DATA MINING CLUSTERING

The DBSCAN algorithm

Evaluation

DBSCAN

DBSCAN: Density-Based Clustering

- DBSCAN is a Density-Based Clustering algorithm

- Reminder: In density-based clustering we partition points into dense regions separated by not-so-dense regions.

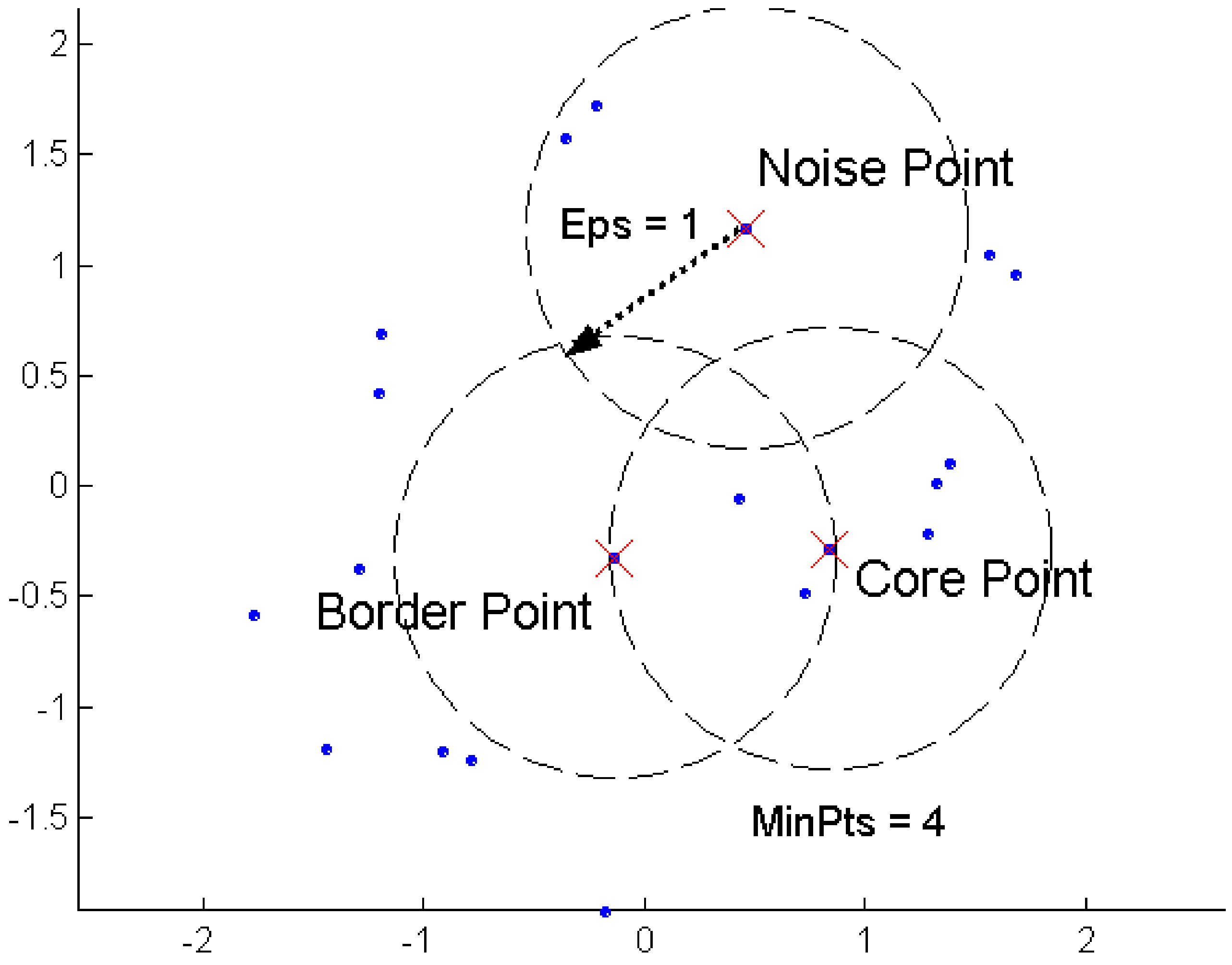
- Important Questions:
  - How do we measure density?
  - What is a dense region?

- DBSCAN:
  - Density at point p: number of points within a circle of radius Eps
  - Dense Region: A circle of radius Eps that contains at least MinPts points

DBSCAN

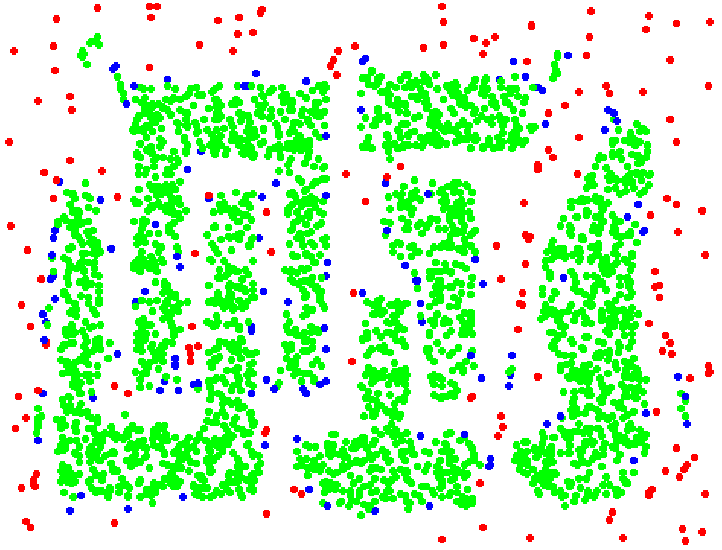
- Characterization of points
  - A point is a core point if it has more than or equal to a specified number of points (MinPts) within Eps
    - These points belong in a dense region and are at the interior of a cluster
  - A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point.
  - A noise point is any point that is not a core point or a border point.

