Data Warehouse and Data Mining

Dhruv Gupta

dhruv.gupta@ntnu.no

28-February-2023

NTNU | Norwegian University of Science and Technology

1. Announcements and References
   - Administrative
   - References for Today's Lecture

2. Clustering
   - Density based Clustering
   - Evaluating Clustering and Clusters

1. Announcements and References
   - Administrative
   - References for Today's Lecture

2. Clustering
   - Density based Clustering
   - Evaluating Clustering and Clusters

Administrative

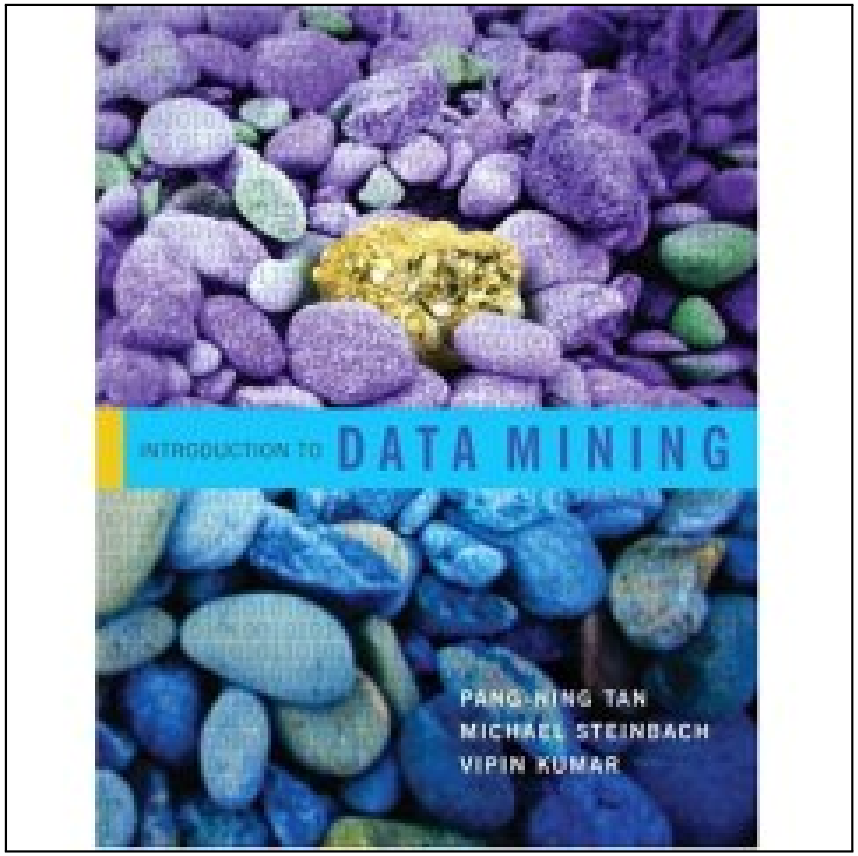
1. Third Assignment
   - Due by 09.March.2023.

1. Announcements and References
   - Administrative
   - References for Today's Lecture

2. Clustering
   - Density based Clustering
   - Evaluating Clustering and Clusters

References for "Clustering"

1. Book: Tan et al. *"Introduction to Data Mining"*, 1st Edition, 2006, Pearson Education Inc.
2. Text and images for majority of slides in "Clustering" are based on the book by Tan et al.



1. Announcements and References
   - Administrative
   - References for Today's Lecture

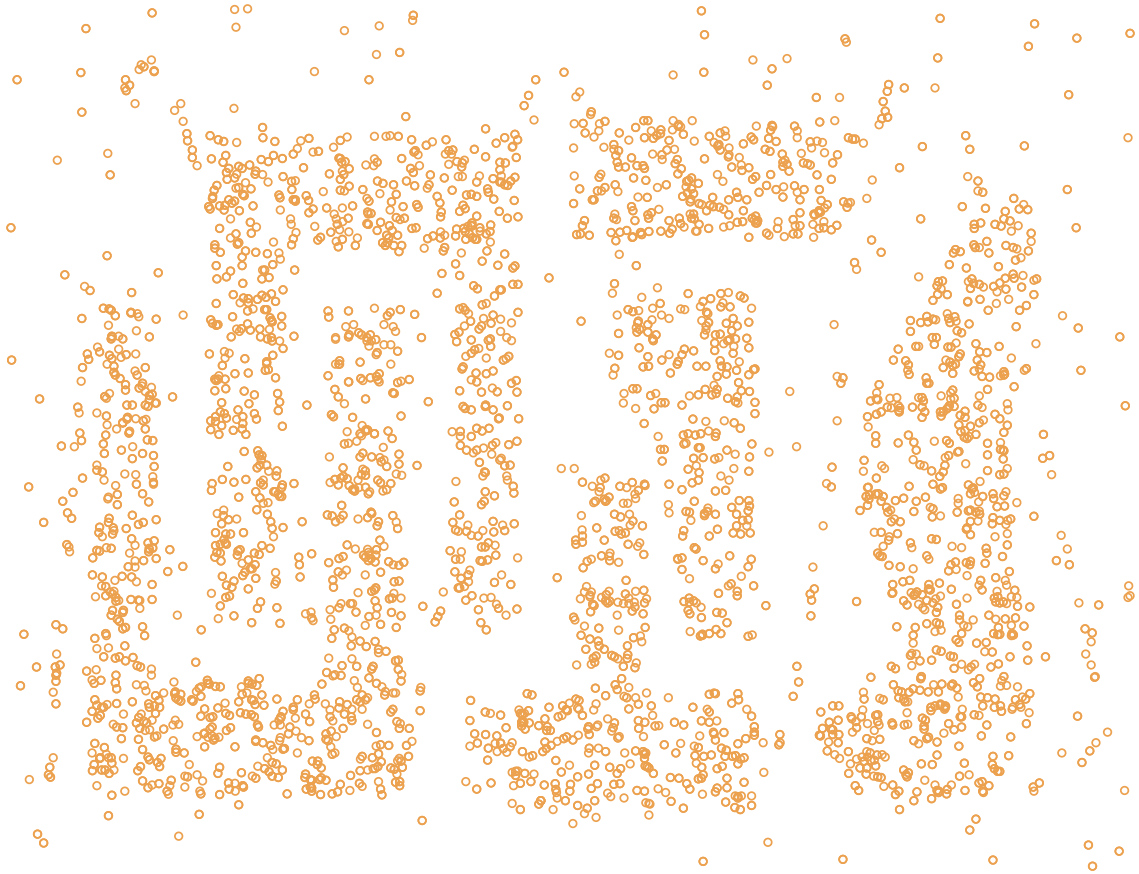
2. Clustering
   - Density based Clustering
   - Evaluating Clustering and Clusters

1. Announcements and References
   - Administrative
   - References for Today's Lecture

2. Clustering
   - Density based Clustering
   - Evaluating Clustering and Clusters

Density based Clustering

- Clusters are regions of high density that are separated from one another by regions on low density.



All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Density based Clustering — DBSCAN Algorithm

- DBSCAN is a density-based algorithm.
  - Density = number of points within a specified radius (Eps).
  - A point is a core point if it has at least a specified number of points (MinPts) within Eps.
    - These are points that are at the interior of a cluster.
    - Counts the point itself.
  - A border point is not a core point, but is in the neighborhood of a core point.
  - A noise point is any point that is not a core point or a border point.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

DBSCAN — Core, Border, and Noise Points

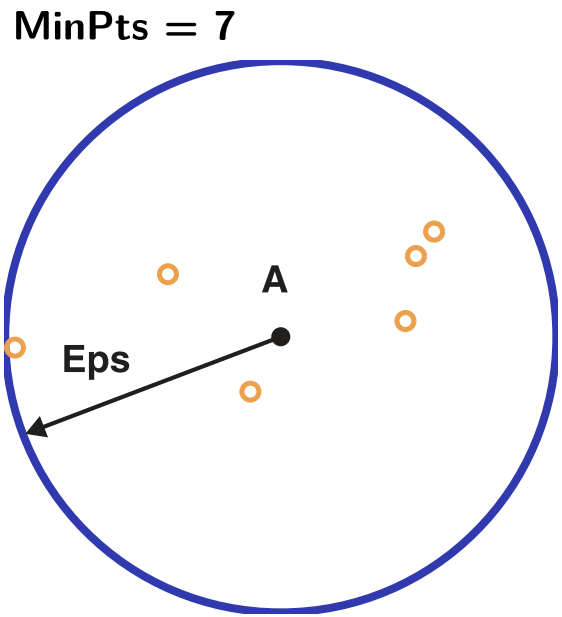


Figure 7.20. Center-based density.

