COSC 3337 : Data Science I

N. Rizk

College of Natural and Applied Sciences

Department of Computer Science

University of Houston

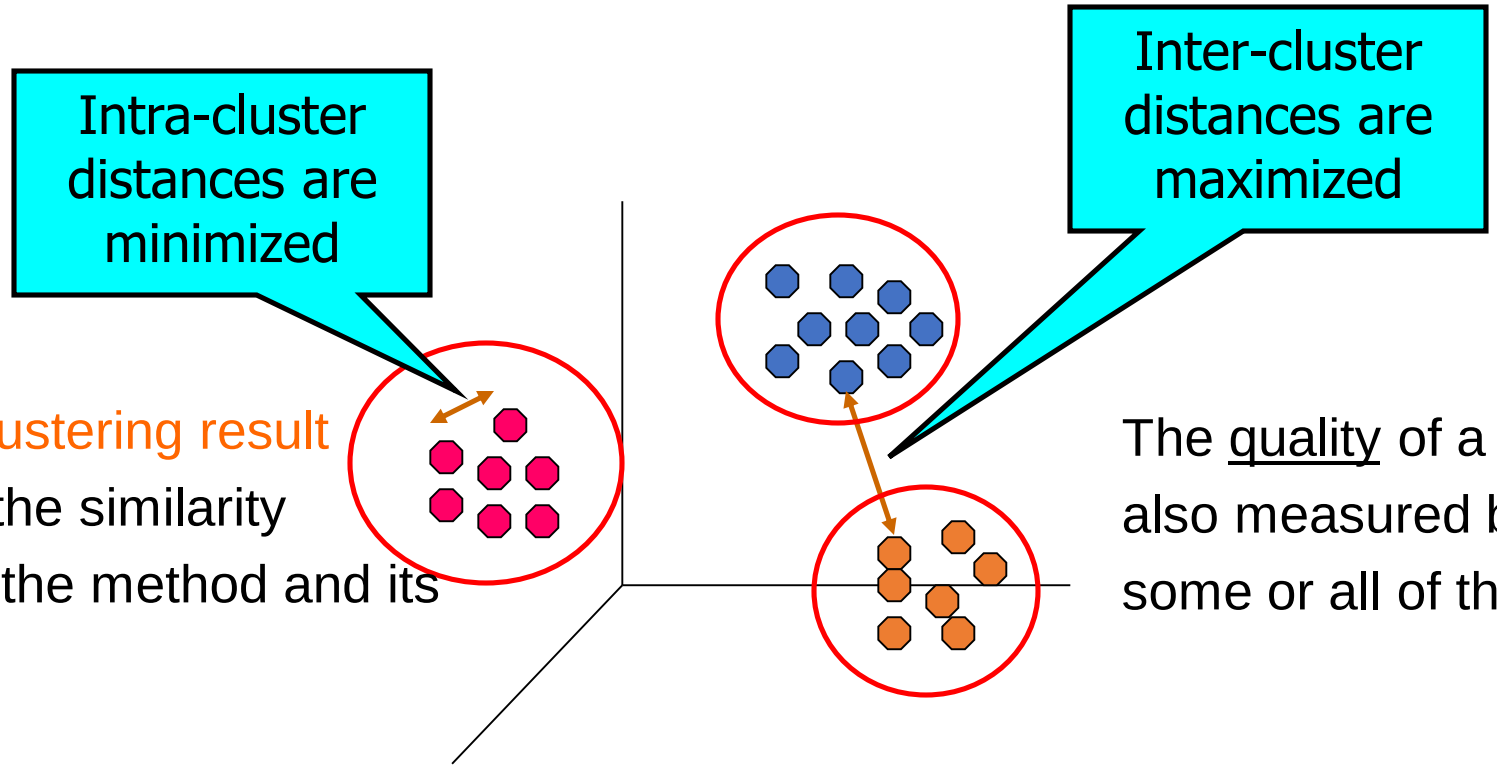
What is Cluster Analysis?

- Cluster: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Cluster analysis
  - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters
- Unsupervised learning: no predefined classes
- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms

Measure the Quality of Clustering

- Dissimilarity/Similarity metric: Similarity is expressed in terms of **a distance function, which is typically metric:**
  $d(i, j)$
- There is a separate "quality" function that measures the "goodness" of a cluster.
- The definitions of distance functions are usually very different for **interval-scaled, boolean, categorical, ordinal and ratio variables.**
- Weights should be associated with different variables based on applications and data semantics.
- It is hard to define "similar enough" or "good enough"
  - the answer is typically highly subjective.

Intra-cluster and Inter-cluster distances



The quality of a clustering result depends on both the similarity measure used by the method and its implementation.

The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns

What is good clustering?