DBSCAN: Core, Border, and Noise Points



DBSCAN: Core, Border and Noise Points

Original Points

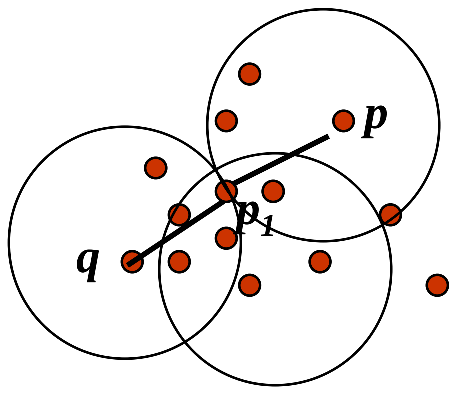


Point types: core, border and noise

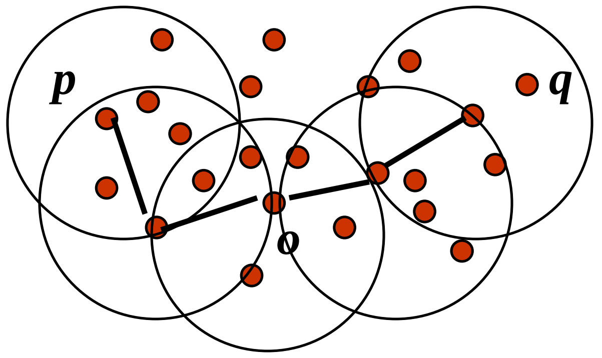
Eps = 10, MinPts = 4

Density-Connected points

- Density edge
  - We place an edge between two core points q and p if they are within distance Eps.



- Density-connected
  - A point p is density-connected to a point q if there is a path of edges from p to q



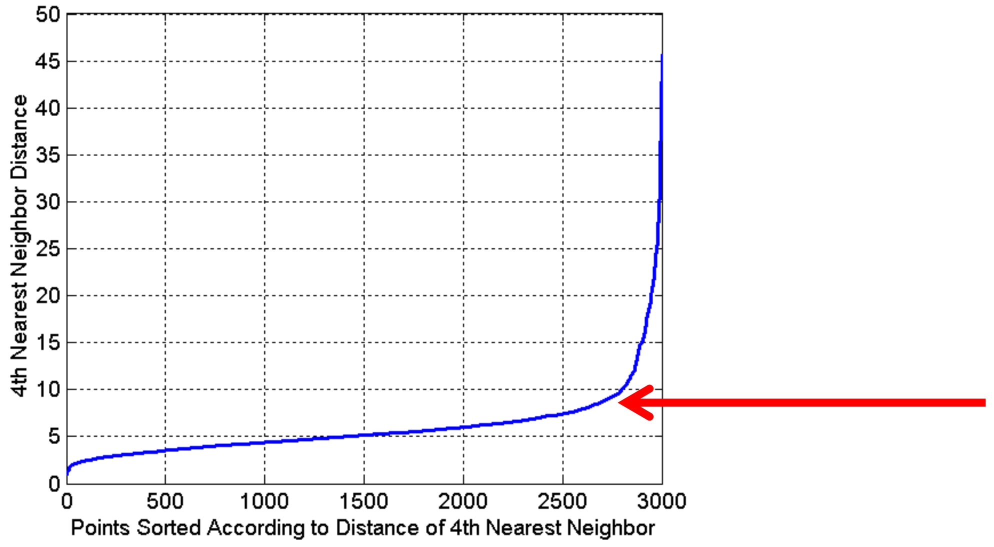
DBSCAN Algorithm

- Label points as core, border and noise
- Eliminate noise points
- For every core point p that has not been assigned to a cluster
  - Create a new cluster with the point p and all the points that are density-connected to p.
- Assign border points to the cluster of the closest core point.

DBSCAN: Determining Eps and MinPts

- Idea: for points in a cluster, their k^{th} nearest neighbors are at roughly the same 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
- Find the distance d where there is a "knee" in the curve
  - Eps = d, MinPts = k

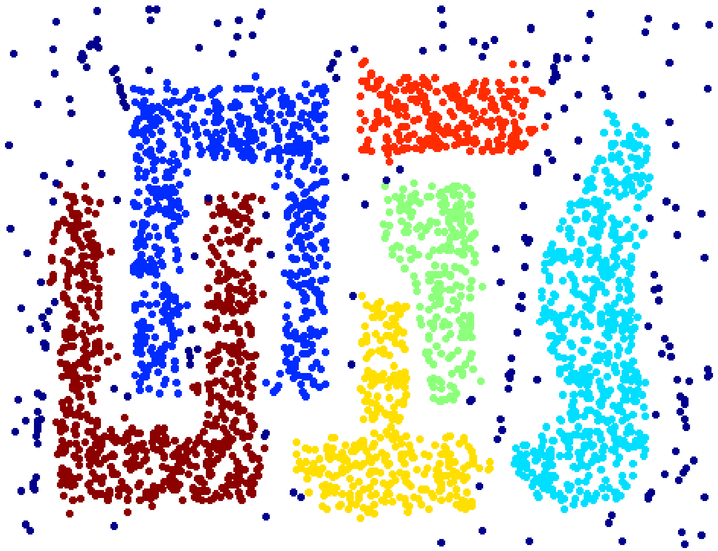
忽然上升



《导论》p567

When DBSCAN Works Well

Original Points



Clusters

- Resistant to Noise
- Can handle clusters of different shapes and sizes

DBSCAN: Sensitive to Parameters

Figure 8. DBScan results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.

