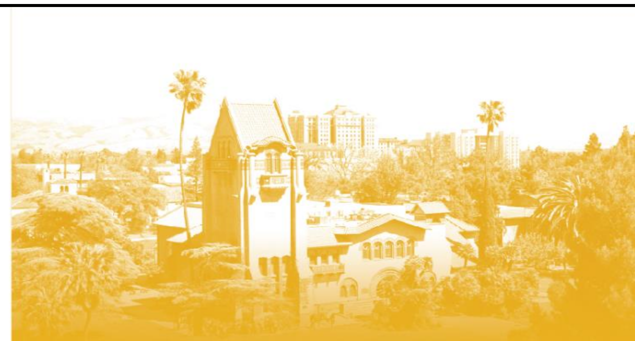


SJSU SAN JOSÉ STATE UNIVERSITY

Vector Data – Clustering

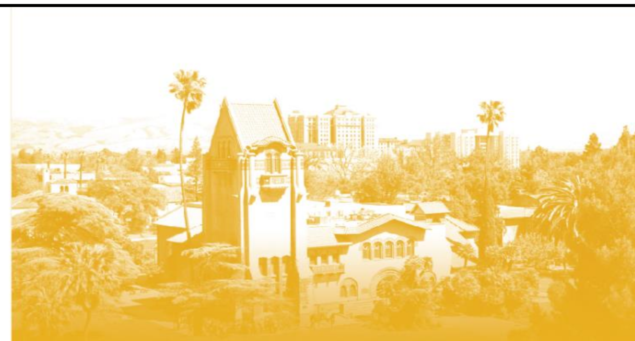
1

SJSU SAN JOSÉ STATE UNIVERSITY

Agenda

- Review of Clustering Basics
- Partitioning methods: K-Means
- Hierarchical Methods
- Cluster Validity Evaluation

2



SJSU SAN JOSÉ STATE UNIVERSITY

Review of Clustering Basics

3

SJSU SAN JOSÉ STATE UNIVERSITY

Clustering

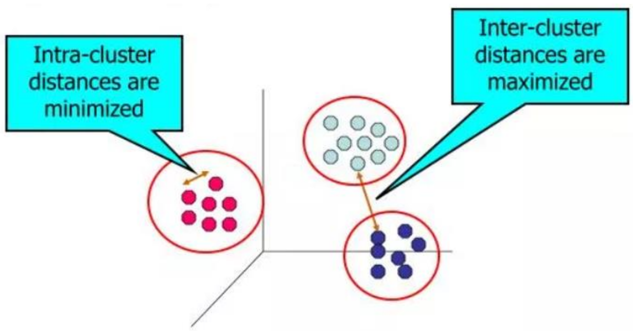
- Clustering refers to a very broad set of techniques for finding subgroups, or clusters, in a data set.
- We seek a partition of the data into distinct groups so that the observations within each group are quite similar to each other.
- To do so, we must define what it means for two or more observations to be similar or different (dissimilar).
- Very often a domain-specific consideration that must be made based on knowledge of the data being studied.

4

SJSU SAN JOSÉ STATE UNIVERSITY

What is Cluster Analysis

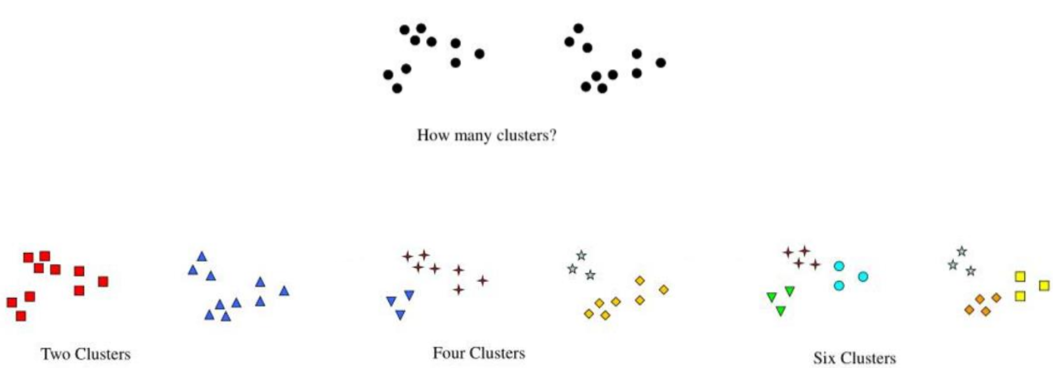
- Cluster Analysis is a statistical method used to group "similar" items based on their characteristics.



5

SJSU SAN JOSÉ STATE UNIVERSITY

Clusters Can Be Ambiguous….



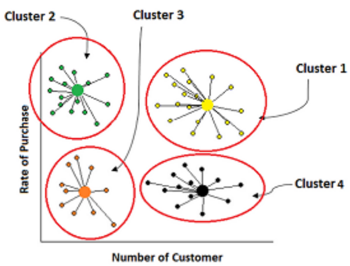
6

SJSU SAN JOSÉ STATE UNIVERSITY

Types of Clustering Methods

Partitioning Methods

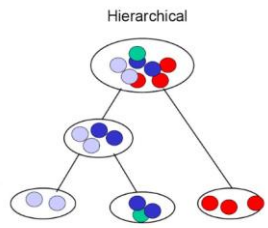
– split the data into distinct groups

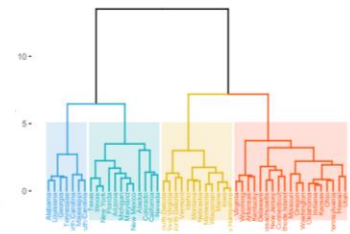


– K-Means, K-Medoids (PAM), CLARA

Hierarchical Methods

– Nested clusters organized as a hierarchical tree





– Agglomerative vs Divisive; Dendrograms

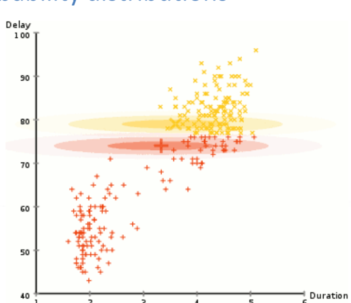
7

SJSU SAN JOSÉ STATE UNIVERSITY

Types of Clustering Methods

Model-Based Methods

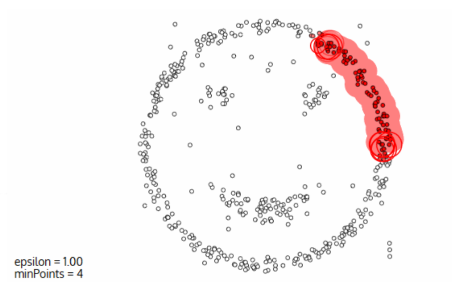
– assume data is based on a mixture of probability distributions



– Gaussian Mixture Models (GMMs), Bayesian Mixture Models, Expectation-Maximization

Density-Based Methods

– clusters based on density of data points



– Density-Based Spatial Clustering of Application with Noise (DBSCAN), OPTICS

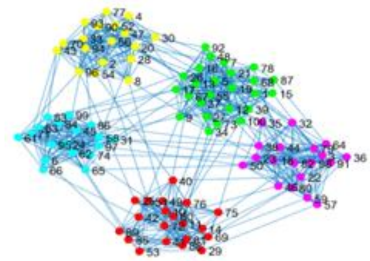
8

SJSU SAN JOSÉ STATE UNIVERSITY

Types of Clustering Methods

Graph-Based Methods

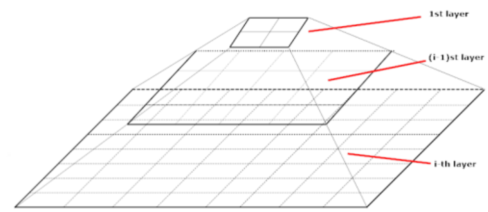
– data as graphs and clusters based on graph properties