A **good clustering** method will produce high quality clusters with

- high **intra-class** similarity
- low **inter-class** similarity

The **quality** of a clustering result depends on

the similarity measure used

implementation of the similarity measure

The **quality** of a clustering method is also measured by its ability to discover some or all of the **hidden** patterns

Requirements of clustering

- Scalability
- Ability to deal with different types of attributes
- Discovery of clusters with arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
- Ability to deal with noise and outliers
- Insensitivity to order of input records
- High dimensionality
- Incorporation of user-specified constraints
- Interpretability and usability

Cohesion and Separation

- **Cohesion** is measured by the within cluster sum of squares(How closely related are objects in a cluster?)

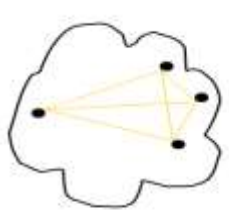
$$WSS = \sum_{i} \sum_{x \in C_i} (x - {}_i m^2)$$

- **Separation** is measured by the between cluster sum of squares(How distant and well separated a cluster is from other clusters)

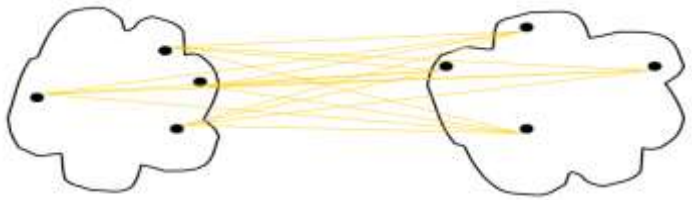
$$BSS = \sum_{i} |C_i| (m - m_i)^2$$

where $|Ci|$ is the size of cluster i, m is the centroid of the whole data set

- BSS + WSS = constant
- WSS (Cohesion) measure is called Sum of Squared Error (SSE)—a commonly used measure
- A larger number of clusters tend to result in smaller SSE



cohesion



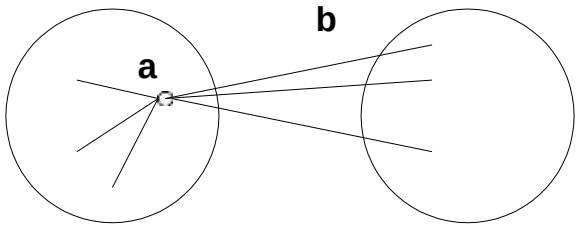
separation

Silhouette Coefficient

- Silhouette Coefficient combines ideas of both cohesion and separation

- For an individual point, *i*
  - Calculate ***a*** = average distance of *i* to the points in its cluster(Smaller the value better the assignment)
  - Calculate ***b*** = min (average distance of *i* to points in another cluster)
  - The **silhouette coefficient** for a point is then given by
    $s = 1 - a/b$ if $a < b$, ($s = b/a - 1$ if $a \geq b$, not the usual case)

  - Typically between 0 and 1
  - The closer to 1 the better



Silhouette index for cohesion and separation

Silhouette Coefficient…continue

S.A. is a way to measure how close each point in a cluster is to the points in its neighboring clusters

Its a neat way to find out the optimum value for k during k-means clustering.

Silhouette values lies in the range of [-1, 1]. A value of +1 indicates that the sample is far away from its neighboring cluster and very close to the cluster its assigned. Similarly, value of -1 indicates that the point is close to its neighboring cluster than to the cluster its assigned. And, a value of 0 means its at the boundary of the distance between the two cluster. Value of +1 is idea and -1 is least preferred.

Correlation with Distance Matrix

- Distance Matrix
  - D_{ij} is the similarity between object O_i and O_j
- Incidence Matrix
  - C_{ij}=1 if O_i and O_j belong to the same cluster, C_{ij}=0 otherwise
- Compute the correlation between the two matrices
  - Only $n(n-1)/2$ entries needs to be calculated
- High correlation indicates good clustering

Correlation with Distance Matrix

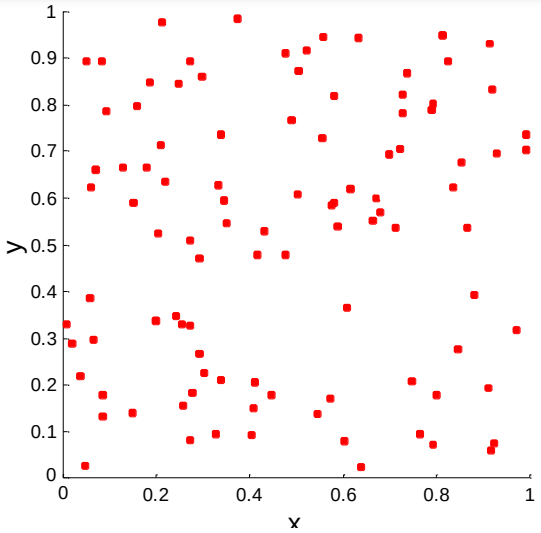
- Given Distance Matrix D = $\{d_{11}, d_{12}, \ldots, d_{nn}\}$ and Incidence Matrix $C = \{c_{11}, c_{12}, \ldots, c_{nn}\}$.