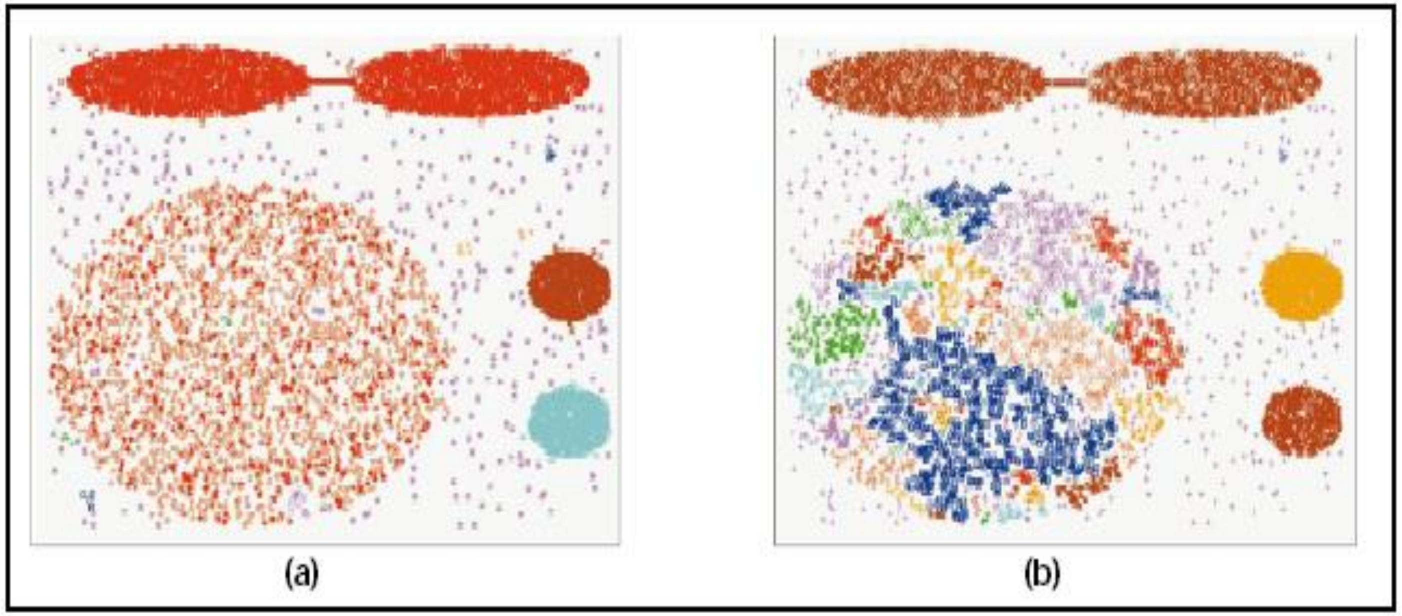
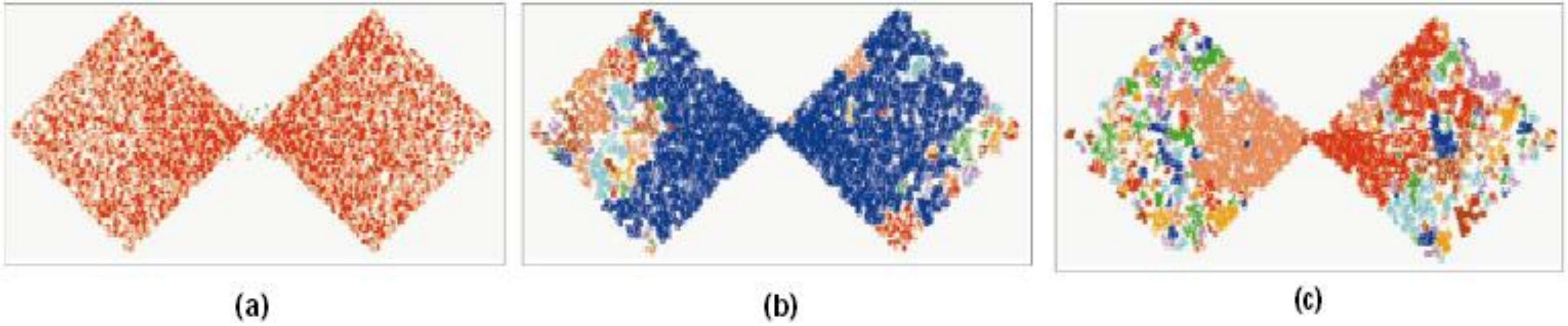
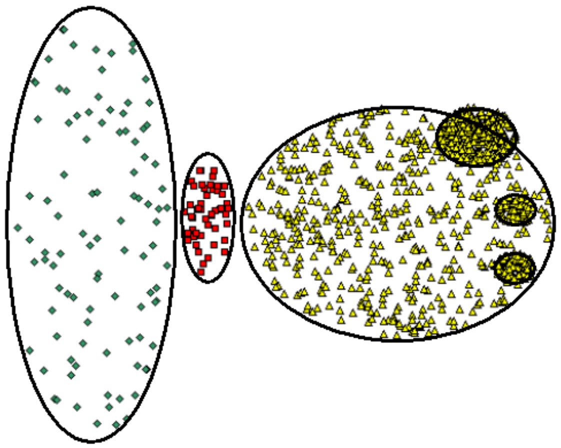


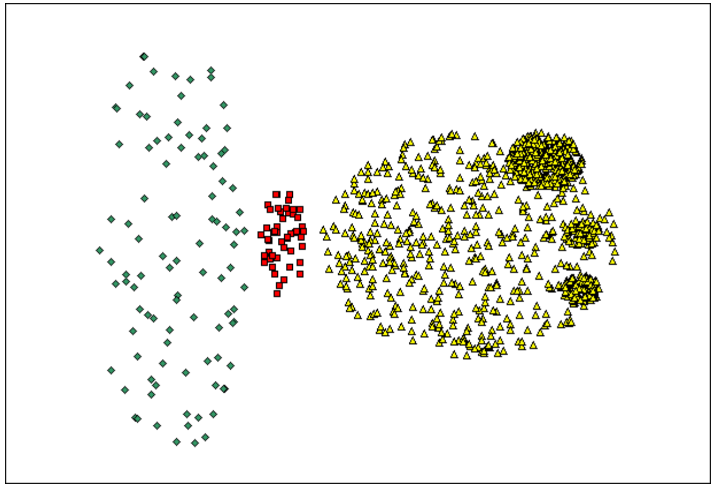
Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.