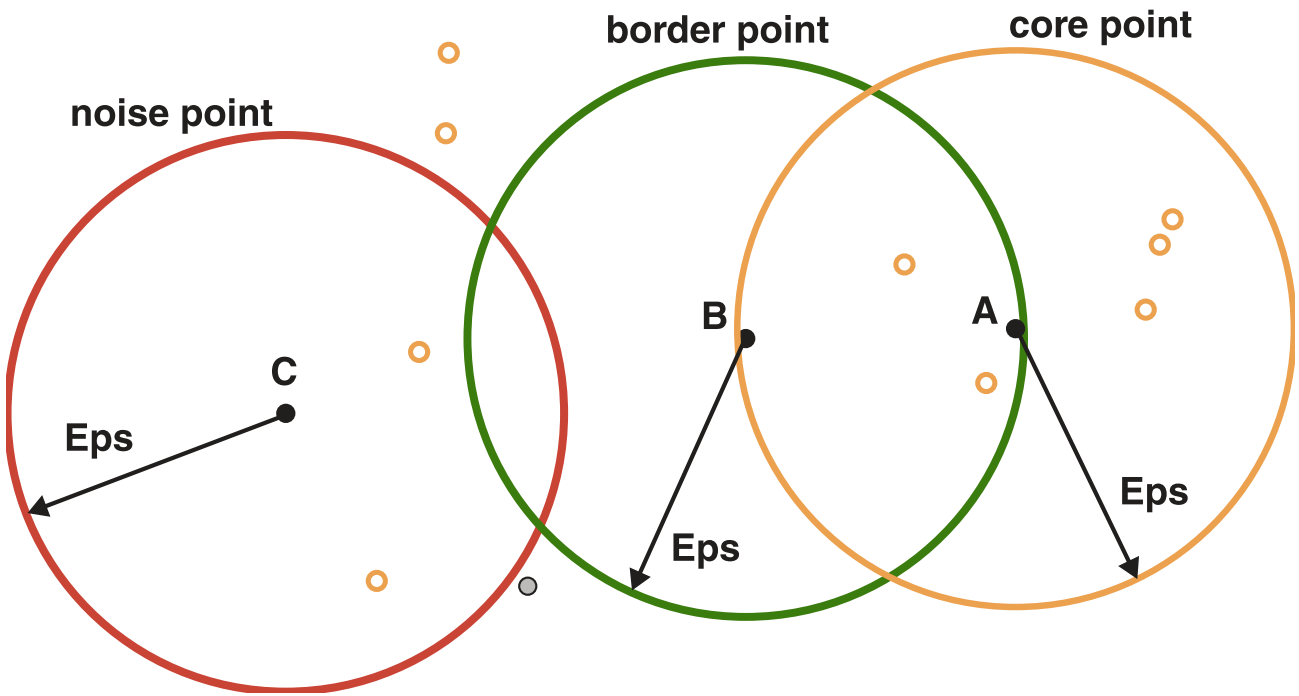
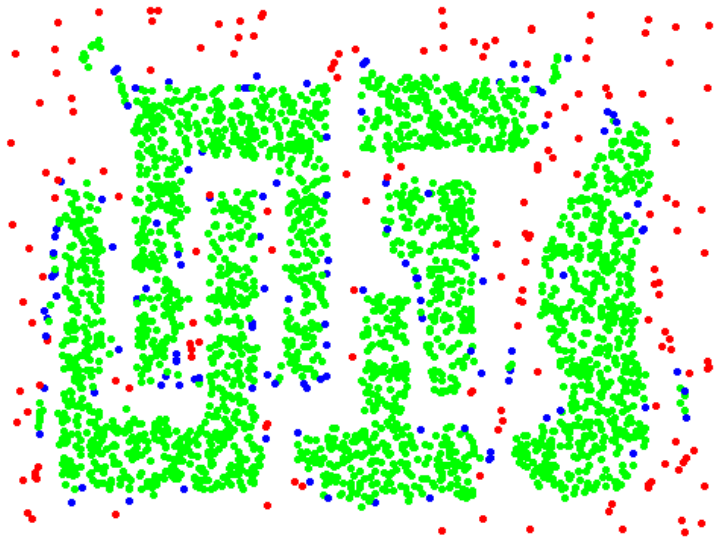


Figure 7.21. Core, border, and noise points.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

DBSCAN — Core, Border, and Noise Points

Original Points



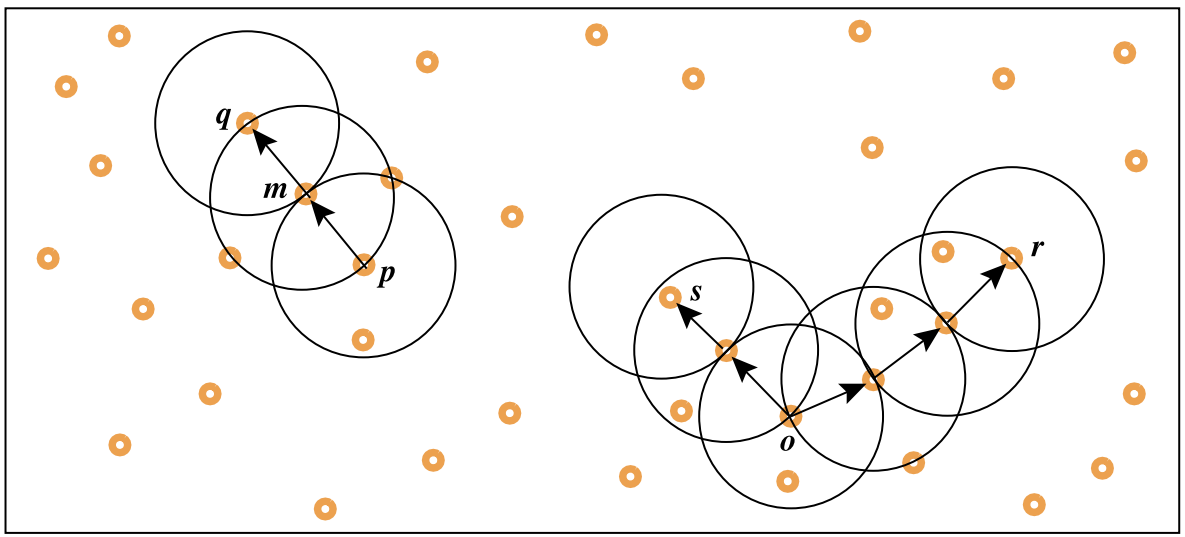
Point types: core, border and noise

Eps = 10, MinPts = 4

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

DBSCAN Algorithm

- Form clusters using core points, and assign border points to one of its neighboring clusters.
- DBSCAN Algorithm:
  1. Label all points as core, border, or noise points.
  2. Eliminate noise points.
  3. Put an edge between all core points within a distance Eps of each other.
  4. Make each group of connected core points into a separate cluster.
  5. Assign each border point to one of the clusters of its associated core points.

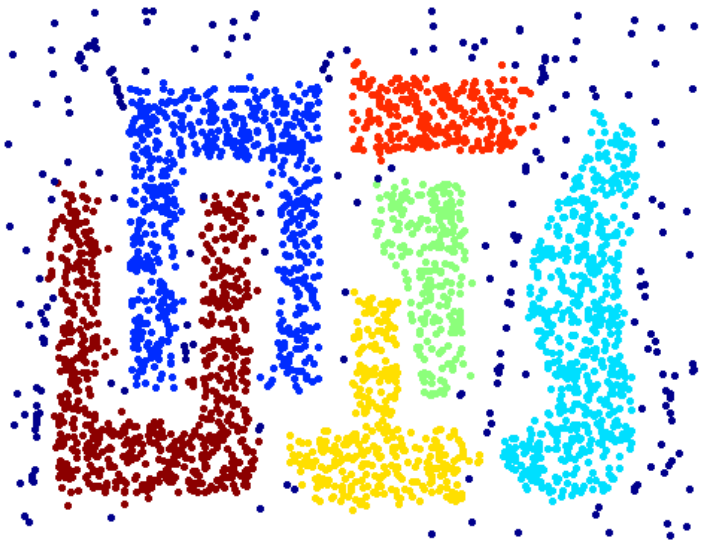


All from: 1. Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.
2. Han et al. "Data Mining Concepts and Techniques", 3rd Edition, 2012, Morgan Kaufmann Publishers.

DBSCAN— Strengths

- Resistant to noise.
- Can handle clusters of different shapes and sizes.

Original Points

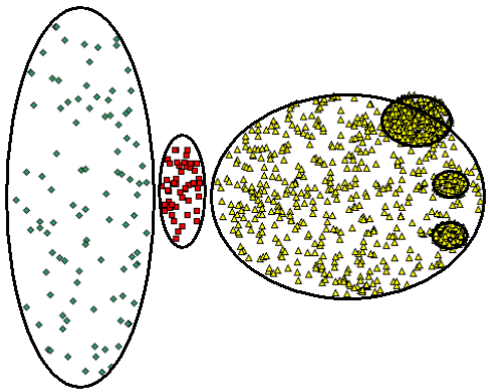


Clusters

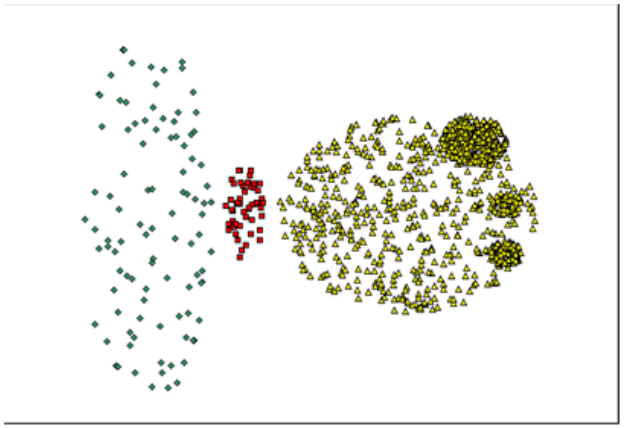
All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

DBSCAN — Weakness

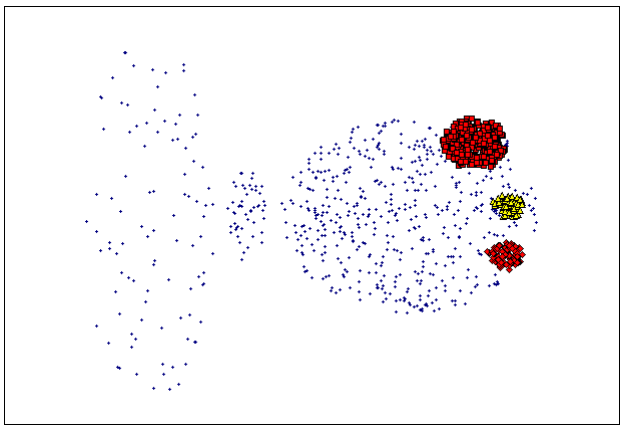
- Varying densities.
- High-dimensional data.