- Correlation r between D and C is given by

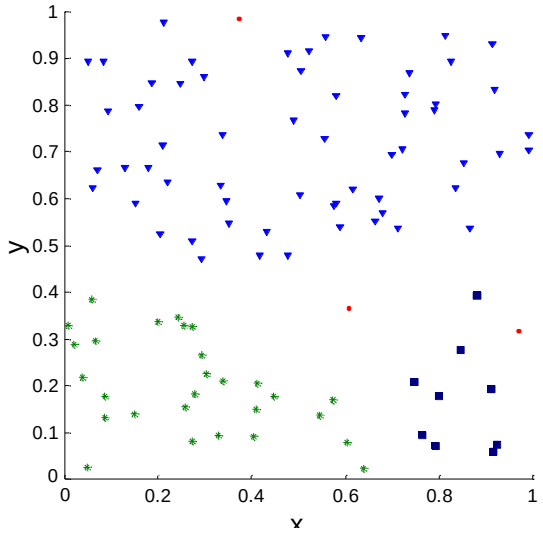
$$r = \frac{\sum_{i=1,j=1}^{n} (d_{ij} - \bar{d})(c_{ij} - \bar{c})}{\sqrt{\sum_{i=1,j=1}^{n} (d_{ij} - \bar{d})^2} \sqrt{\sum_{i=1,j=1}^{n} (c_{\overline{ij}}\ c)^{\ \bar{}2}}}$$

Are There Clusters in the Data?