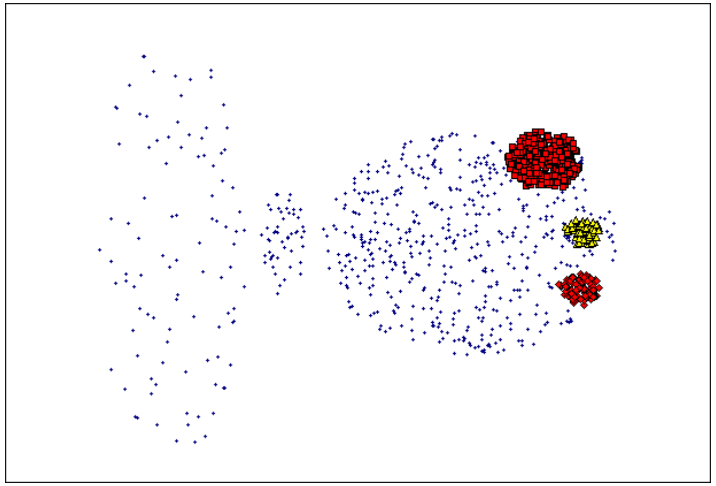
When DBSCAN Does NOT Work Well



Original Points



(MinPts=4, Eps=9.75).



(MinPts=4, Eps=9.92)

- Varying densities
- High-dimensional data

CLUSTERING EVALUATION

Clustering Evaluation

- We need to evaluate the "goodness" of the resulting clusters?
- But "clustering lies in the eye of the beholder"!
- Then why do we want to evaluate them?
  - To avoid finding patterns in noise
  - To compare clusterings, or clustering algorithms
  - To compare against a "ground truth"

Different Aspects of Cluster Validation

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

Measures of Cluster Validity

- Numerical measures to judge various aspects of cluster validity
  - Internal Index: Used to measure the goodness of a clustering structure without reference to external information.
    - E.g., Sum of Squared Error (SSE)
  - External Index: Used to measure the extent to which cluster labels match externally supplied class labels.
    - E.g., precision, recall

CLUSTER VALIDITY WITH INTERNAL CRITERIA

Internal Measures

- Internal Index: Used to measure the goodness of a clustering structure without reference to external information
    - Example: Sum of Squared Error (SSE)
- SSE is good for comparing two clusterings; or two clusters (average SSE, since they may have different sizes).

Cohesion and Separation

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

《导论》P332

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

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

We want this to be small

$|x - c_i|$为两点间距离

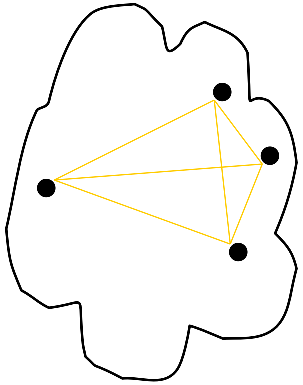
  - Separation is measured by the between cluster sum of squares

$$BSS = \sum_{x \in C_i} \sum_{y \in C_j} (|x - y|)^2$$

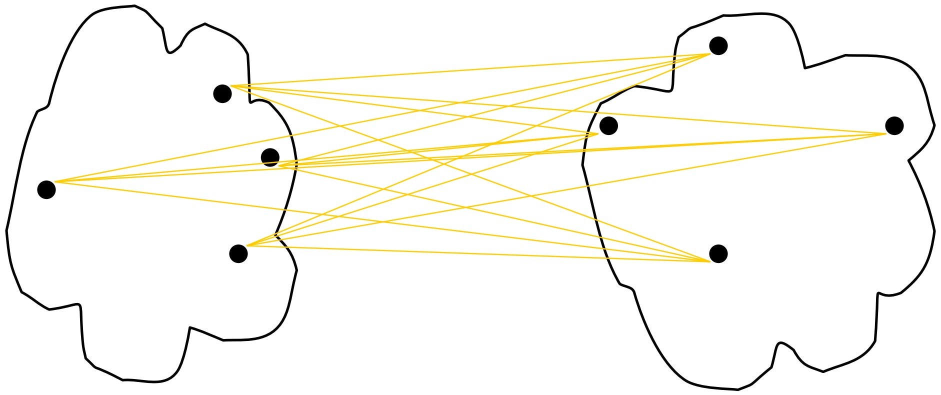
We want this to be large

Cohesion and Separation

- A proximity graph-based approach can also be used for cohesion and separation.
  - Cluster cohesion is the sum of the weight of all links within a cluster.
  - Cluster separation is the sum of the weights between nodes in the cluster and nodes outside the cluster.



cohesion

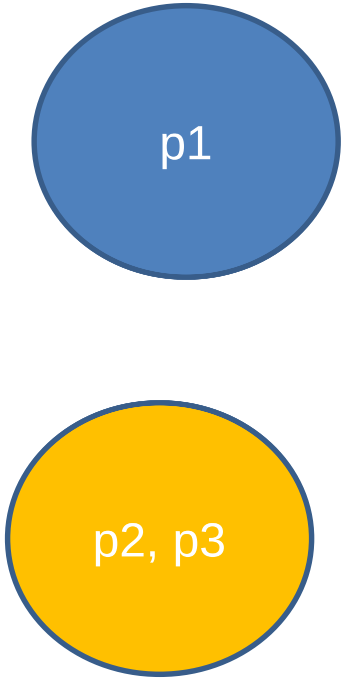


separation

Measuring Cluster Validity Via Correlation

$$CorrCoeff(X,Y) = \frac{\sum_i (x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{\sum_i (x_i - \mu_X)^2}\sqrt{\sum_i (y_i - \mu_Y)^2}}$$

- Two matrices
  - Similarity or Distance Matrix   Pair-wise 相似度或者距离
    - One row and one column for each data point
    - An entry is the similarity or distance of the associated pair of points
  - "Incidence" Matrix   聚类结果   《导论》P337
    - 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
- Compute the correlation between the two matrices
  - Since the matrices are symmetric, only the correlation between n(n-1) / 2 entries needs to be calculated.
- High correlation (positive for similarity, negative for distance) indicates that points that belong to the same cluster are close to each other.