– Spectral clustering

Grid-Based Methods

– use a multi-resolution grid data structure to process data



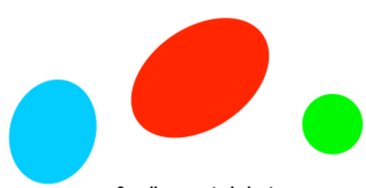
– Statistical Information Grid (STING), Clustering in Quest of the Interesting (CLIQUE)

9

SJSU SAN JOSÉ STATE UNIVERSITY

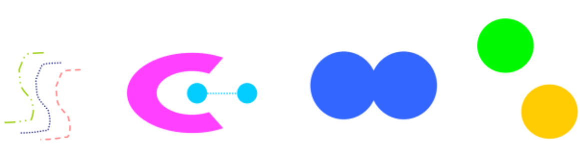
Types of Clusters

- Well-separated clusters
- Prototype-based clusters
- Contiguous clusters
- Density-based clusters

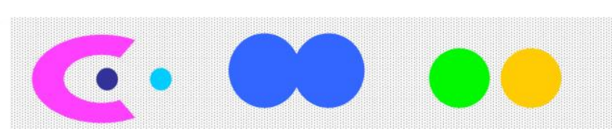


3 well-separated clusters

4 center-based clusters

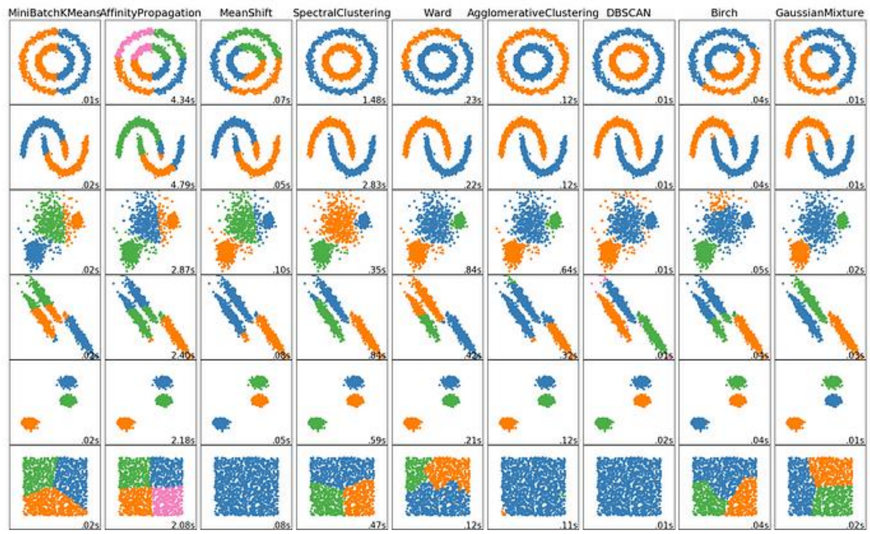


8 contiguous clusters



6 density-based clusters

10



11

SJSU SAN JOSÉ STATE UNIVERSITY

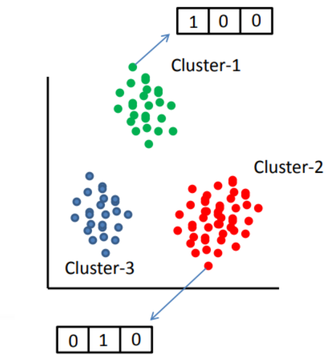
Review of K-Means

12

SJSU SAN JOSÉ STATE UNIVERSITY

K-Means Clustering

- K-Means is one of the most popular clustering algorithms.
- Input:
  - Observations/data points (N): $x_i \quad \forall\, i \in \{1, \dots, N\}$
  - # of clusters: k
- Output:
  - Cluster Assignments: w_{ij}
  - Cluster Centroids: $c_j \quad \forall\, j \in \{1, \dots, k\}$



1-of-k representation for cluster assignment.

- Objective Function: It aims to minimize the within-cluster sum of squares (WCSS):

13

SJSU SAN JOSÉ STATE UNIVERSITY

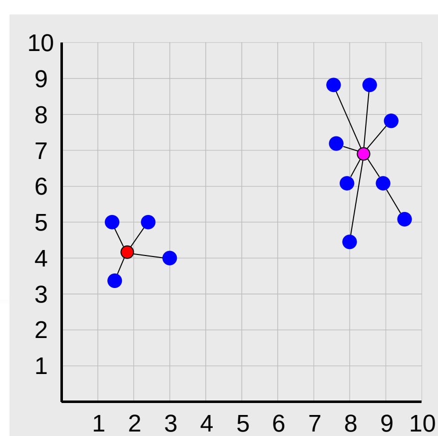
Squared Error

- Squared Error (SE):

$$SE_{C_j} = \sum_{x_i \in C_j}^{N} (x_i - c_j)^2 \qquad \text{for cluster } j$$

- Within-Cluster Sum of Squares (WCSS):

$$WCSS = \sum_{j=1}^{k} \sum_{x_i \in C_j} (x_i - c_j)^2 \qquad \text{sum over all clusters } 1, \dots, k$$



14

SJSU SAN JOSÉ STATE UNIVERSITY

K-Means Clustering

- K-Means is a minimization problem with the objective function:

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} (x_i - c_j)^2 = \sum_{j=1}^{k} \sum_{i=1}^{N} w_{ij}(x_i - c_j)^2 \qquad w_{ij} = \begin{cases} 1 & if\ x_i \in C_{ij} \\ 0 & if\ x_i \notin C_j \end{cases}$$

- Find best assignments of w_{ij} and best cluster centroids c_j ➔ minimize J
- Differentiate wrt c_k and equate to 0:

$$c_k = \frac{\sum_{i=1}^{N} w_{ik} x_i}{\sum_{i=1}^{N} w_{ij}}$$

15

SJSU SAN JOSÉ STATE UNIVERSITY

K-Means Example & Algorithm

each object is randomly assigned to a cluster

cluster centroids are computed

each object is assigned to the nearest centroid

new cluster centroids are computed

results after 10 iterations

Iterative Algorithm For K-Means

- Initialize k centroids
- Repeat till convergence
  - Calculate w_{ij} ← Expectation Step (E-Step)
  - Update $c_j = \frac{\sum_{i=1}^{N} w_{ij} x_i}{\sum_{i=1}^{N} w_{ij}}$ ← Maximization Step (M-Step)

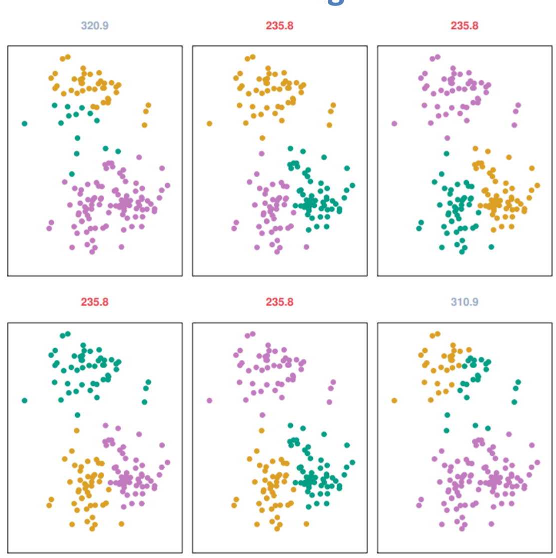
17

Comments on the K-Means Method

- Strength:
  - Efficient: *O(tkn)*, where *n*: # objects, *k*: # clusters, and *t*:# iterations. Normally, *k*, *t*<< *n*.
  - Often terminates at a local optimal
- Weakness:
  - Applicable only when mean is defined, then what about categorical data?
  - Need to specify *k*, the # of clusters, in advance the best k
  - Unable to handle noisy data and outliers well
  - Not suitable to discover clusters with non-convex shapes