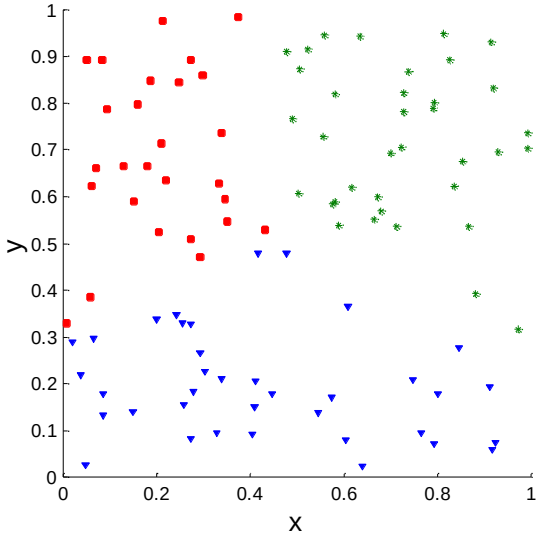
Random Points



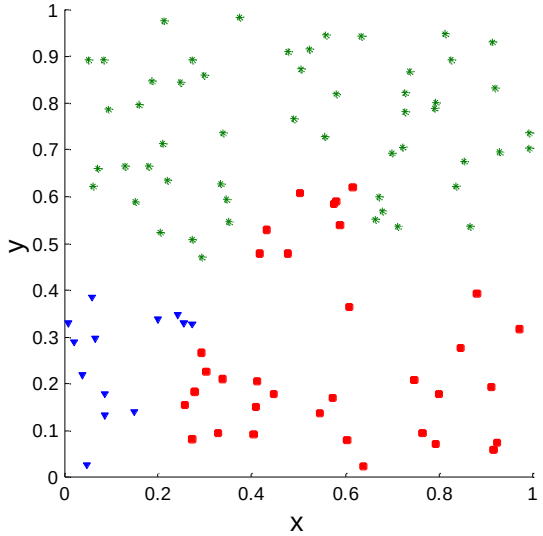
DBSCAN



K-means

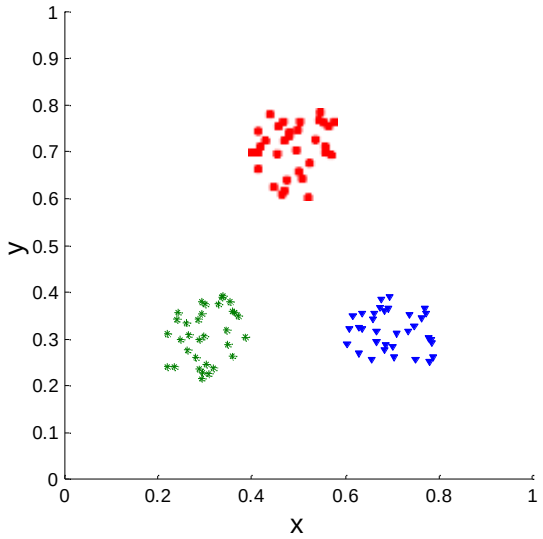


Complete Link

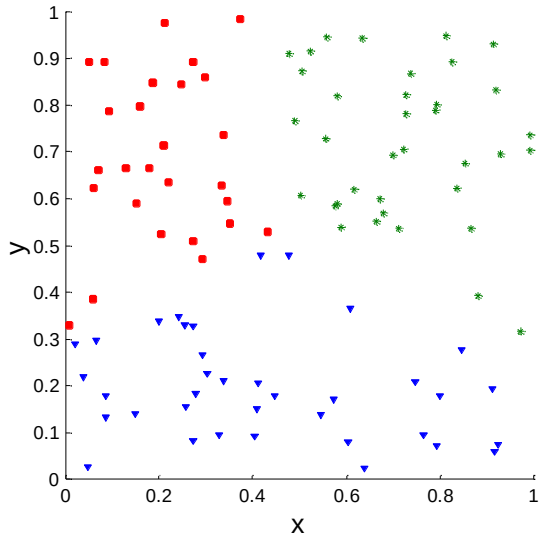


Measuring Cluster Validity Via Correlation

- Correlation of incidence and distance matrices for the K-means clustering of the following two data sets



Corr = -0.9235



Corr = -0.5810

```
Cluster 1 ={{1,0},{1,1}}
Cluster 2 {{1,2},{2,3},{2,2},{1,2}},
Cluster 3 ={{3,1},{3,3},{2,1}}
```

1. Take a point {1,0} in cluster 1
2. Calculate its **average distance** to all other points in **it's cluster**,
   For cluster 1
   =√((1-1)^2 + (0-1)^2) =√(0+1)=√1=1

➔ ***the average distance of point {1,0} in cluster 1 to all the points in cluster 1 = 1***

for the object {1,0} in cluster 1 calculate its average distance from all the objects in cluster 2 and cluster 3.
Of these take the **minimum average distance.**

```
Cluster 1 ={{1,0},{1,1}}
Cluster 2 {{1,2},{2,3},{2,2},{1,2}},
Cluster 3 ={{3,1},{3,3},{2,1}}
```

So for cluster 2

{1,0} ----> {1,2} = distance = √((1-1)^2 + (0-2)^2) =√(0+4)=√4=2
{1,0} ----> {2,3} = distance = √((1-2)^2 + (0-3)^2) =√(1+9)=√10=3.16
{1,0} ----> {2,2} = distance = √((1-2)^2 + (0-2)^2) =√(1+4)=√5=2.24
{1,0} ----> {1,2} = distance = √((1-1)^2 + (0-2)^2) =√(0+4)=√4=2

➔ ***the average distance of point {1,0} in cluster 1 to all the points in cluster 2 = (2+3.16+2.24+2)/4 = 2.325***

```
Cluster 1 ={{1,0},{1,1}}
Cluster 2 {{1,2},{2,3},{2,2},{1,2}},
Cluster 3 ={{3,1},{3,3},{2,1}}
```

So for cluster 3

{1,0} ----> {3,1} = distance = √((1-3)^2 + (0-1)^2) =√(4+1)=√5=2.24
{1,0} ----> {3,3} = distance = √((1-3)^2 + (0-3)^2) =√(4+9)=√13=3.61
{1,0} ----> {2,1} = distance = √((1-2)^2 + (0-1)^2) =√(1+1)=√2=2.24

➔ the average distance of point {1,0} in cluster 1 to all the points in cluster 3 =(2.24+3.61+2.24)/3 = 2.7

Silhouette coefficient for cluster 1

- {1,0} from cluster 1 is 1 (a1)
- {1,0} from cluster 2 is 2.325
- {1,0} from cluster 3 is 2.7

- Since 2.325<2.7➔ the minimum average is 2.325 (b1)

- So the silhouette coefficient of cluster 1

- s1= 1-(a1/b1) = 1- (1/2.325)=1-0.4301=0.5699

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}$$

Silhouette coefficient for cluster 2 and cluster 3

Calculate the silhouette coefficient for cluster 2 and cluster 3 separately by taking any single object point in each of the clusters and repeating the previous steps.

The cluster with the greatest silhouette coefficient is the best as per evaluation.

```
Cluster 1 ={{1,0},{1,1}}
Cluster 2 {{1,2},{2,3},{2,2},{1,2}},
Cluster 3 ={{3,1},{3,3},{2,1}}
```

$$WSS = \sum \sum (x - m)$$

Cohesion(C1) = (1 - 1) ^ 2 + (1 - 1) ^ 2 + (0 - .5) ^ 2 + (1 - .5) ^ 2 = 0.5

(green and brown centroids)

$$BSS = \sum C_i (m - m_i)$$

Separation: {Between clusters i.e. (C1,C2) , (C1,C3) & (C2,C3)}

Separation(C1,C2) = SSE(Centroid(C1), Centroid(C2))
= (1 - 1.5) ^ 2 + (0.5 - 2.25) ^ 2 = 1 + 3.0625 = 4.0625