- Incidence Matrix

	p1	p2	p3
p1		0	0
p2	0		1
p3	0	1	

- Similarity Matrix

	p1	p2	p3
p1		0.2	0.3
p2	0.2		0.8
p3	0.3	0.8	

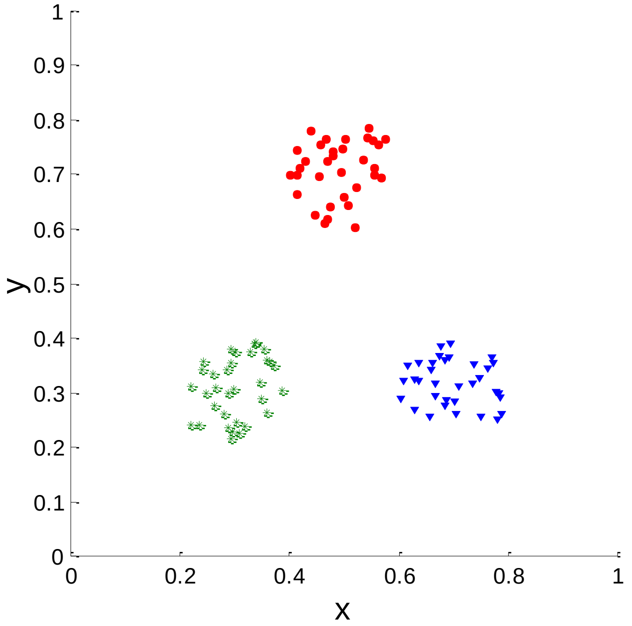
Correlation between [0,0,1] and [0.2,0.3,0.8]

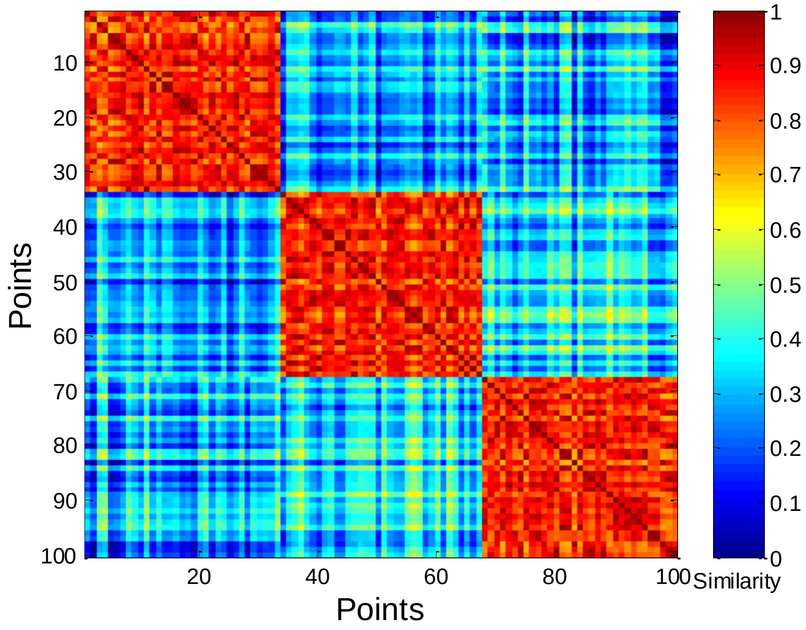
Using Similarity Matrix for Cluster Validation

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

Corr = -0.9235

距离矩阵与聚类结果矩阵的相关系数



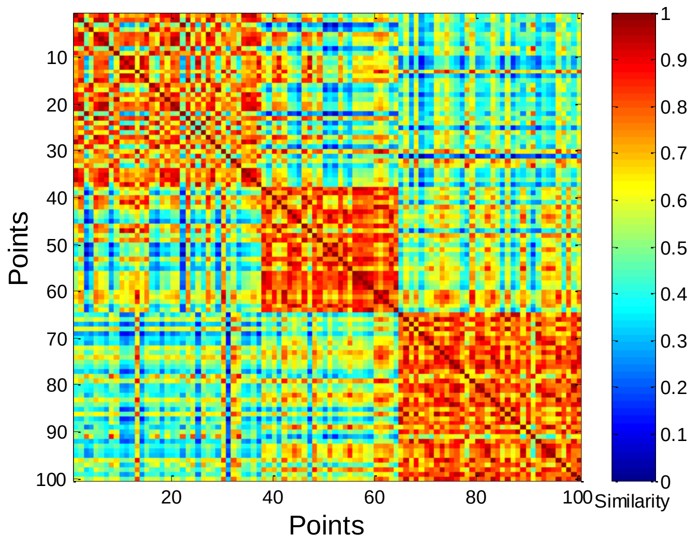


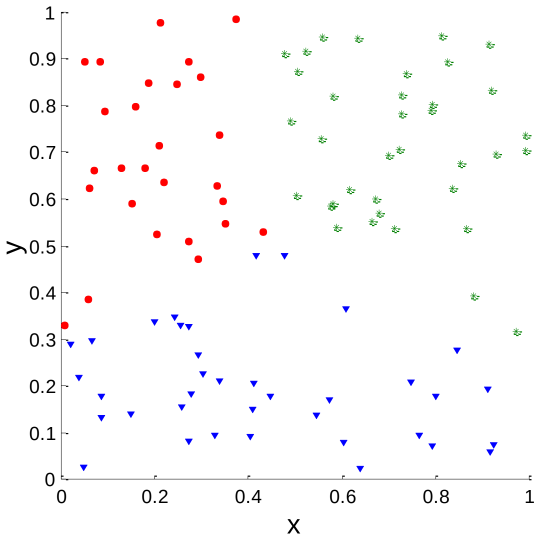
$$sim(i,j) = 1 - \frac{d_{ij} - d_{min}}{d_{max} - d_{min}}$$

