irregular shapes?

- CHAMELEON
- DBSCAN
- **KMeans**
- CURE

Question 12

(Grid-based) Which one of the following statements about STING and CLIQUE is *CORRECT*?

- **For both STING and CLIQUE, the quality of the results is limited by predefined cell sizes and the density threshold**
- Both STING and CLIQUE are subspace clustering algorithms
- Both STING and CLIQUE scale linearly with the size of number of data points
- Both STING and CLIQUE may lose accuracy in query processing

Solution 12

(Grid-based) Which one of the following statements about STING and CLIQUE is *CORRECT*?

- **For both STING and CLIQUE, the quality of the results is limited by predefined cell sizes and the density threshold**
  - This is a weakness of all grid-based algorithms
- Both STING and CLIQUE are subspace clustering algorithms
  - STING is not
- Both STING and CLIQUE scale linearly with the size of number of data points
  - STING scales with number of cells K in the lowest level, K << N
- Both STING and CLIQUE may lose accuracy in query processing
  - CLIQUE does not

Question 13

(Evaluation) Which one of the following measures does *NOT* need ground truth cluster labels?

- Purity
- NMI
- Fowlkes-Mallow measure
- **BetaCV**

Solution 13

(Evaluation) Which one of the following measures does *NOT* need ground truth cluster labels?

- Purity
  - Matching based
- NMI
  - Entropy based
- Fowlkes-Mallow measure
  - Pairwise
- **BetaCV**
  - Internal

Question 14

(External Measures) Which of the following statements about a perfect clustering for a ground truth partition is *INCORRECT*?

- The corresponding purity measure is 1
- The corresponding maximum matching is 1
- **The corresponding conditional entropy is 1**
- The corresponding normalized mutual information is 1

Solution 14

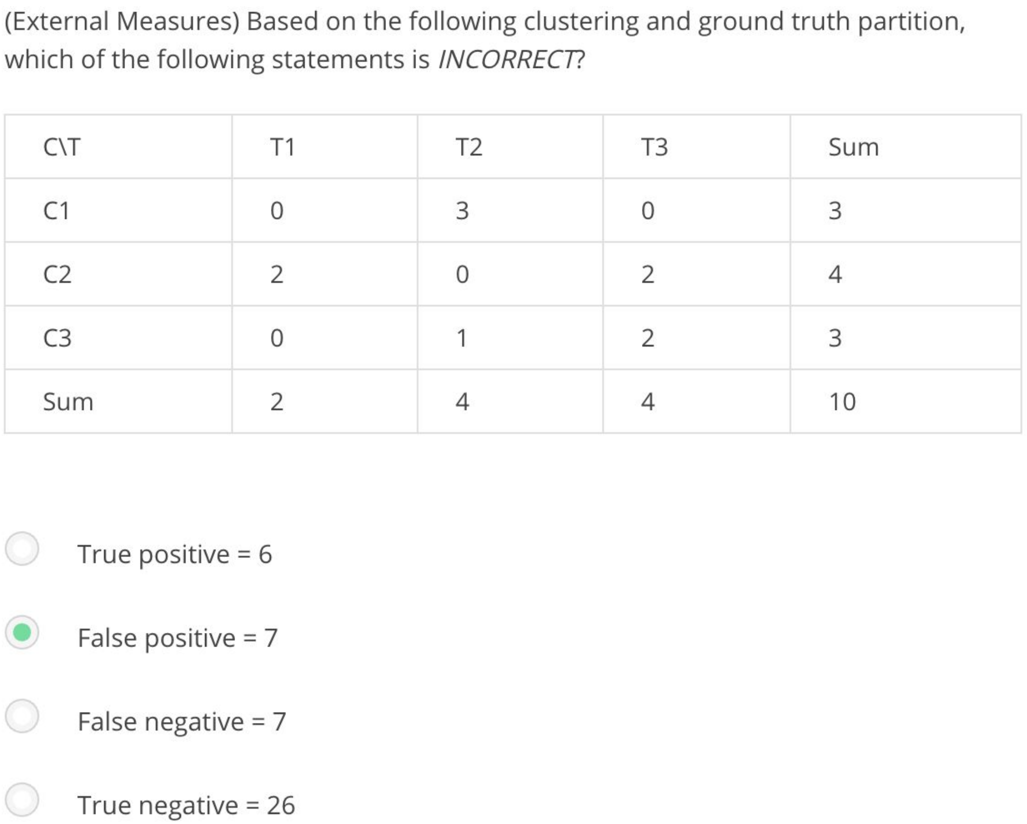
(External Measures) Which of the following statements about a perfect clustering for a ground truth partition is *INCORRECT*?