The silhouette combines idea of cohesion and separation

Cohesion(C1)=0.5
How dissimilar the points are inside a cluster

➔small value is preferred

Separation(C1,C2)=4.06
Large value means cluster c1 is badly matched with c2

➔high value is preferred

Silhouette=0.5699
High separation and low cohesion correponds to values close to 1 for the silhouette
➔how well the clustering algorithm has performed
➔can be used to determine the best number of clusters

```
from sklearn.metrics import silhouette_samples, silhouette_score
```

```
# The silhouette_score gives the average value for all the samples.
  # This gives a perspective into the density and separation of the formed
  # clusters
  silhouette_avg = silhouette_score(X, cluster_labels)
  print("For n_clusters =", n_clusters,
        "The average silhouette_score is :", silhouette_avg)
```

```
For n_clusters = 2 The average silhouette_score is : 0.7049787496083262
For n_clusters = 3 The average silhouette_score is : 0.5882004012129721
For n_clusters = 4 The average silhouette_score is : 0.6505186632729437
For n_clusters = 5 The average silhouette_score is : 0.56376469026194
For n_clusters = 6 The average silhouette_score is : 0.4504666294372765
```

Measuring Clustering Quality

Measuring Clustering Quality

- Two methods: extrinsic vs. intrinsic
- Extrinsic: supervised, i.e., the ground truth is available
  - Compare a clustering against the ground truth using certain clustering quality measure
  - Ex. Purity, precision and recall metrics, normalized mutual information
- Intrinsic: unsupervised, i.e., the ground truth is unavailable
  - Evaluate the goodness of a clustering by considering how well the clusters are separated, and how compact the clusters are
  - Ex. Silhouette coefficient

Entropy and Purity

- **Notation**
  - $|C_k \cap P_j|$ the number of objects in both the *k*-th cluster of the clustering solution and j-th cluster of the groundtruth
  - $|C_k|$ the number of objects in the *k*-th cluster of the clustering solution
  - $|P_j|$ the number of objects in the *j*-th cluster of the groundtruth

- **Purity**

$$Purity = \frac{1}{N}\sum_k \max_j |C_k \cap P_j|$$

- **Normalized Mutual Information**

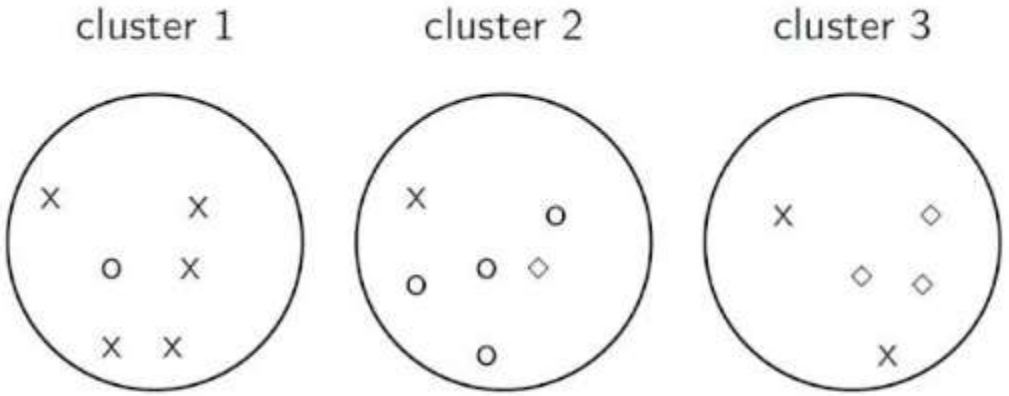
$$NMI = \frac{I(C,P)}{\sqrt{H(C)H(P)}}$$

$$I(C,P) = \sum_k \sum_j \frac{|C_k \cap P_j|}{N} \log \frac{N \cdot |C_k \cap P_j|}{|C_k||P_j|}$$

$$H(C) = \sum_k \frac{|C_k|}{N} \log \frac{|C_k|}{N}$$

$$H(P) = \sum_j \frac{|P_j|}{N} \log \frac{|P_j|}{N}$$

Purity Example



Majority class and number of members of the majority class for the three clusters are: x, 5 (cluster 1); o, 4 (cluster 2); and ◇, 3 (cluster 3). Purity is $(1/17) \times (5 + 4 + 3) \approx 0.71$.

Ground Truth

Each number is an object an each circle or square is a community
Analysis of clustering
{1,3} and {2} maps to 1,2,3 circle

➔ {2} is wrongly clustered

Thus the use of Normalized Mutual Information NMI to validate a cluster

Normalized Mutual Information

- NMI is a good measure for determining the quality of clustering.
- It is an external measure because we need the class labels of the instances to determine the NMI.
- Since it's normalized we can measure and compare the NMI between different clustering having different number of clusters.