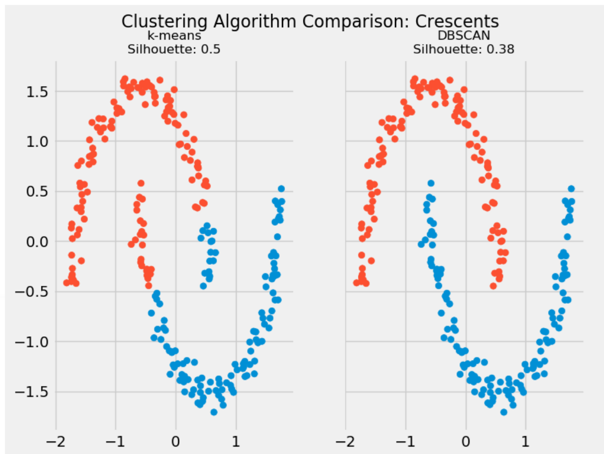
18

Different Starting Values…



19

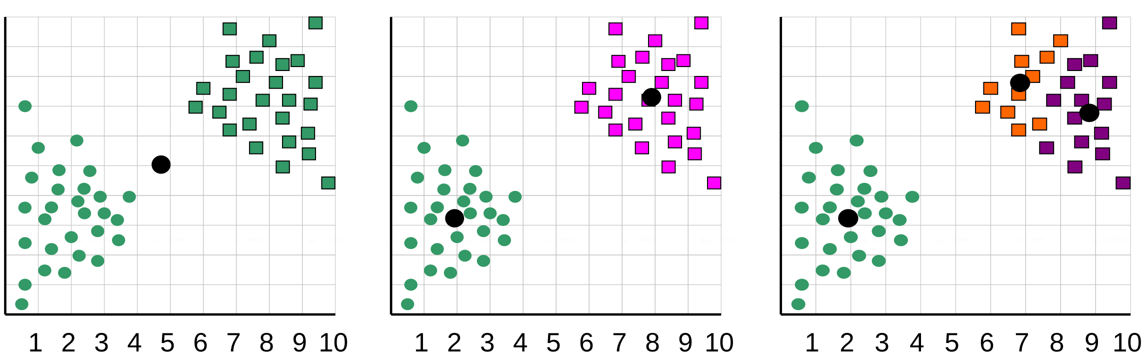
Problems with K-Means



20

Example: Determine *k*

- SE can be used to determine the # of clusters needed

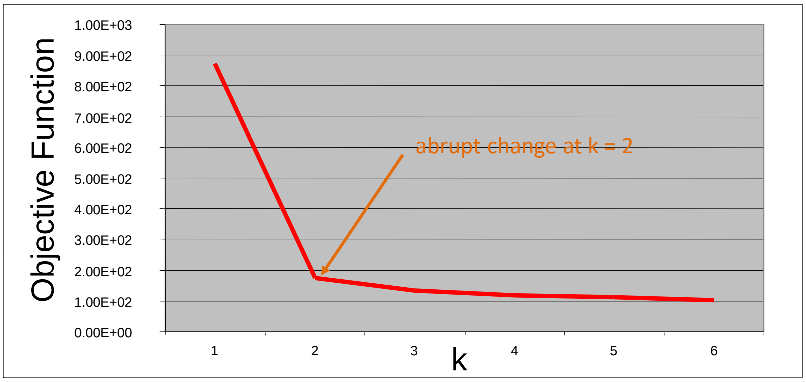


21

SJSU SAN JOSÉ STATE UNIVERSITY

Example: Determine *k* (Knee Finding)

- Plot of Objective Function vs k



22

SJSU SAN JOSÉ STATE UNIVERSITY

Variations of K-Means Method

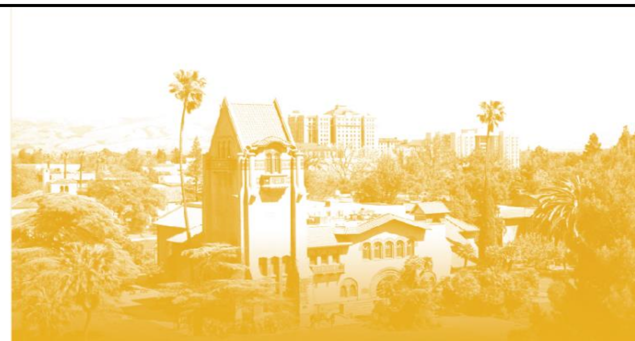
- Most of the variants of the k-means which differ in
  - Selection of the initial k-means
  - Dissimilarity calculations
  - Strategies to calculate cluster means

- Handling categorical data: k-modes
  - Replacing means of clusters with modes
  - Using new dissimilarity measures to deal with categorical objects
  - Using a frequency-based method to update modes of clusters
  - A mixture of categorical and numerical data: k-prototype method

23

SJSU SAN JOSÉ STATE UNIVERSITY

Hierarchical Clustering



24

SJSU SAN JOSÉ STATE UNIVERSITY

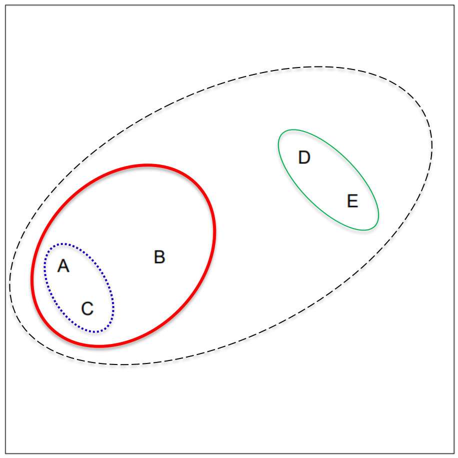
Hierarchical Clustering

- K-means clustering requires us to pre-specify the number of clusters k.
- Hierarchical Clustering is an alternative approach which does not require that we commit to a particular choice of k.

25

SJSU SAN JOSÉ STATE UNIVERSITY

A Simple Hierarchical Clustering

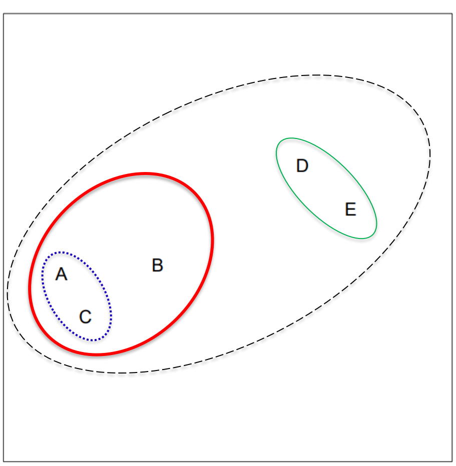


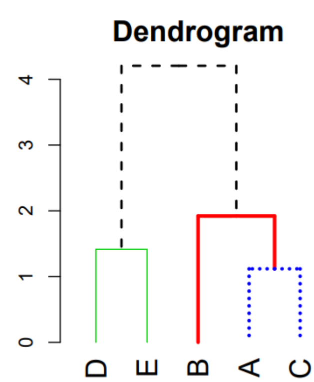
- Start with each point in its own cluster.
- Identify the closest 2 clusters & merge them.
- Repeat.
- Ends when all points are in a single cluster.