距离转换为相似度[0,1]的方式

Using Similarity Matrix for Cluster Validation

- Clusters in random data are not so crisp

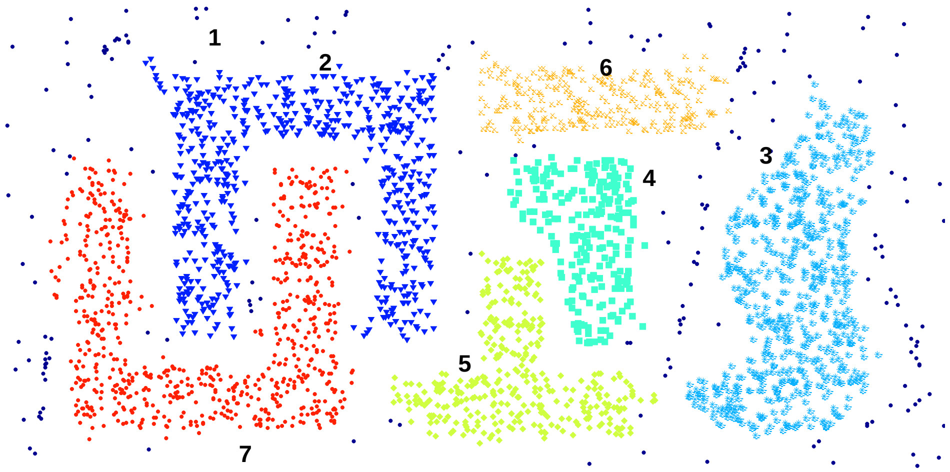


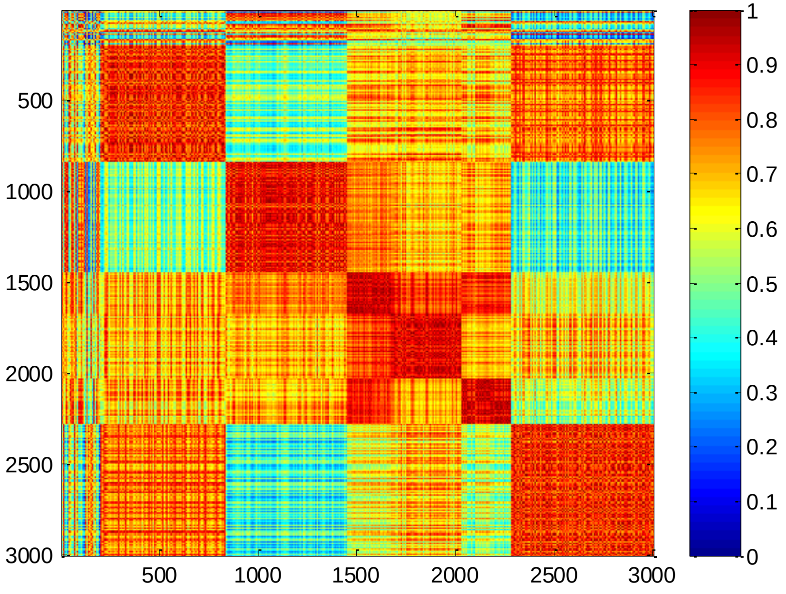


Corr = -0.5810

K-means

Using Similarity Matrix for Cluster Validation





DBSCAN

- Clusters in more complicated figures are not well separated
- This technique can only be used for small datasets since it requires a quadratic computation

STATISTICAL FRAMEWORK FOR CLUSTER(ING) VALIDITY

Framework for Cluster Validity

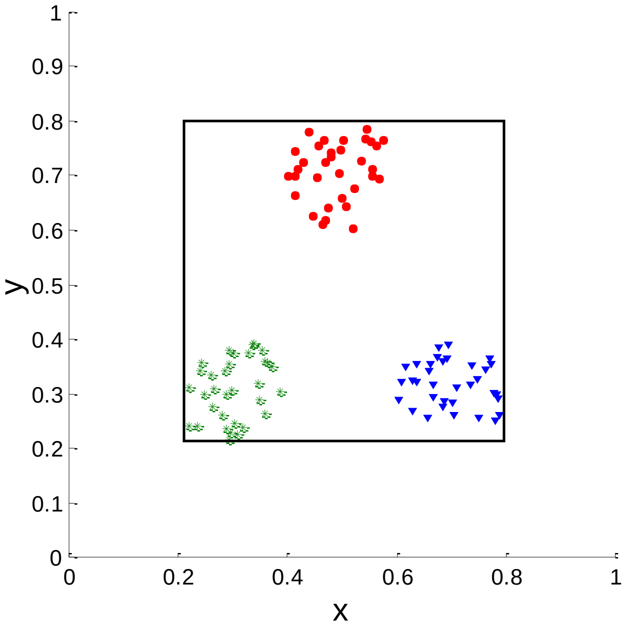
- Need a framework to interpret any measure.
  - For example, if our measure of evaluation has the value, 10, is that good, fair, or poor?
- Statistics provide a framework for cluster validity
  - The more “non-random” a clustering result is, the more likely it represents valid structure in the data
  - 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