Normalized Mutual Information

- Normalized Mutual Information:

$$NMI(Y, C) = \frac{2 \times I(Y; C)}{[H(Y) + H(C)]}$$

- where

1)Y = class labels

2)C = cluster labels

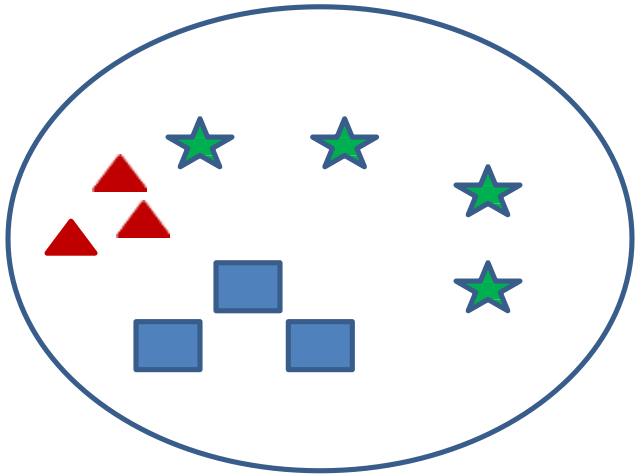
3)H(.) = Entropy

4)I(Y;C) = Mutual Information b/w Y and C

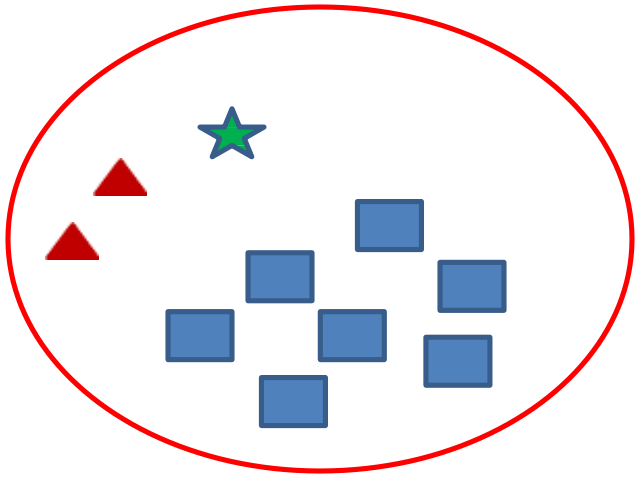
Note: All logs are base-2.

Calculating NMI for Clustering

- Assume m=3 classes and k=2 clusters



Cluster-1 (C=1)



Cluster-2 (C=2)

Class-1 (Y=1) Class-2 (Y=2) Class-3 (Y=3)

H(Y) = Entropy of Class Labels

Total

- P(Y=1) = 5/20 = ¼
- P(Y=2) = 5/20 = ¼
- P(Y=3) = 10/20 = ½
- $H(Y) = -\frac{1}{4}\log\left(\frac{1}{4}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1.5$

This is calculated for the entire dataset and can be calculated prior to clustering, as it will not change depending on the clustering output.

H(C) = Entropy of Cluster Labels

- P(C=1) = 10/20 = 1/2
- P(C=2) = 10/20 = ½
- H(Y) = $-\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1$

This will be calculated every time the clustering changes.

I(Y;C)= Mutual Information

- Mutual information is given as:
  - $$I(Y;C) = H(Y) - H(Y|C)$$
    - We already know H(Y)
    - H(Y|C) is the entropy of class labels within each cluster, how do we calculate this??

Mutual Information tells us the reduction in the entropy of class labels that we get if we know the cluster labels. (Similar to Information gain in deicison trees)

H(Y|C): conditional entropy of class labels for clustering C

- Consider Cluster-1:
  - P(Y=1|C=1)=3/10 (three triangles in cluster-1)
  - P(Y=2|C=1)=3/10 (three rectangles in cluster-1)
  - P(Y=3|C=1)=4/10 (four stars in cluster-1)
  - Calculate conditional entropy as:

$$H(Y|C=2) = -P(C=2) \sum_{y\in\{1,2,3\}} P(Y=y|C=2)\log(P(Y=y|C=2))$$

$$=-\frac{1}{2}\times\left[\frac{2}{10}\log\left(\frac{2}{10}\right)+\frac{7}{10}\log\left(\frac{7}{10}\right)+\frac{1}{10}\log\left(\frac{1}{10}\right)\right] = 0.5784$$

H(Y|C): conditional entropy of class labels for clustering C

- Now, consider Cluster-2:
  - P(Y=1|C=2)=2/10 (two triangles in cluster-1)
  - P(Y=2|C=2)=7/10 (seven rectangles in cluster-1)
  - P(Y=3|C=2)=1/10 (one star in cluster-1)
  - Calculate conditional entropy as:

$$H(Y|C=2) = -P(C=2) \sum_{y\in\{1,2,3\}} P(Y=y|C=2)\log(P(Y=y|C=2))$$

$$= -\frac{1}{2} \times \left[\frac{2}{10}\log\left(\frac{2}{10}\right) + \frac{7}{10}\log\left(\frac{7}{10}\right) + \frac{1}{10}\log\left(\frac{1}{10}\right)\right] = 0.5784$$

I(Y;C)