26

SJSU SAN JOSÉ STATE UNIVERSITY

Hierarchical Clustering: Dendrogram



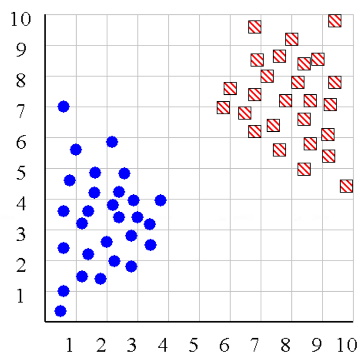


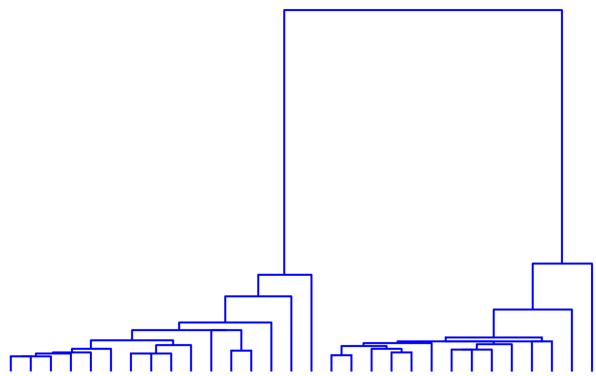
The similarity between two objects in a dendrogram is represented as the height of the lowest internal node they share.

27

Hierarchical Clustering: Dendrogram

- Dendrogram can be used to determine the “correct” number of clusters:



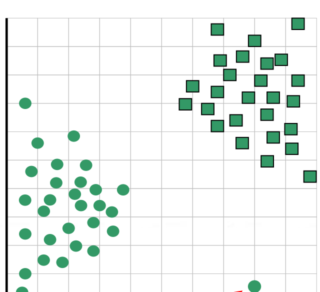


28

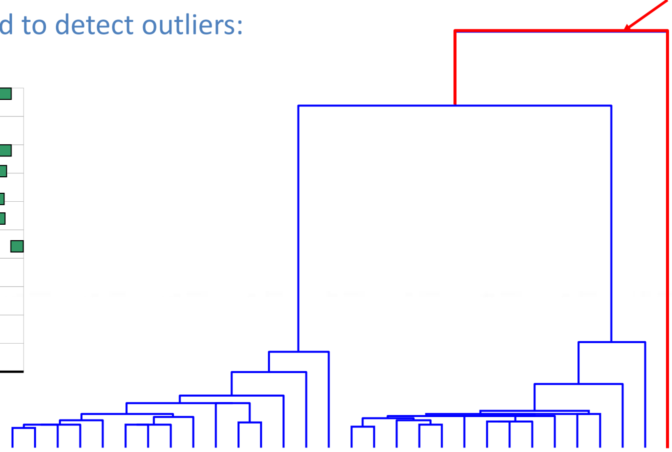
Hierarchical Clustering: Dendrogram

This single isolated branch indicates that the data point that is very different to all others

- Dendrogram can also be used to detect outliers:



Outlier

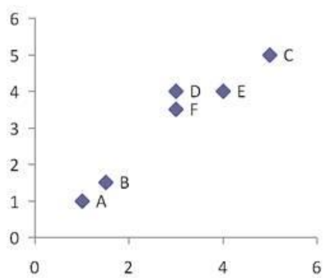


29

SJSU SAN JOSÉ STATE UNIVERSITY

Hierarchical Clustering: Distance Matrix

- Distance matrix is used as clustering criteria. It does not require the number of clusters k as an input, but needs a termination condition
- Different distance or dissimilarity metrics can be used (Euclidean etc):



	X1	X2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5

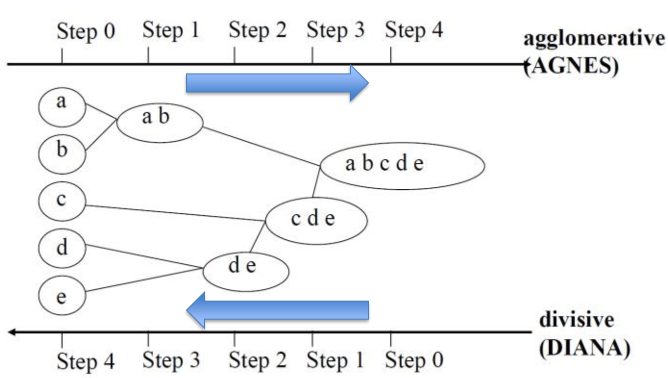
Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

30

SJSU SAN JOSÉ STATE UNIVERSITY

Hierarchical Clustering

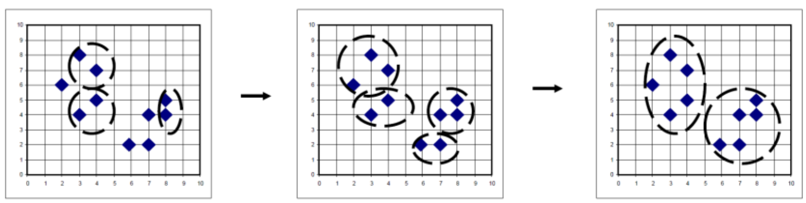
- Distance matrix is used as clustering criteria. It does not require the number of clusters k as an input, but needs a termination condition



31

AGNES (Agglomerative Nesting)

- Introduced in Kaufmann and Rousseeuw (1990)
- Use the single-link method and the dissimilarity matrix
- Merge nodes that have the least dissimilarity
- Go on in a non-descending fashion
- Eventually all nodes belong to the same cluster



32

Hierarchical Clustering: Linkage

Linkage is used to "measure" the distance between an object and a cluster. Here are some options:

- Single Linkage (nearest neighbor)
- Complete Linkage (furthest neighbor)
- Group Average Linkage
- Distance Between Centroids
- Ward's Method

33

Linkage Criteria – Single Linkage

Single or Minimum Linkage (nearest neighbors)

- Measures the distance between the closest points of 2 clusters: $D(c_1, c_2) = \min_{x_1 \in c_1, x_2 \in c_2} D(x_1, x_2)$
- Can handle elongated shapes well

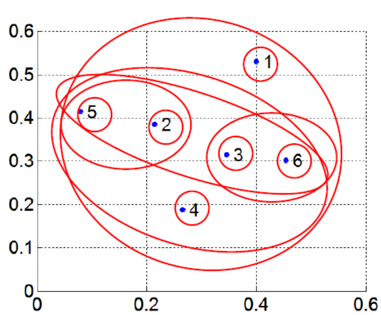
can handle non-elliptical shapes

sensitive to noise

34

Single or MIN Link

- Proximity of two clusters is based on the two closest points in the different clusters



Distance Matrix:

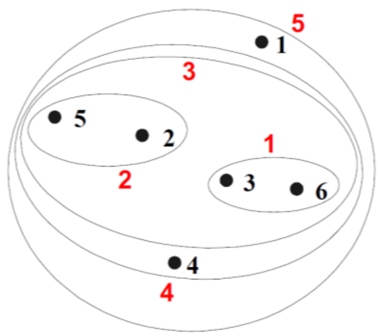
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