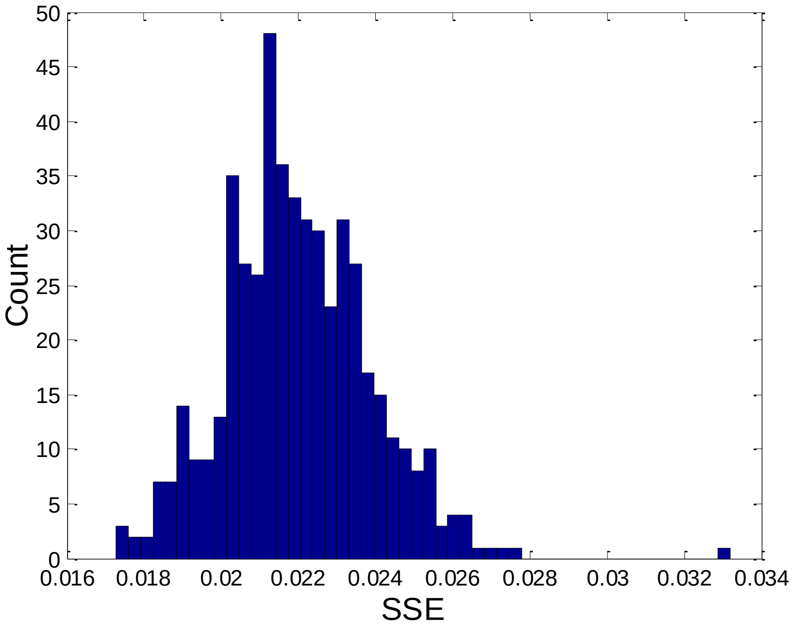
Statistical Framework for SSE

- Example
  - Compare SSE of 0.005 against three clusters in random data
  - Histogram of SSE for three clusters in 500 random data sets of 100 random points distributed in the range 0.2 – 0.8 for x and y
    - Value 0.005 is very unlikely

《导论》P344

SSE = 0.005





Empirical p-value

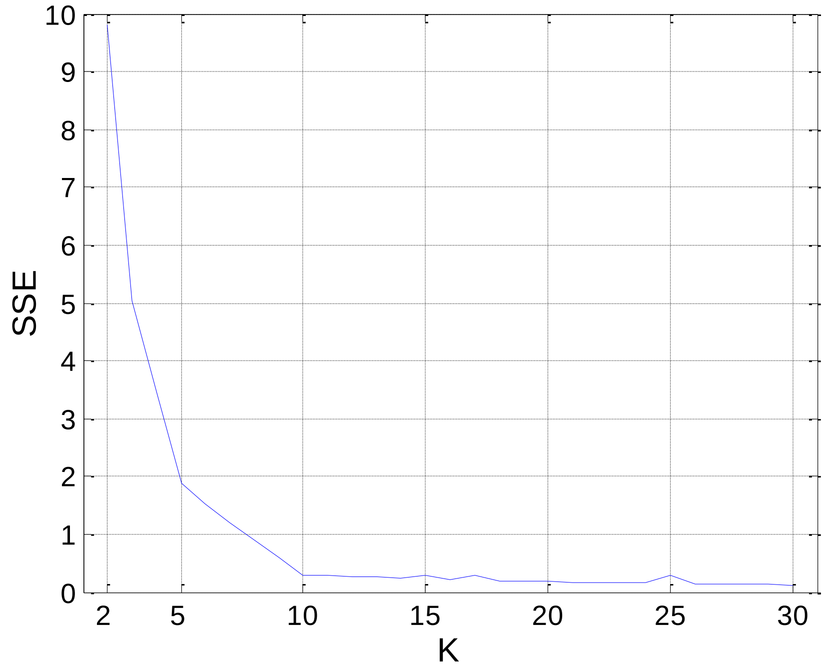
- If we have a measurement v (e.g., the SSE value)
- ..and we have N measurements on random datasets
- …the empirical p-value is the fraction of measurements in the random data that have value less or equal than value v (or greater or equal if we want to maximize)
  - i.e., the value in the random dataset is at least as good as that in the real data

- We usually require that p-value ≤ 0.05

ESTIMATING THE "RIGHT" NUMBER OF CLUSTERS

Estimating the “right” number of clusters

- Typical approach: find a “knee” in an internal measure curve.