- Finally the mutual information is:

$$I(Y;C) = H(Y) - H(Y|C)$$
$$= 1.5 - [0.7855 + 0.5784]$$
$$= 0.1361$$

The NMI is therefore,

$$NMI(Y,C) = \frac{2 \times I(Y;C)}{[H(Y) + H(C)]}$$

$$NMI(Y,C) = \frac{2 \times 0.1361}{[1.5 + 1]} = 0.1089$$

Example

	P 1	P 2	P 3	P 4	P5	P6	Total
C1	3	5	40	506	96	27	677
C 2	4	7	280	29	39	2	361
C 3	1	1	1	7	4	671	685
C 4	10	162	3	119	73	2	369
C 5	331	22	5	70	13	23	464
C 6	5	358	12	212	48	13	648
total	354	555	341	943	273	738	3204

$$Purity = \frac{1}{N}\sum_{k}\max_{j} |C_k \cap P_j|$$

$$Purity = \frac{506+280+671+162+331+358}{3204}$$

$$= 0.7203$$

$$NMI = \frac{I(C,P)}{\sqrt{H(C)H(P)}}$$

$$I(C,P) = \sum_{k}\sum_{j}\frac{|C_k \cap P_j|}{N}\log\frac{N\cdot|C_k \cap P_j|}{|C_k||P_j|}$$

$$H(C) = \sum_{k}\frac{|C_k|}{N}\log\frac{|C_k|}{N}$$

$$H(P) = \sum_{j}\frac{|P_j|}{N}\log\frac{|P_j|}{N}$$

```
from sklearn.metrics.cluster import normalized_mutual_info_score
normalized_mutual_info_score([0, 0, 1, 1], [0, 0, 1, 1])
```

Output :1 (perfect labeling)

```
normalized_mutual_info_score([0, 0, 1, 1], [1, 1, 0, 0])
```

Output :1 (perfect labeling)

```
normalized_mutual_info_score([0, 0, 0, 0], [0, 1, 2, 3])
```

Output :0 (classes members are completely split across different clusters, the assignment is totally in-complete**)**

Internal Index

- "Ground truth" may be unavailable
- Use only the data to measure cluster quality
  - Measure the "*cohesion*" and "*separation*" of clusters
  - Calculate the *correlation* between clustering results and distance matrix

Correlation with Distance Matrix

- Distance Matrix
  - D_{ij} is the similarity between object O_i and O_j
- Incidence Matrix
  - C_{ij}=1 if O_i and O_j belong to the same cluster, C_{ij}=0 otherwise
- Compute the correlation between the two matrices
  - Only $n(n-1)/2$ entries needs to be calculated
- High correlation indicates good clustering

Correlation with Distance Matrix

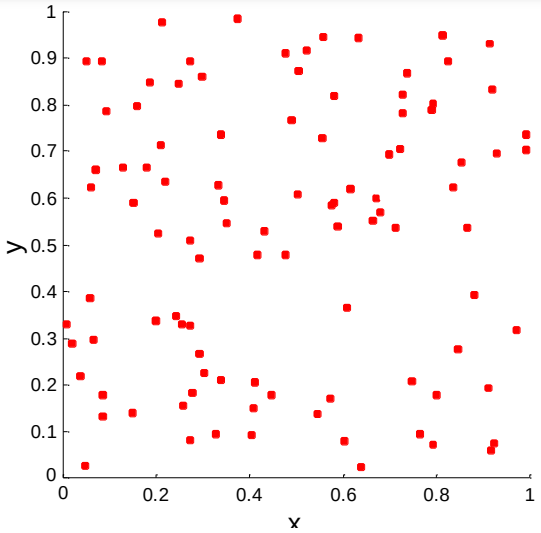
- Given Distance Matrix D = $\{d_{11}, d_{12}, \ldots, d_{nn}\}$ and Incidence Matrix $C = \{c_{11}, c_{12}, \ldots, c_{nn}\}$.

- Correlation r between D and C is given by

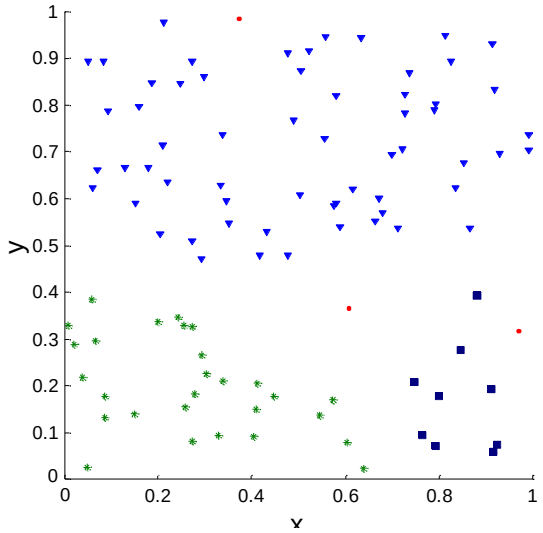
$$r = \frac{\sum_{i=1,j=1}^{n} (d_{ij} - \bar{d})(c_{ij} - \bar{c})}{\sqrt{\sum_{i=1,j=1}^{n} (d_{ij} - \bar{d})^2} \sqrt{\sum_{i=1,j=1}^{n} (c_{ij} - \bar{c})^2}}$$

Are There Clusters in the Data?