Original Points



(MinPts=4, Eps=9.92).



(MinPts=4, Eps=9.75)

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

DBSCAN — Determining Parameters (EPS and MinPts)

- Idea is that for points in a cluster, their k^{th} nearest neighbors are at close distance.
- Noise points have the k^{th} nearest neighbor at farther distance.
- So, plot sorted distance of every point to its k^{th} nearest neighbor.

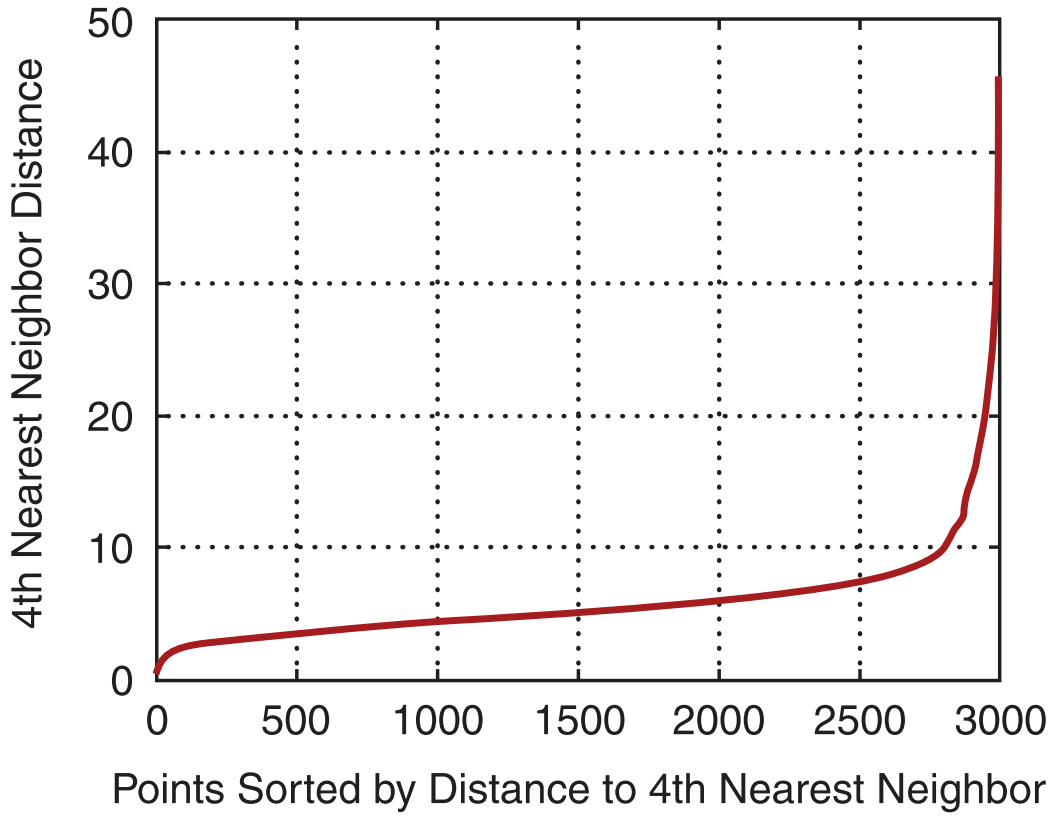


Figure 7.23. K-dist plot for sample data.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

DBSCAN — Time and Space Complexity

- Time complexity: $\mathcal{O}(m \cdot$ time to find points in the Eps-neighborhood), where m is the number of points.
- Worst case, this complexity is $\mathcal{O}(m^2)$.
- Using indexes time complexity can be as low as $\mathcal{O}(m \cdot \log(m))$.
- Space requirement is $\mathcal{O}(m)$ only a small amount of data for each point is needed (i.e., the cluster label and the identification of each point as a core, border, or noise point).

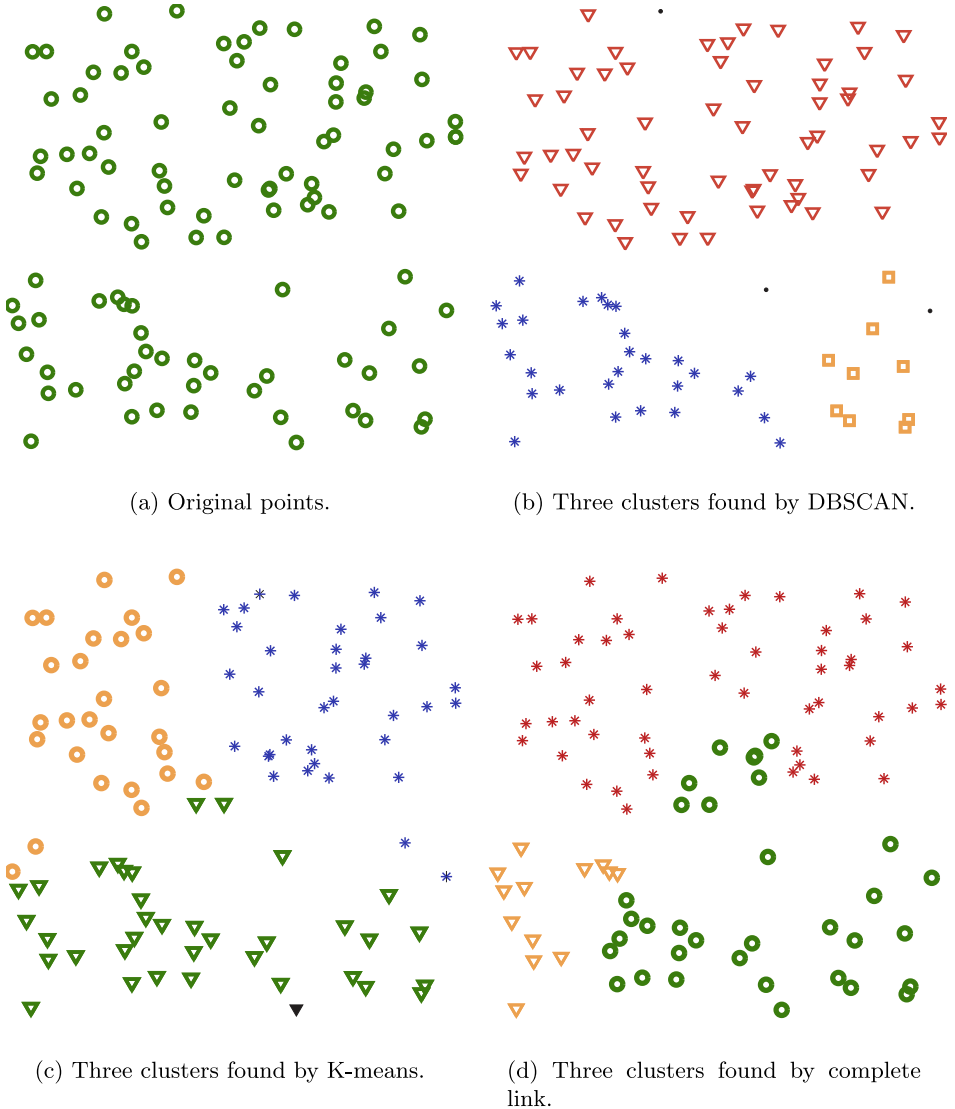
All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Evaluation — Cluster Validity

- For supervised classification we have a variety of measures to evaluate how good our model is: accuracy, precision, recall.
- For cluster analysis, the analogous question is how to evaluate the "goodness" of the resulting clusters?
- But "clusters are in the eye of the beholder"!
- Then why do we want to evaluate them?
  - To avoid finding patterns in noise.
  - To compare clustering algorithms.
  - To compare two sets of clusters.
  - To compare two clusters.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Cluster that can be found in Random Data



(a) Original points.

(b) Three clusters found by DBSCAN.

(c) Three clusters found by K-means.

(d) Three clusters found by complete link.

Figure 7.26. Clustering of 100 uniformly distributed points.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Different Aspects of Cluster Validation

1. Determining the clustering tendency of a set of data, i.e., distinguishing whether non-random structure actually exists in the data.
2. Comparing the results of a cluster analysis to externally known results, e.g., to externally given class labels.
3. Evaluating how well the results of a cluster analysis fit the data without reference to external information.
   - Use only the data.
4. Comparing the results of two different sets of cluster analyses to determine which is better.
5. Determining the 'correct' number of clusters.

- For 2, 3, and 4, we can further distinguish whether we want to evaluate the entire clustering or just individual clusters.