- D({3,6} , {2,5}) = min(D(3,2), D((6,2) , D(3,5), D(6,5)) = min(0.15, 0.25, 0.28, 0.39) = 0.15
- D({3,6} , {1 }) = min(D(3,1), D(6,1)) = min(0.22, 0.23) = 0.22
- D({3,6} , {4 }) = min(D(3,4), D(6,4)) = min(0.15, 0.22) = 0.15
- D({2,5} , {1 }) = min(D(2,1), D(5,1)) = min(0.24, 0.34) = 0.24
- D({2,5} , {4 }) = min(D(2,4), D(5,4)) = min(0.20, 0.29) = 0.20

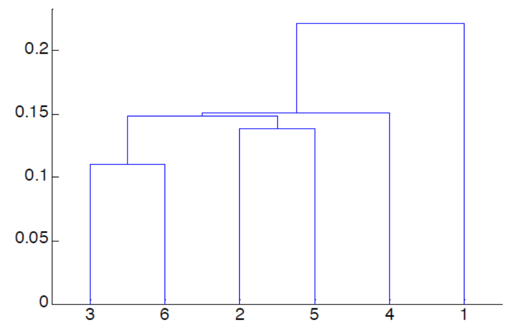
35

SJSU SAN JOSÉ STATE UNIVERSITY

Single or MIN Link



Nested Clusters



Dendrogram

36

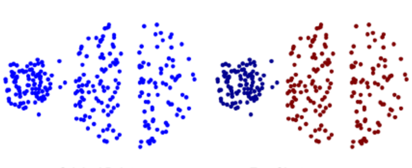
SJSU SAN JOSÉ STATE UNIVERSITY

Linkage Criteria – Complete Link

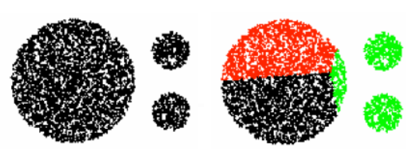
Complete or Maximum Linkage

- Measures the distance between the furthest points of 2 clusters.
- Tends to produce compact and spherical clusters.

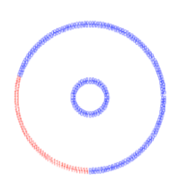
$$D(c_1, c_2) = \max_{x_1 \in c_1, x_2 \in c_2} D(x_1, x_2)$$



less susceptible to noise



tends to break large clusters



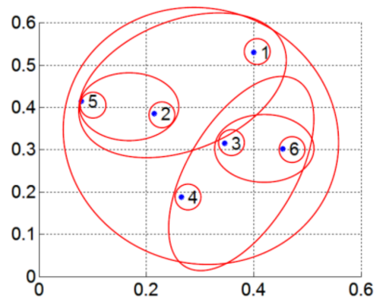
biased towards globular clusters

37

SJSU SAN JOSÉ STATE UNIVERSITY

Complete or MAX Link

- Proximity of two clusters is based on the two furthest points in the different clusters



Distance Matrix:

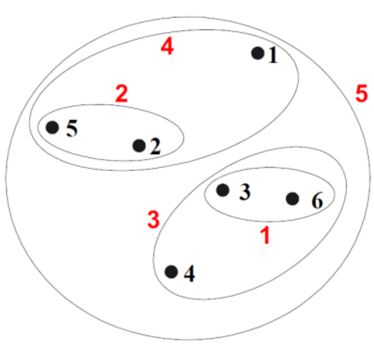
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

- D({3,6} , {2,5}) = max($D(3,2)$, $D((6,2)$, D(3,5), D(6,5)) = max(0.15, 0.25, 0.28, 0.39) = 0.39
- D({3,6} , {1 }) = max($D(3,1)$, $D(6,1)$) = max(0.22, 0.23) = 0.23
- D({3,6} , {4 }) = max($D(3,4)$, $D(6,4)$) = max(0.15, 0.22) = 0.22

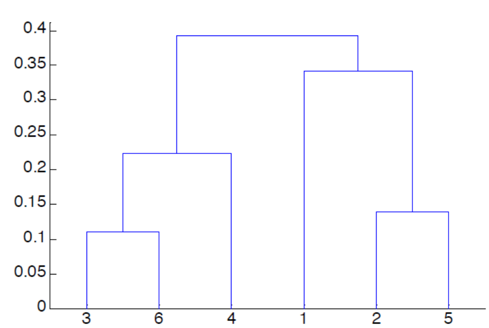
38

SJSU SAN JOSÉ STATE UNIVERSITY

Complete or MAX Link



Nested Clusters



Dendrogram

39

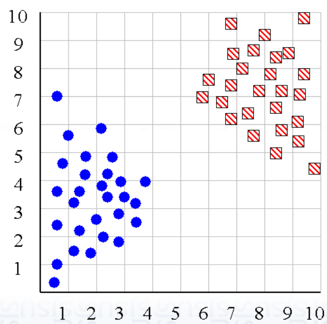
SJSU SAN JOSÉ STATE UNIVERSITY

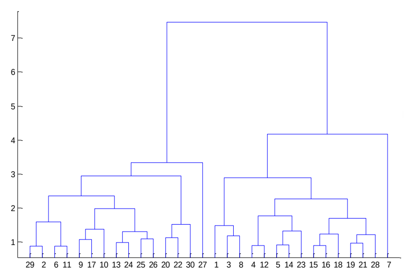
Linkage Criteria: Group Average

Group Average or Mean Linkage

- Measures the **average distance** between all points of 2 clusters.
- Balances between single and complete linkage.

$$D(c_1, c_2) = \frac{1}{|C_1|}\frac{1}{|c_2|}\sum_{x_1 \in c_1}\sum_{x_2 \in c_2} D(x_1, x_2)$$





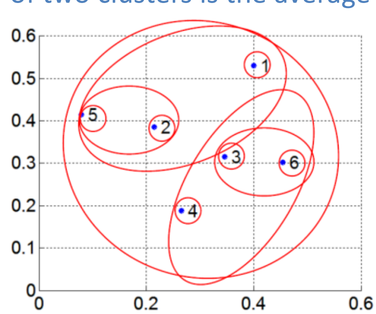
Average linkage

40

SJSU SAN JOSÉ STATE UNIVERSITY

Group Average Link

- Proximity of two clusters is the average of pairwise proximity betweenin the different clusters



Distance Matrix:

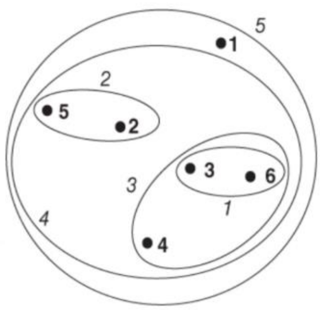
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

- D({3,6} , {2,5}) = $(D(3,2) + D((6,2) + D(3,5) + D(6,5))/2x2 = (0.15 + 0.25 + 0.28 + 0.39)/4 = 0.2675$
- D({3,6} , {1 }) $= (D(3,1) + D(6,1))/2x1 = (0.22 + 0.23)/2 = 0.225$
- D({3,6} , {4 }) $= (D(3,4) + D(6,4))/2x1 = (0.15 + 0.22)/2 = 0.185$