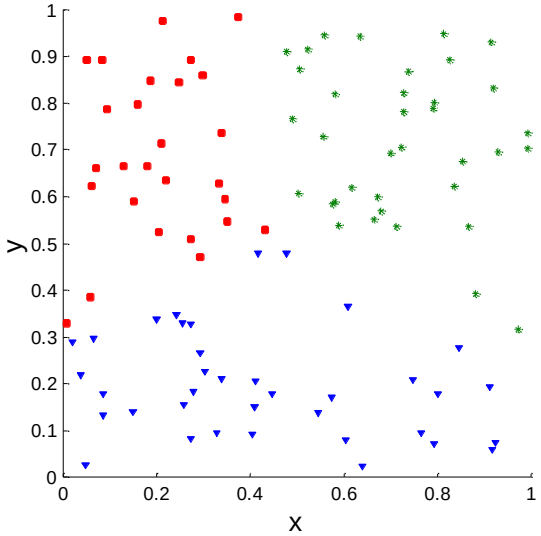
Random Points



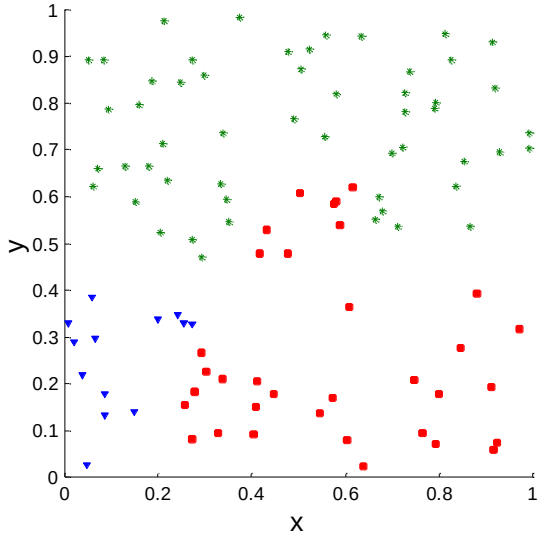
DBSCAN



K-means

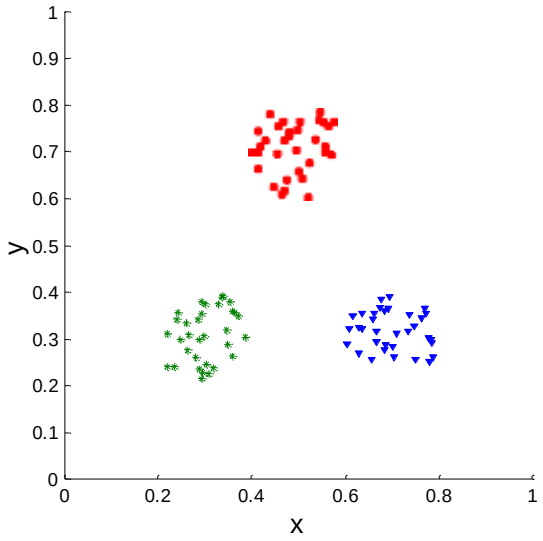


Complete Link

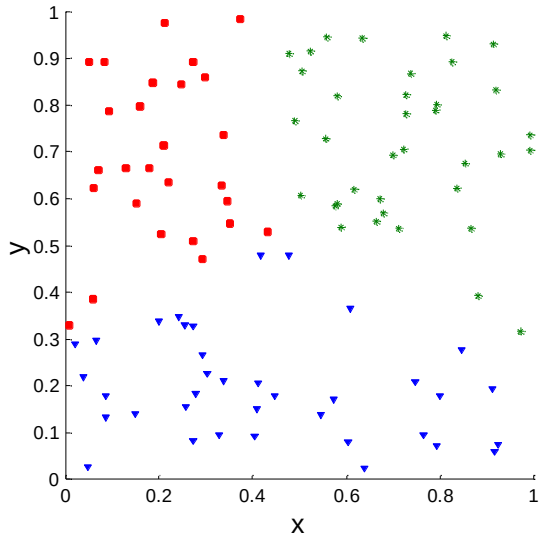


Measuring Cluster Validity Via Correlation

- Correlation of incidence and distance matrices for the K-means clustering of the following two data sets



Corr = -0.9235



Corr = -0.5810

Measuring Clustering Quality

Extrinsic	Intrinsic
Supervised	Unsupervised
The ground truth is available	The ground truth is unavailable
Compare a clustering against the ground truth using certain clustering quality measure	Evaluate the goodness of a clustering by considering how well the clusters are separated, and how compact the clusters are
Quality measure: Purity, precision and recall metrics, normalized mutual information	Quality Measure : Silhouette coefficient