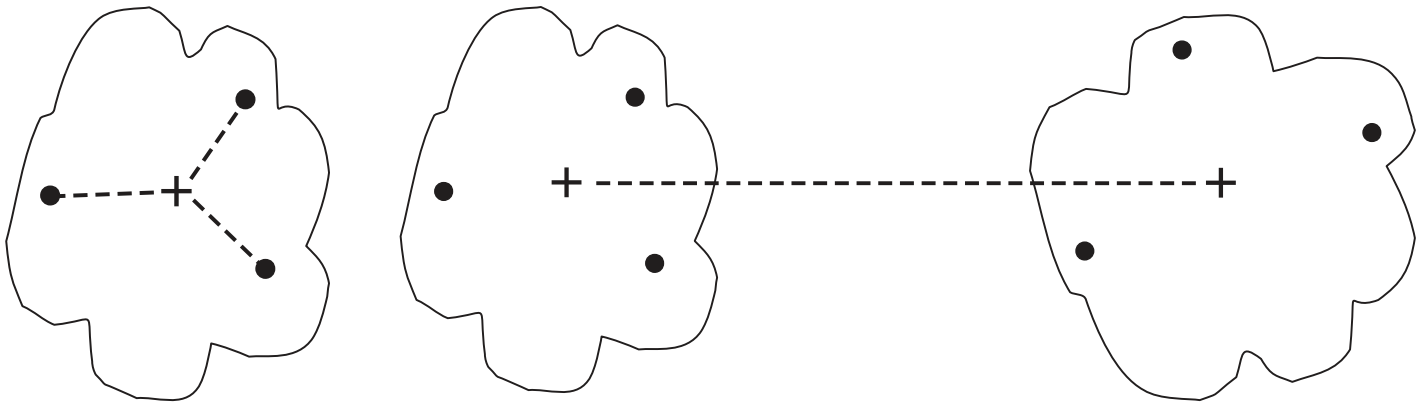
All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Measures of Cluster Validity

- Numerical measures to determine cluster validity, are of three types.
- Internal Index (Unsupervised): Used to measure the goodness of a clustering structure without respect to external information.
  - Sum of Squared Error (SSE)
- External Index (Supervised): Used to measure the extent to which cluster labels match externally supplied class labels.
  - Entropy
- 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.
- Sometimes these are referred to as criteria instead of indices.
  - However, sometimes criterion is the general strategy and index is the numerical measure that implements the criterion.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Unsupervised Cluster Evaluation: Cohesion and Separation



(a) Cohesion. (b) Separation.

Figure 7.28. Prototype-based view of cluster cohesion and separation.

- Cluster Cohesion: Measures how closely related are objects in a cluster.
  - Example: SSE.
- Cluster Separation: Measure how distinct or well-separated a cluster is from other clusters.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Unsupervised Cluster Evaluation: Cohesion and Separation

- Example: Squared Error
  - Cohesion is measured by the within cluster sum of squares (SSE):

$$\text{SSE} = \sum_i \sum_{x \in C_i} (x - m_i)^2. \tag{1}$$

  - Separation is measured by the between cluster sum of squares.

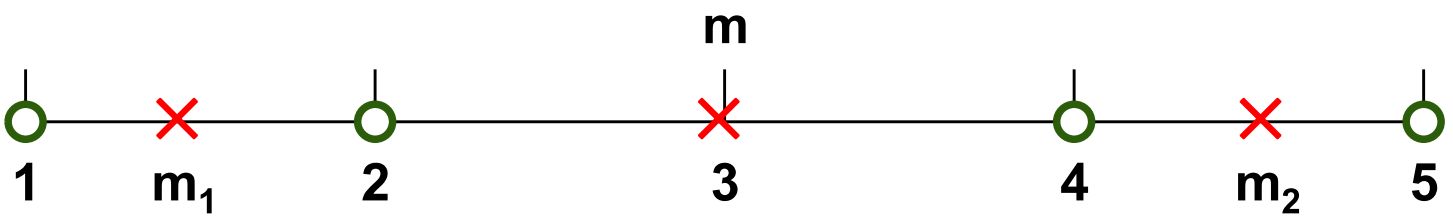
$$\text{SSB} = \sum_i |C_i| (m - m_i)^2, \tag{2}$$

  - where, $|C_i|$ is the size of cluster i.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Unsupervised Cluster Evaluation: Cohesion and Separation

- Example: SSE
  - SSB + SSE = constant.



K=1 cluster:

$$SSE = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$

$$SSB = 4 \times (3-3)^2 = 0$$

$$Total = 10 + 0 = 10$$

K=2 clusters:

$$SSE = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$

$$SSB = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$$

$$Total = 1 + 9 = 10$$

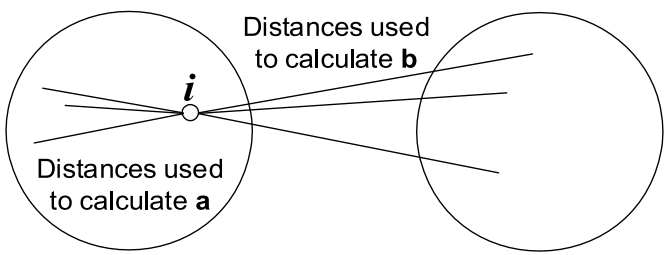
All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Unsupervised Cluster Evaluation: Silhouette Coefficient

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

$$s = \frac{(b - a)}{max(a, b)} \tag{3}$$

  - Value can vary between -1 and 1
  - Typically ranges between 0 and 1.
  - The closer to 1 the better.
- Can calculate the average silhouette coefficient for a cluster or a clustering.



All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Unsupervised Cluster Evaluation: Silhouette Coefficient

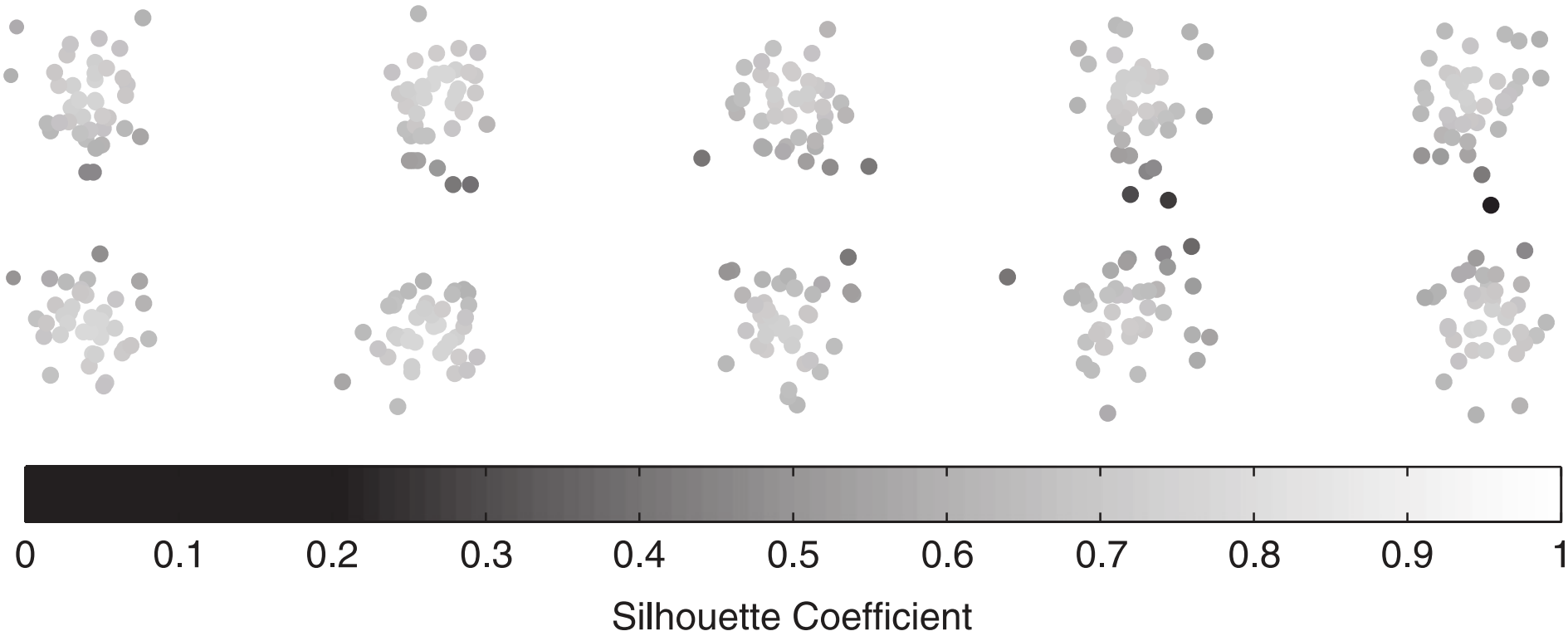


Figure 7.29. Silhouette coefficients for points in ten clusters.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Unsupervised Cluster Evaluation: Correlation