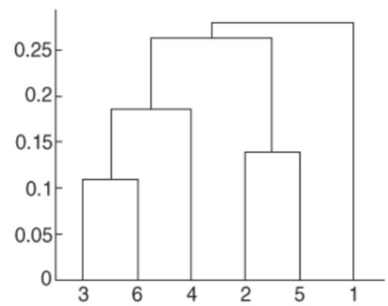
41

SJSU SAN JOSÉ STATE UNIVERSITY

Group Average Link



Nested Clusters

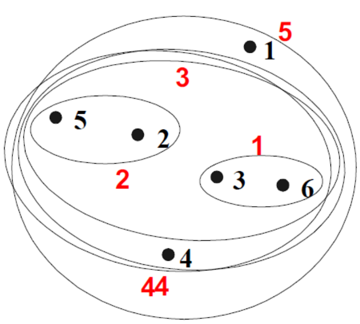


Dendrogram

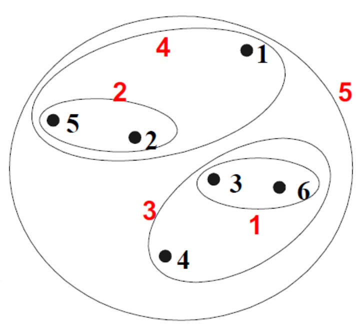
42

SJSU SAN JOSÉ STATE UNIVERSITY

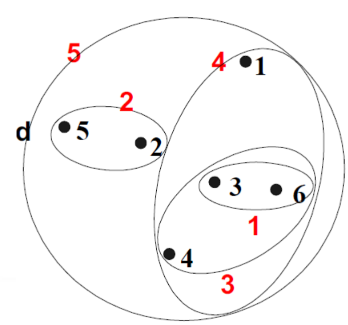
Linkage Comparison



MIN



MAX



Group Average

43

SJSU SAN JOSÉ STATE UNIVERSITY

Hierarchical Clustering: Problems and Limitations

- Computationally heavy with large datasets.
- Once a decision is made to merge or split two clusters, it cannot be undone.
- No global objective function is directly minimized.
- Different schemes have problems with one or more of the following:
  - Sensitivity to noise (MIN)
  - Difficulty handling clusters of different sizes and non-globular shapes (MAX, Group Average)
  - Breaking large clusters(MAX)

44

SJSU SAN JOSÉ STATE UNIVERSITY

Evaluation of Clustering

45

SJSU SAN JOSÉ STATE UNIVERSITY

Measures of Cluster Validity

- Measures of cluster validity can be classified as follows.
  - External Index: Measure the extent to which cluster labels match externally supplied class labels ➔ Entropy, Purity
  - Internal index: Measure the goodness of a clustering structure without respect to external information ➔ SSE
  - Relative Index: Used to compare two different clusterings or clusters.
    - Often an external or internal index is used for this function, e.g., SSE or entropy scaled dot product
- Both supervised or unsupervised measures can be used to compare clusters

46

SJSU SAN JOSÉ STATE UNIVERSITY

Unsupervised Measures: Cohesion and Separation

- Cluster Cohesion (within-cluster sum of squares, WCSS): Measure how closely data points in a cluster are to each other.
  - Lower values indicate that the points within the cluster are tightly packed.

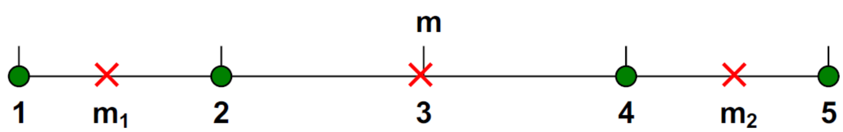
$$WCSS = \sum_{j=1}^{k} \sum_{x_i \in C_j} (x_i - c_j)^2 \quad \text{sum over all clusters } 1, \dots, k$$

- Cluster Separation (between-clusters sum of squares, BCSS): Measure how distinct or well-separated a cluster is from other clusters.
  - Higher values indicate that the clusters are well-separated from each other

$$BCSS = \sum_{j=1}^{k} |size(C_j)| (c - c_j)^2 \quad \text{sum over all clusters } 1, \dots, k$$

47

Example: Cluster Cohesion and Separation



K=1 cluster:

$$SSE = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$
$$SSB = 4\times(3-3)^2 = 0$$

K=2 clusters:

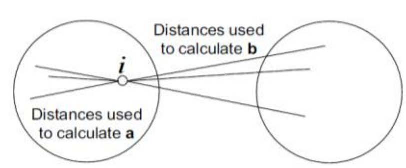
$$SSE = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$
$$SSB = 2\times(3-1.5)^2 + 2\times(4.5-3)^2 = 9$$

48

Unsupervised Measures: Silhouette Coefficient

- Silhouette coefficient combines ideas of both cohesion and separation, but for individual points, as well as clusters and clustering
- For an individual point, *i*
  - Calculate a_i = average distance of *i* to the points in its cluster
  - Calculate b_i = min (average distance of *i* to points in another cluster)
  - The silhouette coefficient for a point is then given by

$$s_i = \frac{(b_i - a_i)}{\max(a_i, b_i)}$$



- Can calculate the average silhouette coefficient for a cluster or a clustering

49

SJSU SAN JOSÉ STATE UNIVERSITY

Cluster Validity Using Correlation

- Proximity Matrix
  - D_{ij} is the similarity between object O_i and O_j
- Ideal Similarity Matrix
  - One row and one column for each data point
  - An entry is 1 if the associated pair of points belong to the same cluster
  - An entry is 0 if the associated pair of points belongs to different clusters

50

SJSU SAN JOSÉ STATE UNIVERSITY

Cluster Validity Using Correlation

- Compute the correlation between the two matrices
  - Given proximity Matrix $D = \{d_{11}, d_{12}, \ldots, d_{nn}\}$ and Incidence Matrix $C = \{c_{11}, c_{12}, \ldots, c_{nn}\}$

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

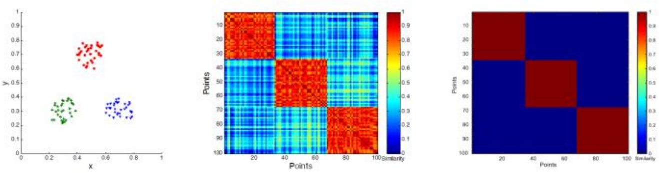
- High magnitude of correlation indicates that points that belong to the same cluster are close to each other

51

SJSU SAN JOSÉ STATE UNIVERSITY

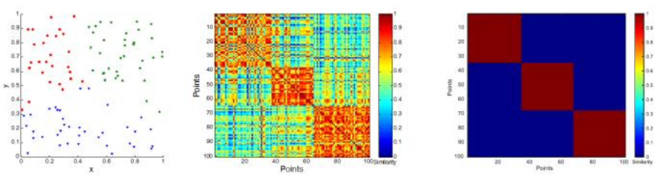
Example: Cluster Validity Using Correlation

- Correlation for a well-clustered data set:



Corr = 0.9235

- Correlation for a random data set:



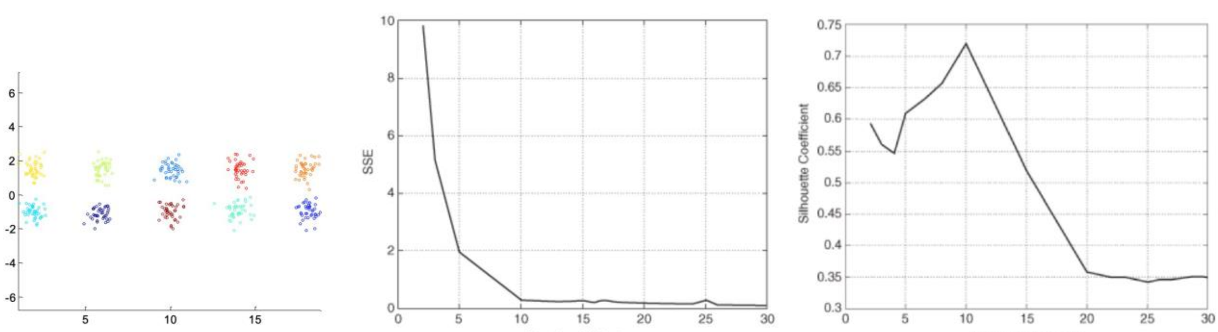
K-means

Corr = 0.5810

SJSU SAN JOSÉ STATE UNIVERSITY

Determining the Number of Clusters

- From SSE and silhouette coefficient, you can determine the (optimal?) # of clusters needed.