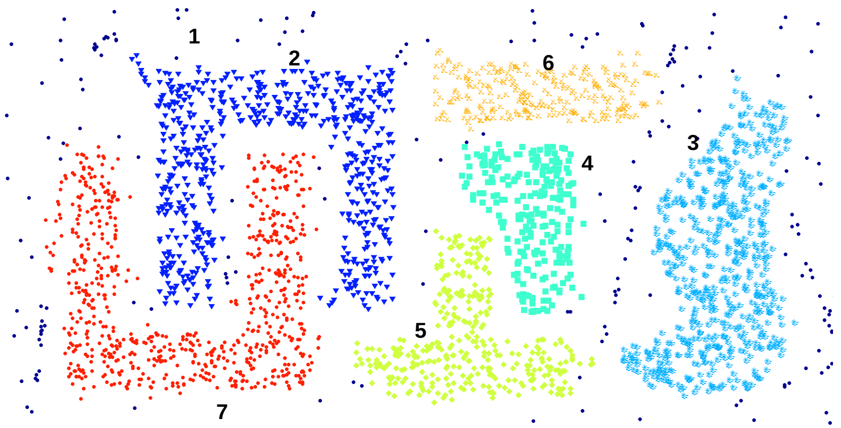


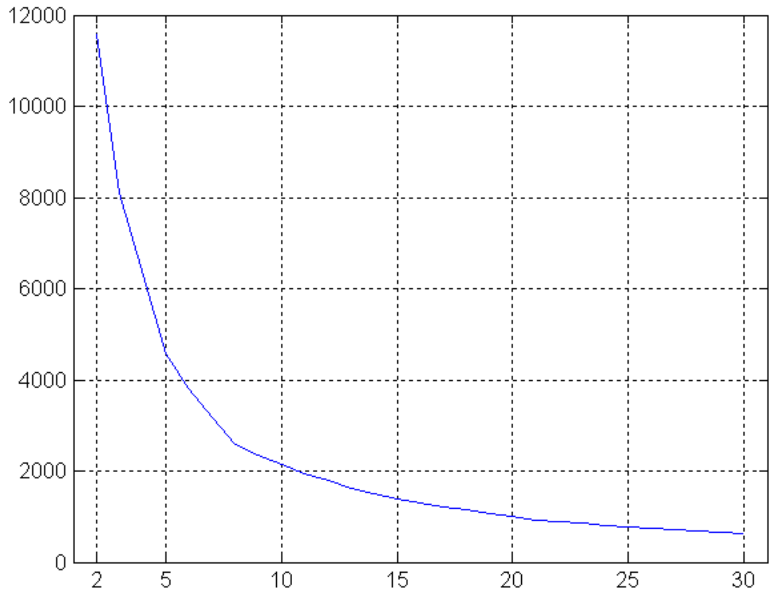
For more, see 《导论》 P339

- Desirable property: the clustering algorithm does not require the number of clusters to be specified (e.g., DBSCAN)

Estimating the “right” number of clusters

- SSE curve for a more complicated data set





SSE of clusters found using K-means

EVALUATION WITH EXTERNAL "GROUND TRUTH"

External Measures for Clustering Validity

- Assume that the data is labeled with some class labels
  - E.g., documents are classified into topics, people classified according to their income.
  - This is called the "ground truth"
- In this case we want the clusters to be homogeneous with respect to classes
  - Each cluster should contain elements of mostly one class
  - Each class should ideally be assigned to a single cluster

Confusion matrix

- Rows: clusters
- Columns: classes
- Entries: counts/probability of cluster-class pair
- n = number of points
- m_i = points in cluster i
- c_j = points in class j
- n_{ij}= points in cluster i coming from class j
- $p_{ij} = n_{ij}/m_i$= probability of element from cluster i to be assigned in class j

Confusion matrix of clusters/classes (counts)

	Class 1	Class 2	Class 3	
Cluster 1	n_{11}	n_{12}	n_{13}	m_1
Cluster 2	n_{21}	n_{22}	n_{23}	m_2
Cluster 3	n_{31}	n_{32}	n_{33}	m_3
	c_1	c_2	c_3	n

Joint distribution of clusters/classes

	Class 1	Class 2	Class 3	
Cluster 1	p_{11}	p_{12}	p_{13}	m_1
Cluster 2	p_{21}	p_{22}	p_{23}	m_2
Cluster 3	p_{31}	p_{32}	p_{33}	m_3
	c_1	c_2	c_3	n

Measures

- Precision:
  - Of cluster i with respect to class j: $Prec(i,j) = \frac{n_{ij}}{m_i} = p_{ij}$
- Recall:
  - Of cluster i with respect to class j: $Rec(i,j) = \frac{n_{ij}}{c_j}$
- F-measure:
  - Harmonic Mean of Precision and Recall:

$$F(i,j) = \frac{2 * Prec(i,j) * Rec(i,j)}{Prec(i,j) + Rec(i,j)}$$

	Class 1	Class 2	Class 3	
Cluster 1	n_{11}	n_{12}	n_{13}	m_1
Cluster 2	n_{21}	n_{22}	n_{23}	m_2
Cluster 3	n_{31}	n_{32}	n_{33}	m_3
	c_1	c_2	c_3	n

	Class 1	Class 2	Class 3	
Cluster 1	p_{11}	p_{12}	p_{13}	m_1
Cluster 2	p_{21}	p_{22}	p_{23}	m_2
Cluster 3	p_{31}	p_{32}	p_{33}	m_3
	c_1	c_2	c_3	n

Measures

Precision/Recall for clusters and clusterings

- Assign to cluster i the class k_i such that $k_i = \arg\max_j n_{ij}$
- Precision:
  - Of cluster i: $Prec(i) = \frac{n_{ik_i}}{m_i}$
  - Of the clustering: $Prec(C) = \sum_i \frac{m_i}{n} Prec(i)$
- Recall:
  - Of cluster i: $Rec(i) = \frac{n_{ik_i}}{c_{k_i}}$
  - Of the clustering: $Rec(C) = \sum_i \frac{m_i}{n} Rec(i)$
- F-measure:
  - Harmonic Mean of Precision and Recall

加权平均，权重是cluster中元素占总体的比例

	Class 1	Class 2	Class 3	
Cluster 1	n_{11}	n_{12}	n_{13}	m_1
Cluster 2	n_{21}	n_{22}	n_{23}	m_2
Cluster 3	n_{31}	n_{32}	n_{33}	m_3
	c_1	c_2	c_3	n

Good and bad clustering

	Class 1	Class 2	Class 3	
Cluster 1	2	3	85	90
Cluster 2	90	12	8	110
Cluster 3	8	85	7	100
	100	100	100	300

Precision: (0.94, 0.81, 0.85)
– overall 0.86
Recall: (0.85, 0.9, 0.85)
- overall 0.87

	Class 1	Class 2	Class 3	
Cluster 1	20	35	35	90
Cluster 2	30	42	38	110
Cluster 3	38	35	27	100
	100	100	100	300

Precision: (0.38, 0.38, 0.38)
– overall 0.38
Recall: (0.35, 0.42, 0.38)
– overall 0.39

Another clustering

	Class 1	Class 2	Class 3	
Cluster 1	0	0	35	35
Cluster 2	50	77	38	165
Cluster 3	38	35	27	100
	100	100	100	300

Cluster 1:

Precision: 1

Recall: 0.35