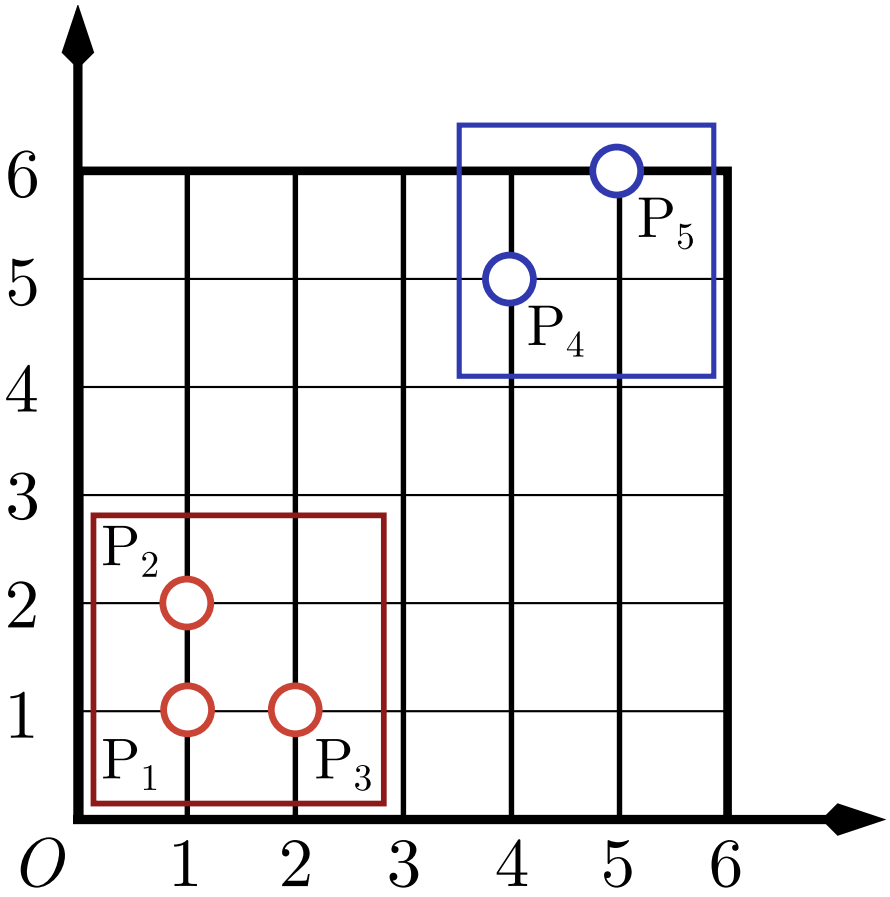
- Two matrices:
  1. Proximity Matrix
     - A similarity matrix can also be obtained by transforming / normalizing the distances using the formula:

$$s = 1 - \frac{d - d_{min}}{d_{max} - d_{min}} \tag{4}$$

  2. 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.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Unsupervised Cluster Evaluation: Correlation



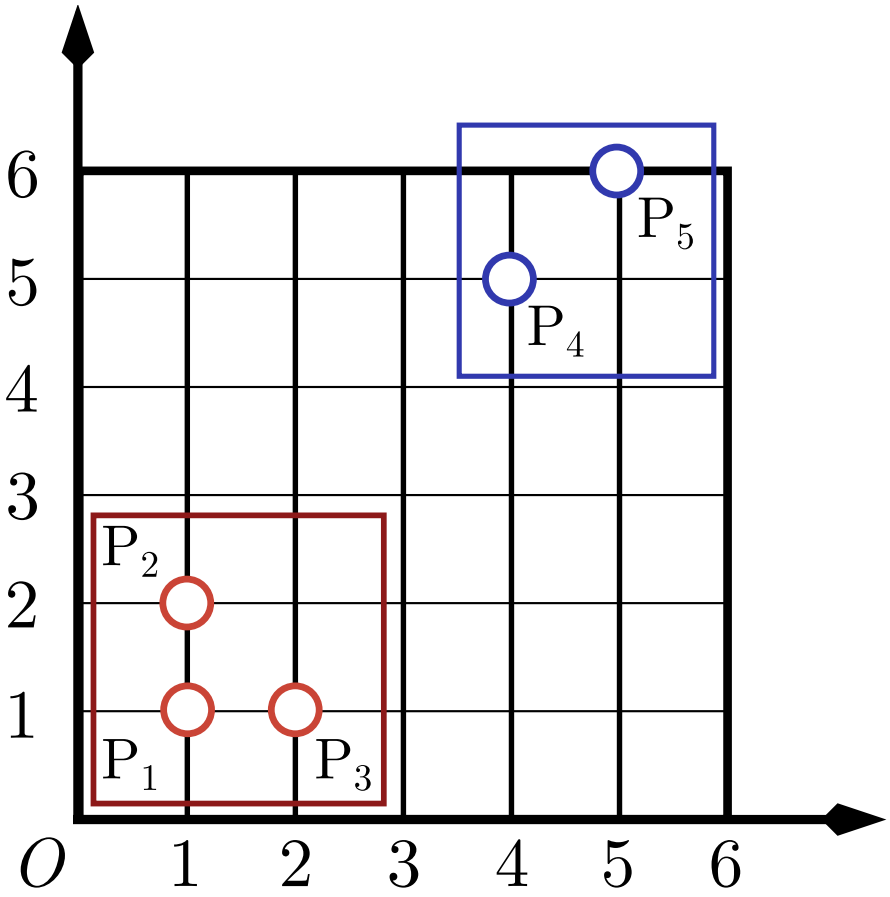
P	X	Y
1	1	1
2	1	2
3	2	1
4	4	5
5	5	6

	1	2	3	4	5
1	1	1	1	0	0
2	1	1	1	0	0
3	1	1	1	0	0
4	0	0	0	1	1
5	0	0	0	1	1

Ideal Similarity Matrix

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Unsupervised Cluster Evaluation: Correlation



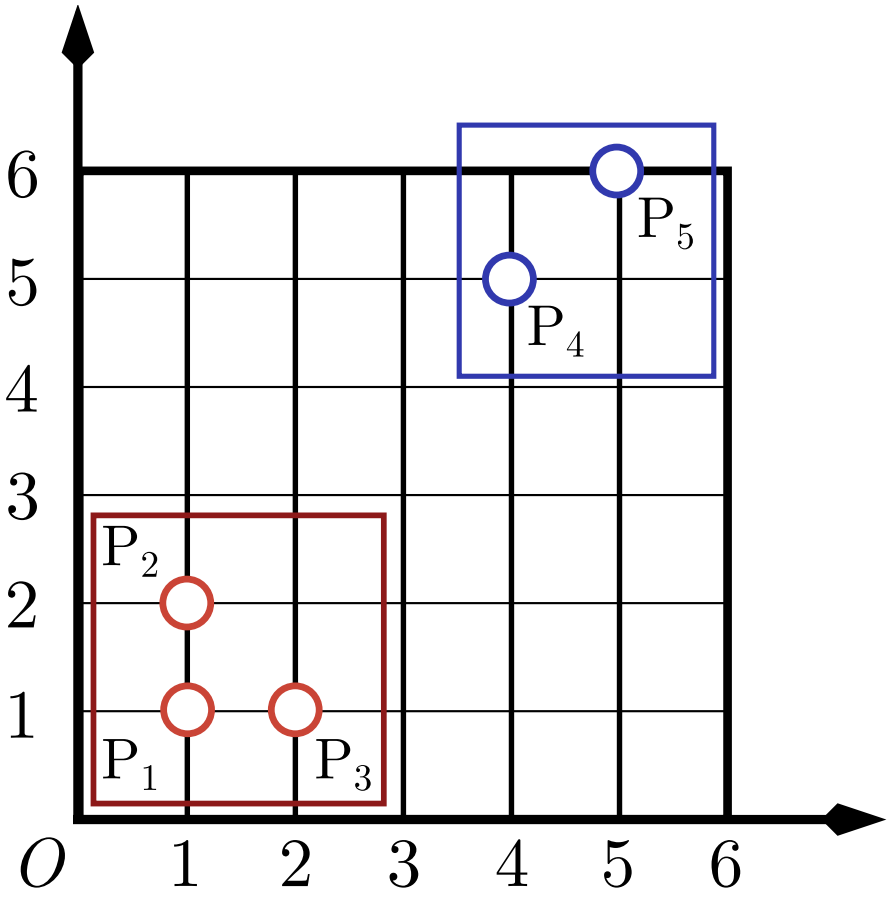
P	X	Y
1	1	1
2	1	2
3	2	1
4	4	5
5	5	6

	1	2	3	4	5
1	0	1	1	5	6.4
2	1	0	1.4	4.2	5.7
3	1	1.4	0	4.5	5.8
4	5	4.2	4.5	0	1.4
5	6.4	5.7	5.8	1.4	0

Proximity Matrix

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Unsupervised Cluster Evaluation: Correlation



P	X	Y
1	1	1
2	1	2
3	2	1
4	4	5
5	5	6

	1	2	3	4	5
1	1	0.8	0.8	0.2	0
2	0.8	1	0.8	0.3	0.1
3	0.8	0.8	1	0.3	0.1
4	0.2	0.3	0.3	1	0.8
5	0	0.1	0.1	0.8	1

Similarity Matrix

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

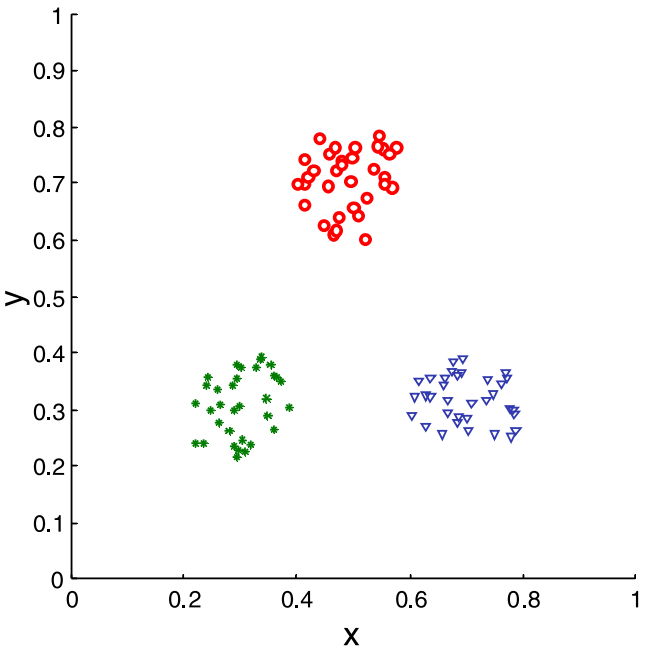
Unsupervised Cluster Evaluation: Correlation

- Compute the correlation between the two matrices.
  - Since the matrices are symmetric, only the correlation between $\frac{n \cdot (n-1)}{2}$ entries needs to be calculated.
- High magnitude of correlation indicates that points that belong to the same cluster are close to each other.
  - Correlation may be positive or negative depending on whether the similarity matrix is a similarity or dissimilarity matrix.
- Not a good measure for some density or contiguity based clusters.