Supervised Measure of Cluster Validity

- Measure the degree of correspondence between the cluster labels and the class labels.
- Classification-oriented: measures from classification, such as entropy, purity, and the F-measure. These measures evaluate the extent to which a cluster contains objects of a single class.
- Similarity-oriented: measure the extent to which two objects that are in the same class are in the same cluster and vice versa

54

Classification-Oriented Measures of Cluster Validity (Entropy)

Entropy: The degree of impurity within clusters or how mixed the cluster is.

- For each cluster i, compute the class distribution of the data:

$$p_{ij} = \frac{m_{ij}}{m_i}$$

m_i # objects in cluster i,
m_{ij} # of objects of class j in cluster i.

- Calculate entropy of each cluster i:

$$e_i = -\sum_{j=1}^{L} p_{ij} \log_2 p_{ij}$$

L = # of classes

- Calculate total entropy for a set of clusters:

$$e = \sum_{i=1}^{k} \frac{m_i}{m} e_i$$

k: # of clusters, m: total # of data points

55

SJSU SAN JOSÉ STATE UNIVERSITY

Classification-Oriented Measures of Cluster Validity (Purity)

Purity: The degree to which each cluster consists of objects of a single class.

- For each cluster i, compute the class distribution of the data:

$$p_{ij} = \frac{m_{ij}}{m_i}$$

m_i # objects in cluster i,
m_{ij} # of objects of class j in cluster i.

- Calculate the purity of each cluster i

$$purity(i) = \max_j p_{ij}$$

- Calculate the overall purity:

$$purity(total) = \sum_{i=1}^{k} \frac{m_i}{m} purity(i)$$

k: # of clusters, m: total # of data points

56

SJSU SAN JOSÉ STATE UNIVERSITY

Classification-Oriented Measures of Cluster Validity (F-measure)

- Precision: The fraction of a cluster that consists of objects of a specified class.
  The precision of cluster i with respect to class j is

$$Precision(i,j) = \frac{m_{ij}}{m_i}$$

m_i # objects in cluster i,
m_{ij} # of objects of class j in cluster i.

- Recall: The extent to which a cluster contains all objects of a specified class.
  The recall of cluster i with respect to class j is:

$$Recall(i,j) = \frac{m_{ij}}{m_j}$$

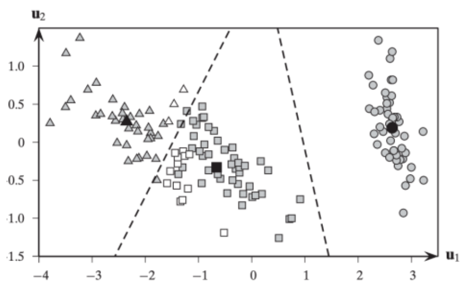
m_j # objects in cluster j,
m_{ij} # of objects of class j in cluster i.

- F-measure: A combination of both precision and recall that measures the extent to which a cluster contains only objects of a particular class and all objects of that class.
  The F-measure of cluster i with respect to class j is

$$F(i,j) = 2 \times \frac{Precision(i,j) \times Recall(i,j)}{Precision(i,j) + Recall(i,j)}$$

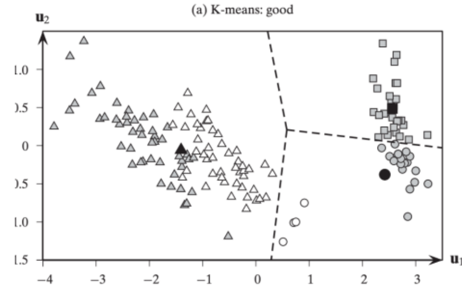
57

Example: Classification-Oriented Evaluation Measures



(a) K-means: good

(a) K-means: good



(b) K-means: bad

	iris-setosa T_1	iris-versicolor T_2	iris-virginica T_3	n_i
C_1(squares)	0	47	14	61
C_2(circles)	50	0	0	50
C_3(triangles)	0	3	36	39
m_j	50	50	50	$n = 100$

	iris-setosa T_1	iris-versicolor T_2	iris-virginica T_3	n_i
C_1	30	0	0	30
C_2	20	4	0	24
C_3	0	46	50	96
m_j	50	50	50	$n = 150$

Example: Classification-Oriented Evaluation Measures

	iris-setosa T_1	iris-versicolor T_2	iris-virginica T_3	n_i
C_1(squares)	0	47	14	61
C_2(circles)	50	0	0	50
C_3(triangles)	0	3	36	39
m_j	50	50	50	$n = 100$

$$e_i = -\Sigma_{j=1}^{L} p_{ij} \log_2 p_{ij} \qquad purity(i) = \max_j p_{ij}$$

$$e = -\Sigma_{i=1}^{k} \frac{m_i}{m} e_i \qquad purity(total) = \sum_{i=1}^{k} \frac{m_i}{m} purity(i)$$

$$e(C_1) = -(p_{11}\log_2 p_{11} + p_{12}\log_2 p_{12} + p_{13}\log_2 p_{13}) = -\left(\frac{0}{61}\log_2\frac{0}{61} + \frac{47}{61}\log_2\frac{47}{61} + \frac{14}{61}\log_2\frac{14}{61}\right) = 0.75$$

$$e(C_2) = -(p_{21}\log_2 p_{21} + p_{22}\log_2 p_{22} + p_{23}\log_2 p_{23}) = -\left(\frac{50}{50}\log_2\frac{50}{50} + \frac{0}{50}\log_2\frac{0}{50} + \frac{0}{50}\log_2\frac{0}{50}\right) = 0$$

$$e(C_3) = -(p_{31}\log_2 p_{31} + p_{32}\log_2 p_{32} + p_{33}\log_2 p_{33}) = -\left(\frac{0}{39}\log_2\frac{0}{39} + \frac{3}{39}\log_2\frac{3}{39} + \frac{36}{39}\log_2\frac{36}{39}\right) = 0.39$$

$$Purity(C_1) = max(p_{11}, p_{12}, p_{13}) = max\left(\frac{47}{61}, \frac{14}{61}, 0\right) = \frac{47}{61}$$

$$Purity(C_1) = max(p_{21}, p_{22}, p_{23}) = max\left(\frac{50}{50}, 0, 0\right) = 1$$

$$Purity(C_1) = max(p_{31}, p_{32}, p_{33}) = max\left(0, \frac{3}{39}, \frac{36}{39}\right) = \frac{36}{39}$$

$$Entropy(total) = -\left(\frac{61}{150} \times 0.75 + \frac{39}{150} \times 0.39\right) = 0.40$$

$$Purity(total) = \frac{61}{150} \times \frac{47}{61} + \frac{50}{150} \times 1 + \frac{39}{150} \times \frac{36}{39} = 0.89$$