- The corresponding purity measure is 1
- The corresponding maximum matching is 1
- **The corresponding conditional entropy is 1**
  - The additional information needed to describe one RV given another RV
  - Since clusters and partitions perfectly match, the additional info is 0
- The corresponding normalized mutual information is 1
  - Independent -> 0
  - Perfectly correlated -> 1

5	0	0
0	6	0
0	0	7

Verify on this perfect confusion matrix

Question 15



(External Measures) Based on the following clustering and ground truth partition, which of the following statements is *INCORRECT*?

C\T	T1	T2	T3	Sum
C1	0	3	0	3
C2	2	0	2	4
C3	0	1	2	3
Sum	2	4	4	10

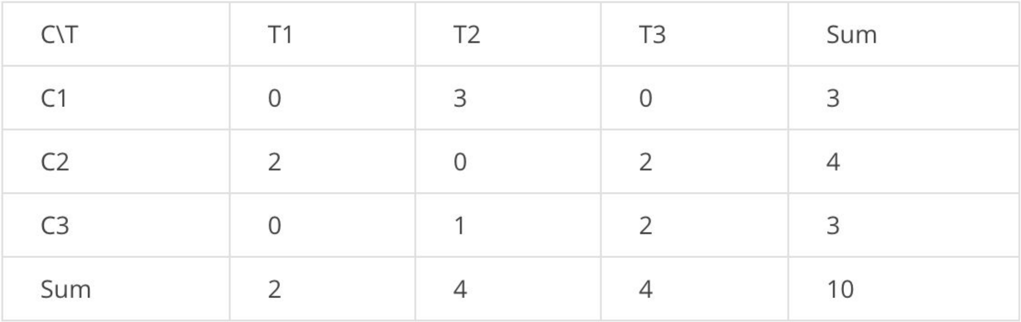
○ True positive = 6

● False positive = 7

○ False negative = 7

○ True negative = 26

Solution 15



C\T	T1	T2	T3	Sum
C1	0	3	0	3
C2	2	0	2	4
C3	0	1	2	3
Sum	2	4	4	10

TP = sum of (n_ij choose 2) = 3 + 1 + 1 + 1 = 6
FP = sum of (n_i choose 2) - TP = 3 + 6 + 3 - 6 = 6   (not 7)
FN = sum of (m_j choose 2) - TP = 1 + 6 + 6 - 6 = 7
TN = (n choose 2) - TP - FP - FN = 45 - 6 - 6 - 7 = 26

Question 16

(Measures) Regarding clustering evaluation measures, which of the following statements is *INCORRECT*?

- For an obtained clustering, the larger the F-measure is, the better the clustering is
- For an obtained clustering, the smaller the Beta-CV measure is, the better the clustering is
- For an obtained clustering, the smaller the conditional entropy is, the better the clustering is
- For an obtained clustering, the smaller the Silhouette coefficient is, the better the clustering is.

Solution 16

(Measures) Regarding clustering evaluation measures, which of the following statements is *INCORRECT*?

- For an obtained clustering, the larger the F-measure is, the better the clustering is
    - How well cluster matches with ground truth
- For an obtained clustering, the smaller the Beta-CV measure is, the better the clustering is
    - Intra-cluster distance / Inter-cluster distance
    - Good results indicates points inside cluster are close, while points across clusters are far
- For an obtained clustering, the smaller the conditional entropy is, the better the clustering is
    - Smaller conditional entropy means more correlation on cluster result and ground truth
- **For an obtained clustering, the smaller the Silhouette coefficient is, the better the clustering is**

Solution 16

- **Silhouette coefficient** as an **internal measure**: Check cluster cohesion and separation
  - For each point ***x**$_i$*, its silhouette coefficient s_i is: $s_i = \frac{\mu_{out}^{\min}(\mathbf{x}_i) - \mu_{in}(\mathbf{x}_i)}{\max\{\mu_{out}^{\min}(\mathbf{x}_i), \mu_{in}(\mathbf{x}_i)\}}$

    where $\mu_{in}(\mathbf{x}_i)$ is the mean distance from ***x**$_i$* to points in its own cluster

    $\mu_{out}^{\min}(\mathbf{x}_i)$ is the mean distance from ***x**$_i$* to points in its closest cluster
  - Silhouette coefficient (*SC*) is the mean values of s_i across all the points: $SC = \frac{1}{n}\sum_{i=1}^{n} s_i$
  - *SC* close to +1 implies good clustering
    - Points are close to their own clusters but far from other clusters