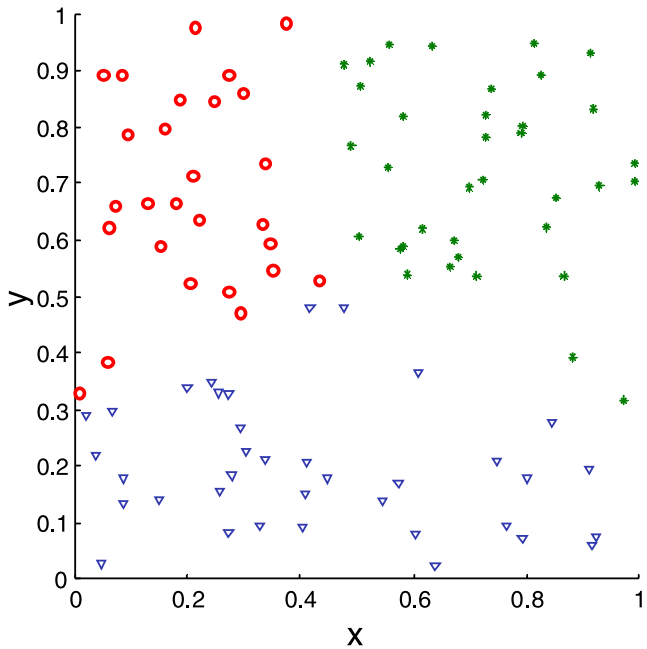
All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Unsupervised Cluster Evaluation: Correlation

- Correlation of ideal similarity and proximity matrices for the K-means clusterings of the following two data sets.



Corr = -0.9235

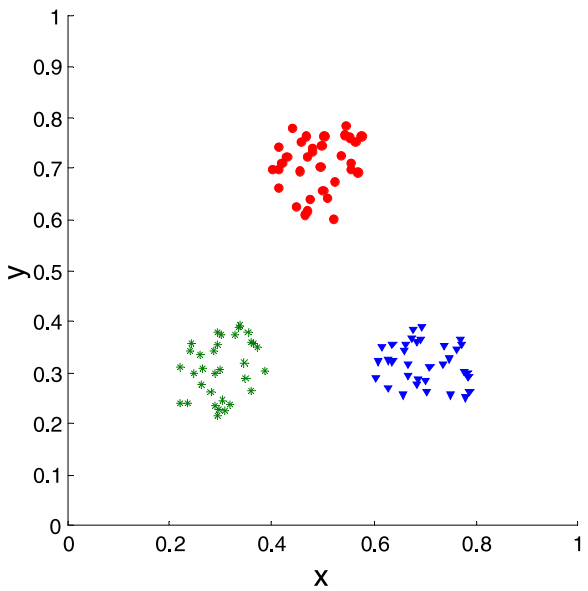


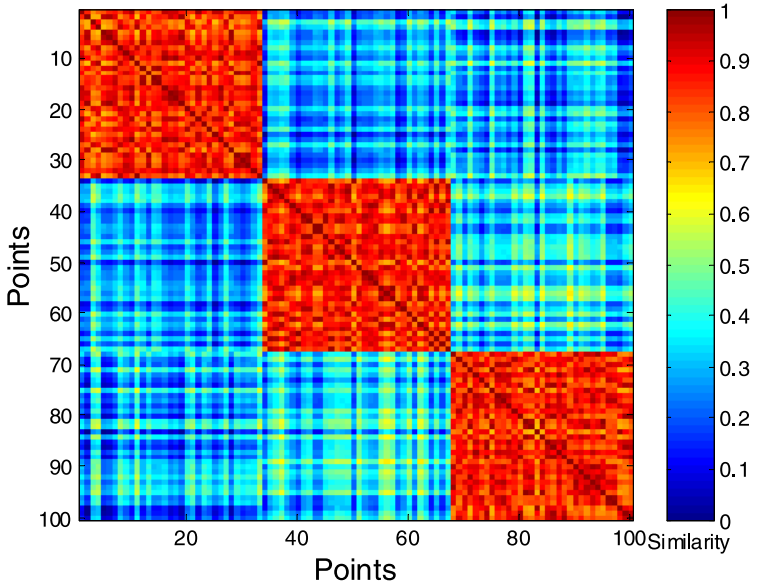
Corr = -0.5810

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Unsupervised Cluster Evaluation: Similarity Matrix

- Order the similarity matrix with respect to cluster labels and inspect visually.

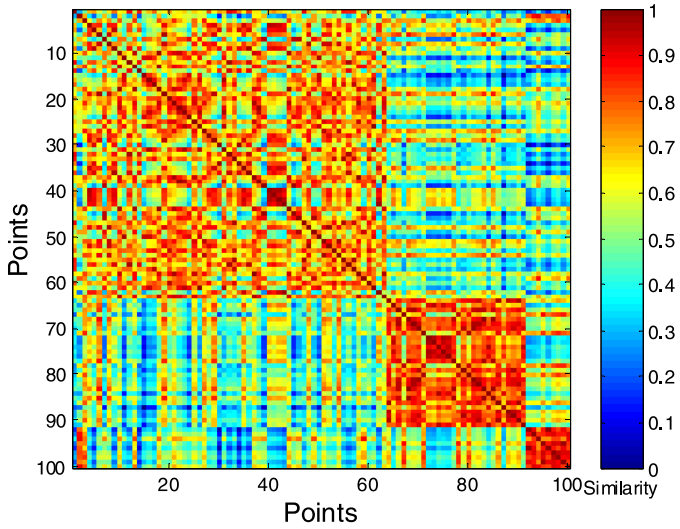


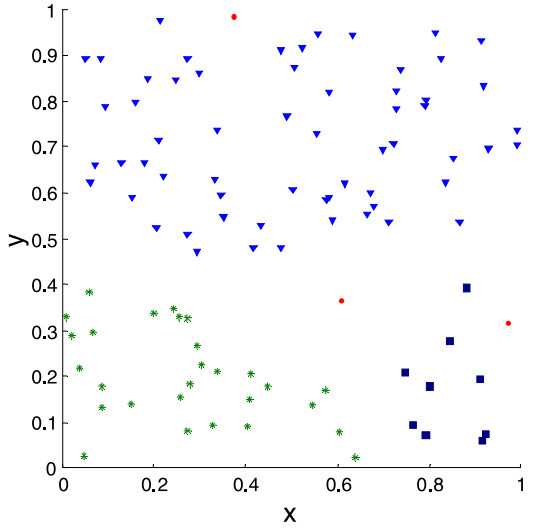


All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Unsupervised Cluster Evaluation: Similarity Matrix

- Clusters in random data are not so crisp.



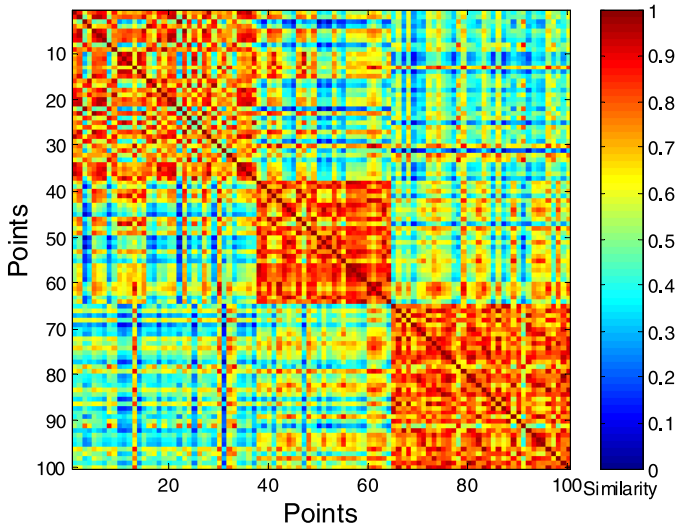


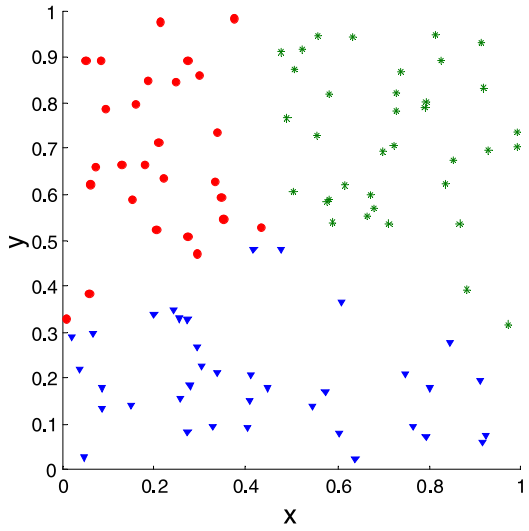
DBSCAN

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Unsupervised Cluster Evaluation: Similarity Matrix

- Clusters in random data are not so crisp.