SJSU SAN JOSÉ STATE UNIVERSITY

Example: Classification-Oriented Evaluation Measures

	iris-setosa T_1	iris-versicolor T_2	iris-virginica T_3	n_i
C_1	30	0	0	30
C_2	20	4	0	24
C_3	0	46	50	96
m_j	50	50	50	$n = 150$

$$e_i = -\Sigma_{j=1}^{L} p_{ij} \log_2 p_{ij}$$

$$e = -\Sigma_{i=1}^{k} \frac{m_i}{m} e_i$$

$$purity(i) = \max_j p_{ij}$$

$$purity(total) = \sum_{i=1}^{k} \frac{m_i}{m} purity(i)$$

$$e(C_1) = -(p_{11}\log_2 p_{11} + p_{12}\log_2 p_{12} + p_{13}\log_2 p_{13}) = -\left(\frac{30}{30}\log_2\frac{30}{30} + \frac{0}{30}\log_2\frac{0}{30} + \frac{0}{30}\log_2\frac{0}{30}\right) = 0$$

$$e(C_2) = -(p_{21}\log_2 p_{21} + p_{22}\log_2 p_{22} + p_{23}\log_2 p_{23}) = -\left(\frac{20}{24}\log_2\frac{20}{24} + \frac{4}{24}\log_2\frac{4}{24} + \frac{0}{24}\log_2\frac{0}{24}\right) = 0.65$$

$$e(C_3) = -(p_{31}\log_2 p_{31} + p_{32}\log_2 p_{32} + p_{33}\log_2 p_{33}) = -\left(\frac{0}{96}\log_2\frac{0}{96} + \frac{46}{96}\log_2\frac{46}{96} + \frac{50}{96}\log_2\frac{50}{96}\right) = 1$$

$$Purity(C_1) = max(p_{11}, p_{12}, p_{13}) = max\left(\frac{30}{30}, 0, 0\right) = 1$$

$$Purity(C_1) = max(p_{21}, p_{22}, p_{23}) = max\left(\frac{20}{24}, \frac{4}{24}, 0\right) = \frac{20}{24}$$

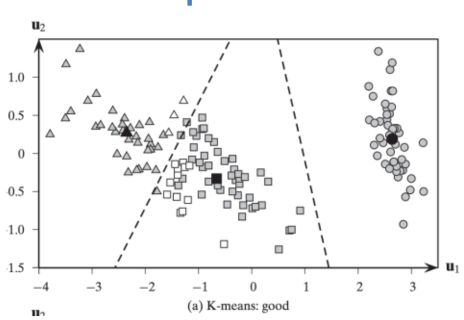
$$Purity(C_1) = max(p_{31}, p_{32}, p_{33}) = max\left(\frac{0}{96}, \frac{46}{96}, \frac{50}{96}\right) = \frac{50}{96}$$

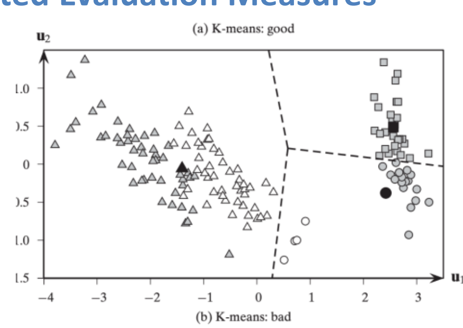
$$Entropy(total) = -\left(\frac{46}{96} \times 0.65 + \frac{50}{96} \times 1\right) = 0.74$$

$$Purity(total) = \frac{30}{150} \times 1 + \frac{24}{150} \times \frac{20}{24} + \frac{96}{150} \times \frac{50}{96} = 0.66$$

SJSU SAN JOSÉ STATE UNIVERSITY

Example: Classification-Oriented Evaluation Measures





	iris-setosa T_1	iris-versicolor T_2	iris-virginica T_3	n_i
C_1(squares)	0	47	14	61
C_2(circles)	50	0	0	50
C_3(triangles)	0	3	36	39
m_j	50	50	50	$n = 100$

$Entropy(total) = 0.40$
$Purity(total) = 0.89$
$F\ Measure = 0.88$

	iris-setosa T_1	iris-versicolor T_2	iris-virginica T_3	n_i
C_1	30	0	0	30
C_2	20	4	0	24
C_3	0	46	50	96
m_j	50	50	50	$n = 150$

$Entropy(total) = 0.74$
$Purity(total) = 0.66$
$F\ Measure = 0.65$

SJSU SAN JOSÉ STATE UNIVERSITY

Supervised Similarity-Oriented Evaluation Measure

Ideal cluster similarity matrix

- $C_{ij} = 1$ if O_i and O_j belong to the same cluster, $C_{ij} = 0$ otherwise

Class similarity matrix

- $C_{ij} = 1$ if O_i and O_j belong to the same class, $C_{ij} = 0$ otherwise

- Calculate correlation of these two matrices as the measure of cluster validity.

62

SJSU SAN JOSÉ STATE UNIVERSITY

Supervised Similarity-Oriented Evaluation measure

- Five data points, p_1, p_2, p_3, p_4, p_5, two clusters, $C_1 = \{p_1, p_2, p_3\}$ and $C_2 = \{p_4, p_5\}$, and two classes, $L_1 = \{p_1, p_2\}$ *and* $L_2 = \{p_3, p_4, p_5\}$.

Point	p_1	p_2	p_3	p_4	p_5
p_1	1	1	1	0	0
p_2	1	1	1	0	0
p_3	1	1	1	0	0
p_4	0	0	0	1	1
p_5	0	0	0	1	1

Ideal cluster similarity matrix

Point	p_1	p_2	p_3	p_4	p_5
p_1	1	1	0	0	0
p_2	1	1	0	0	0
p_3	0	0	1	1	1
p_4	0	0	1	1	1
p_5	0	0	1	1	1

Class similarity matrix

Correlation = 0.359

63

Supervised Similarity-Oriented Evaluation Measure

- Five data points, p_1, p_2, p_3, p_4, p_5, two clusters, $C_1 = \{p_1, p_2, p_3\}$ and $C_2 = \{p_4, p_5\}$, and two classes, $L_1 = \{p_1, p_2\}$ *and* $L_2 = \{p_3, p_4, p_5\}$.

	Same Cluster	Different Cluster
Same Class	f11	f10
Different Class	f01	f00

f_{00} = number of pairs of objects having a different class and a different cluster
f_{01} = number of pairs of objects having a different class and the same cluster
f_{10} = number of pairs of objects having the same class and a different cluster
f_{11} = number of pairs of objects having the same class and the same cluster

$$\text{Rand statistic} = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

$$\text{Jaccard coefficient} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

f_{00}: $\{p_1, p_4\},\{p_1, p_5\},\{p_2, p_4\}, \{p_2, p_5\} = 4$
f_{01}: $\{p_3, p_4\},\{p_3, p_5\} = 2$
f_{10}: $\{p_3, p_4\},\{p_3, p_5\} = 2$
f_{11}: $\{p_1, p'\},\{p_4, p_5\} = 2$

Rand statistic = (2 + 4)/10 = 0.6

Jaccard Coefficient = 2/6 = 0.66

64

Assessing the Significance of Cluster Validity Measures

- How to interpret a single number provided by validity measures?
- Using minimum and Maximum value:

  e.g., a purity of 0 is bad, while a purity of 1 is good Likewise, an entropy of 0 is good, as is an SSE of 0

- Use absolute standard:

  e.g., clustering for utility, we are often willing to tolerate only a certain level of error in the approximation of our points by a cluster centroid.

- Interpreting the value of our validity measure in statistical terms.

65

SJSU SAN JOSÉ STATE UNIVERSITY

Interpreting Validity Measure in Statistical Terms

- Judging how likely it is that our observed value was achieved by random chance.
  - The value is good if it is unusual; i.e., if it is unlikely to be the result of random chance.
- The motivation is that we are interested only in clusters that reflect non-random structure in the data
- Such structures should generate unusually high values of our cluster validity measures, at least if the validity measures are designed to reflect the presence of strong cluster structure.

66