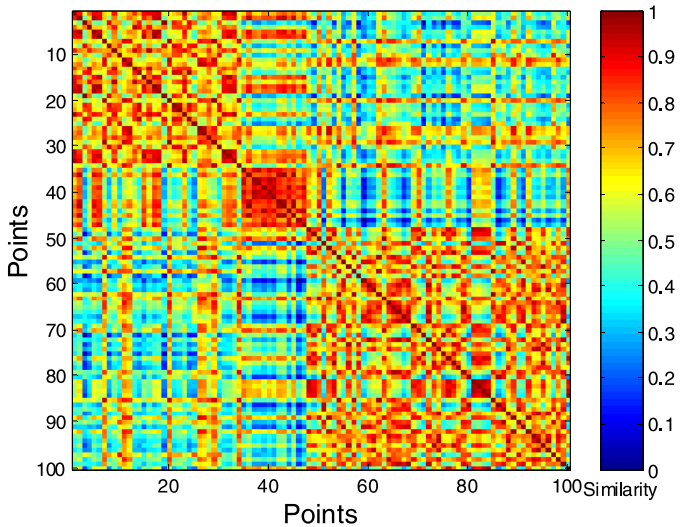


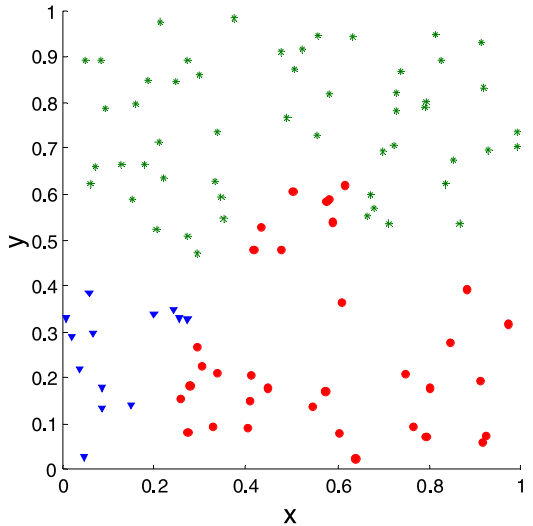
K-means

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Unsupervised Cluster Evaluation: Similarity Matrix

- Clusters in random data are not so crisp.

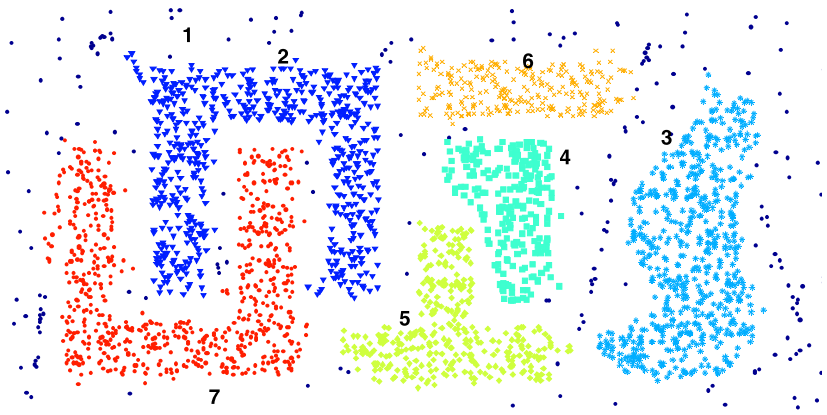


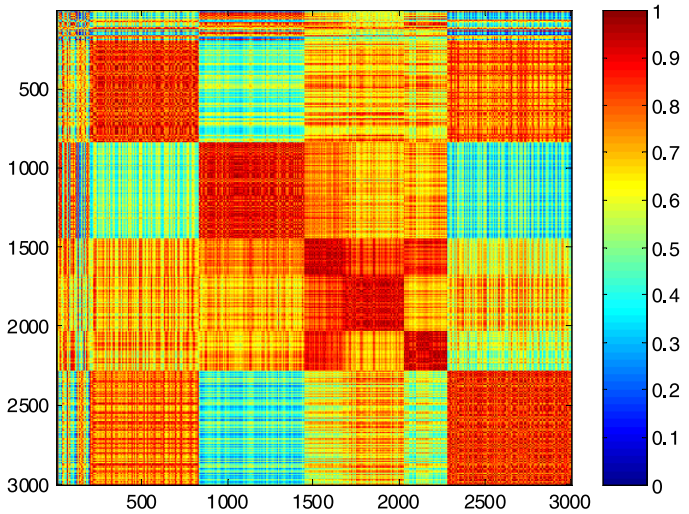


Complete Link

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Unsupervised Cluster Evaluation: Similarity Matrix





DBSCAN

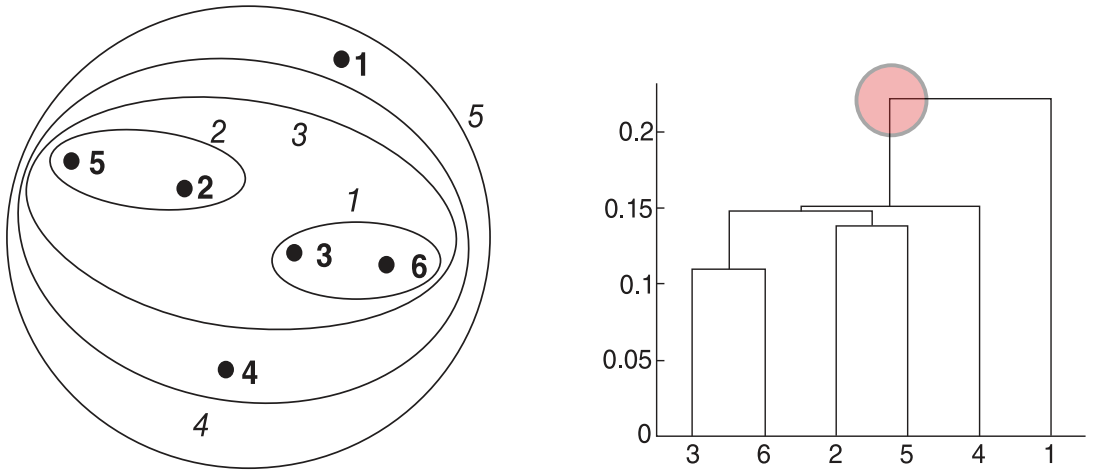
All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Unsupervised Cluster Evaluation: Cophenetic Correlation

- Cophenetic Distance between two objects is the proximity at which an agglomerative hierarchical clustering technique puts the objects in the same cluster for the first time.
- Cophenetic distance matrix, the entries are the cophenetic distances between each pair of objects.
- Cophenetic Correlation Coefficient (CPCC) is the correlation between the entries of this matrix and the original dissimilarity matrix.
- It is a standard measure of how well a hierarchical clustering fits the data.
- Common use is to evaluate which type of hierarchical clustering is best for a particular type of data.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Unsupervised Cluster Evaluation: Cophenetic Correlation



(a) Single link clustering. (b) Single link dendrogram.

Figure 7.16. Single link clustering of the six points shown in Figure 7.15.

Table 7.7. Cophenetic distance matrix for single link and data in Table 2.14 on page 90.

Point	P1	P2	P3	P4	P5	P6
P1	0	0.222	0.222	0.222	0.222	0.222
P2	0.222	0	0.148	0.151	0.139	0.148
P3	0.222	0.148	0	0.151	0.148	0.110
P4	0.222	0.151	0.151	0	0.151	0.151
P5	0.222	0.139	0.148	0.151	0	0.148
P6	0.222	0.148	0.110	0.151	0.148	0

Technique	CPCC
Single Link	0.44
Complete Link	0.63
Group Average	0.66
Ward's	0.64

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Clustering Tendency — Hopkins Statistic

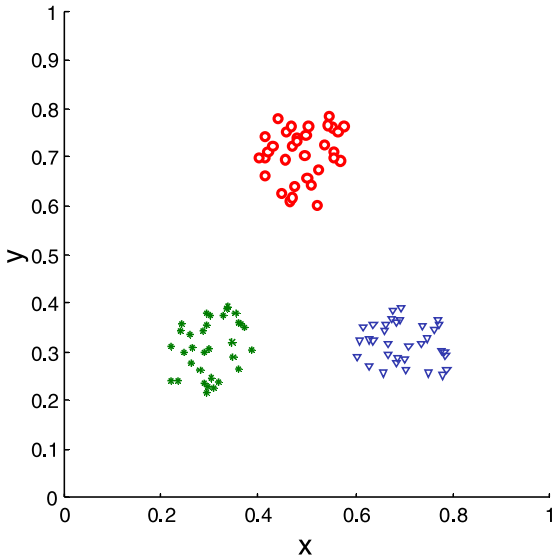
1. Generate p points that are randomly distributed across the data space.
2. Sample p actual data points.
3. For both sets of points, we find the distance to the nearest neighbor in the original data set.
4. Let the u_i be the nearest neighbor distances of the artificially generated points, while the w_i are the nearest neighbor distances of the sample of points from the original data set.
5. The Hopkins statistic H is then defined by:

$$H = \frac{\sum_{i=1}^{p} w_i}{\sum_{i=1}^{p} u_i + \sum_{i=1}^{p} w_i} \tag{5}$$

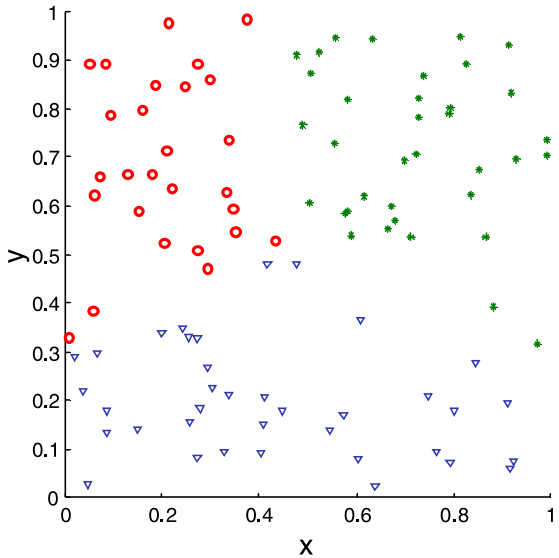
All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Clustering Tendency — Hopkins Statistic

- $H \approx 0.5$: randomly generated points and the sample of data points have roughly the same nearest neighbor distances.
- $H \approx 1.0$: data that is regularly distributed in the data space.
- $H \approx 0.0$: data that is highly clustered.



Average Value of H = 0.95



Average Value of H = 0.56

Figure: Statistic accompanying the figures are wrong in the book.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Clustering Tendency — Hopkins Statistic

$$HS_i = \frac{\sum_{y_j \in \mathbf{R}_i} (\delta_{\min}(\mathbf{y}_j))^d}{\sum_{y_j \in \mathbf{R}_i} (\delta_{\min}(\mathbf{y}_j))^d + \sum_{x_j \in \mathbf{D}_i} (\delta_{\min}(\mathbf{x}_j))^d} \tag{6}$$

- This statistic compares the nearest-neighbor distribution of randomly generated points to the same distribution for random subsets of points from **D**.
- $\delta_{\min}(x_j) < \delta_{\min}(y_j)$: HS_i tends to 1.0: data is well clustered.
- $\delta_{\min}(x_j) \approx \delta_{\min}(y_j)$: HS_i tends to 0.5: data is essentially random.
- $\delta_{\min}(x_j) > \delta_{\min}(y_j)$: HS_i tends to 0.0: indicates point repulsion, with no clustering.

All from: Zaki and Meira, "Data Mining and Machine Learning: Fundamental Concepts and Algorithms", 2nd Edition, 2020, Cambridge University Press.

Clustering Tendency — Hopkins Statistic

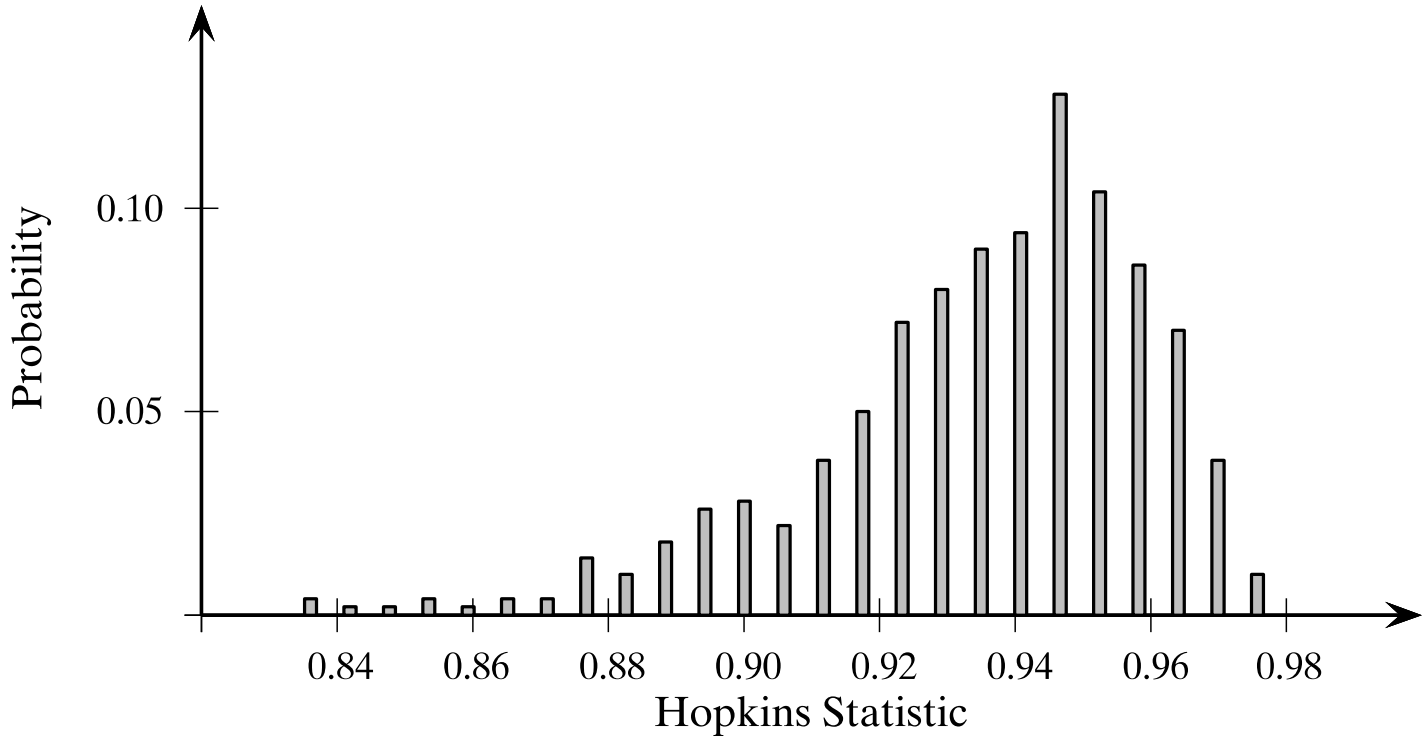


Figure 17.9. Iris dataset: Hopkins statistic distribution.

All from: Zaki and Meira, "Data Mining and Machine Learning: Fundamental Concepts and Algorithms", 2nd Edition, 2020, Cambridge University Press.

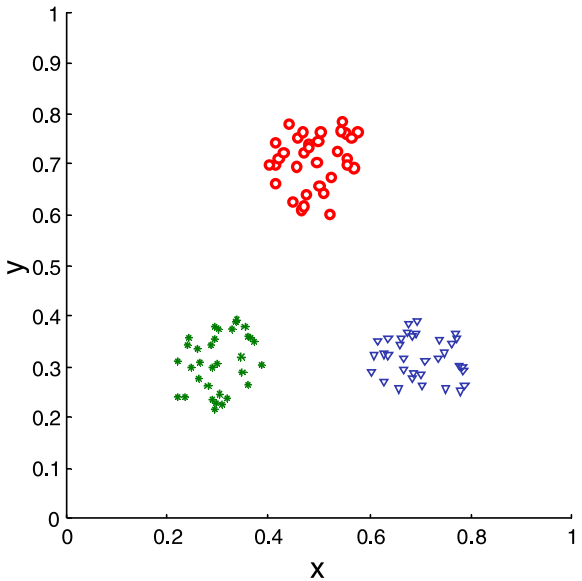
Framework for Cluster Validity

- Need a framework to interpret any measure.
  - For example, if our measure of evaluation has the value, 10, is that good, fair, or poor?
- Statistics provide a framework for cluster validity.
  - The more "atypical" a clustering result is, the more likely it represents valid structure in the data.
  - Can compare the values of an index that result from random data or clusterings to those of a clustering result.
  - If the value of the index is unlikely, then the cluster results are valid.
- For comparing the results of two different sets of cluster analyses, a framework is less necessary.
  - However, there is the question of whether the difference between two index values is significant.

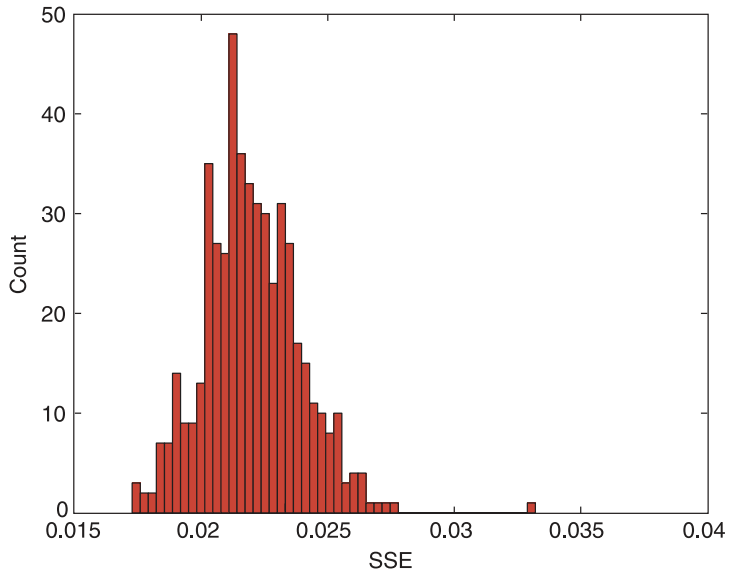
All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Statistical Framework for SSE

- Example: Compare SSE of three cohesive clusters against three clusters in random data.



SSE = 0.005



Histogram shows SSE of three clusters in 500 sets of random data points of size 100 distributed over the range 0.2 – 0.8 for x and y values.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Supervised Cluster Evaluation

- We will cover supervised evaluation metrics (e.g., Precision, Recall, Entropy) in classification.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.