CLUSTERING VALIDITY

CLUSTER VALIDITY

How do we evaluate the "goodness" of the resulting clusters?

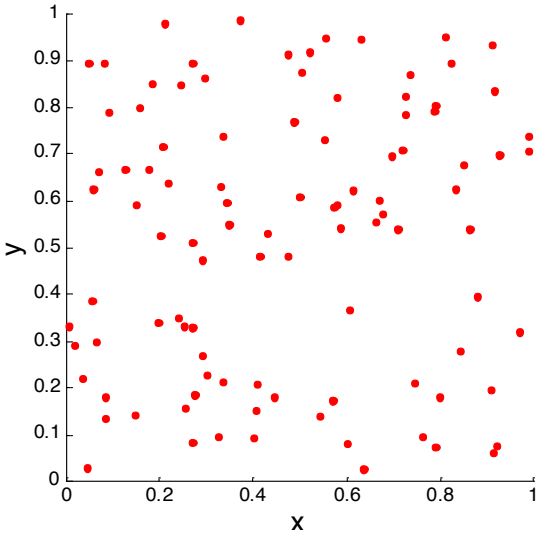
But "clustering lies in the eye of the beholder"!

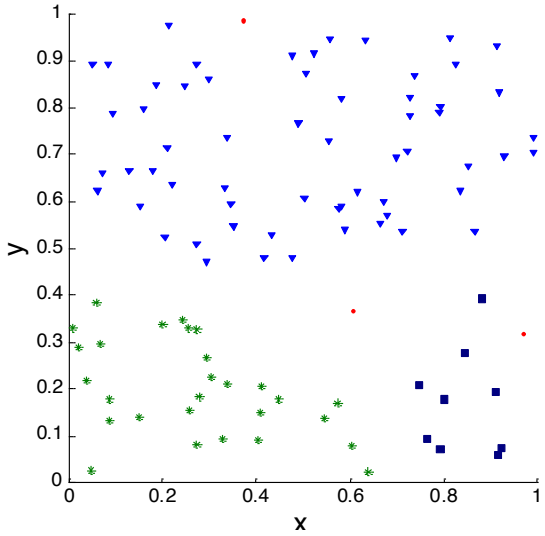
Then why do we want to evaluate them?

- To avoid finding patterns in noise
- To compare clustering algorithms
- To compare two clusterings
- To compare two clusters

CLUSTERS FOUND IN RANDOM DATA

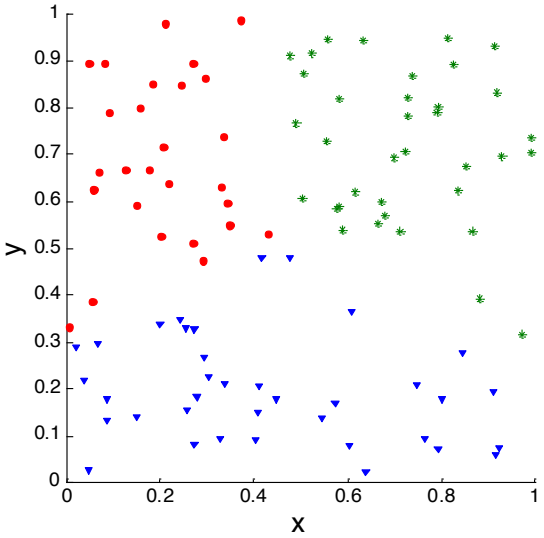
Random Points

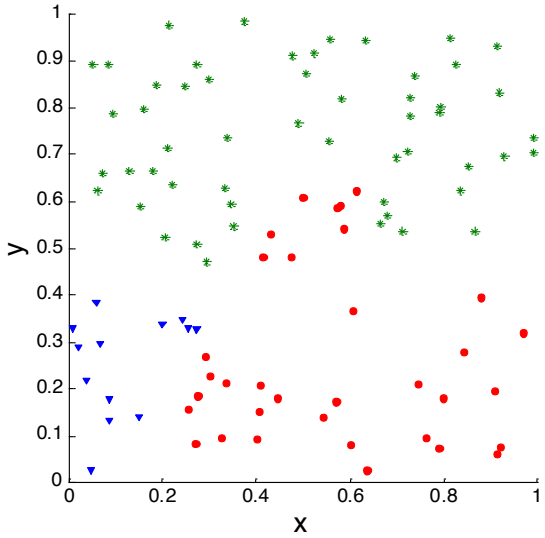




DBSCAN

K-means





Complete Link

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.

MEASURES OF CLUSTER VALIDITY

Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following three types.

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

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.

MEASURING CLUSTER VALIDITY VIA CORRELATION

Two matrices

- Similarity or Distance Matrix
  - One row and one column for each data point
  - An entry is the similarity or distance of the associated pair of points
- "Incidence" 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

Compute the correlation between the two matrices

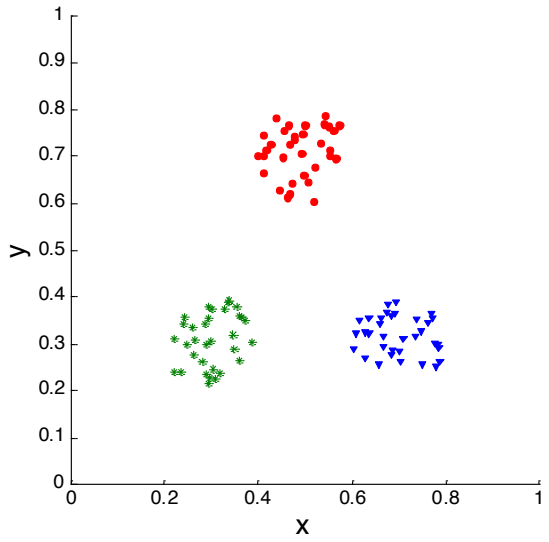
- Since the matrices are symmetric, only the correlation between n(n-1) / 2 entries needs to be calculated.

High correlation (positive for similarity, negative for distance) indicates that points that belong to the same cluster are close to each other.

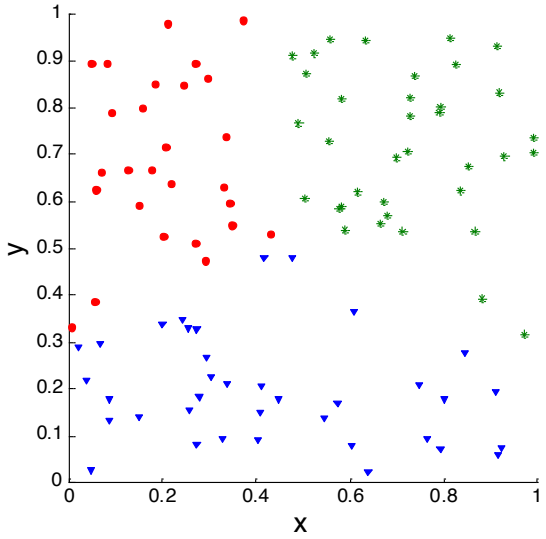
Not a good measure for some density or contiguity based clusters.

MEASURING CLUSTER VALIDITY VIA CORRELATION

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



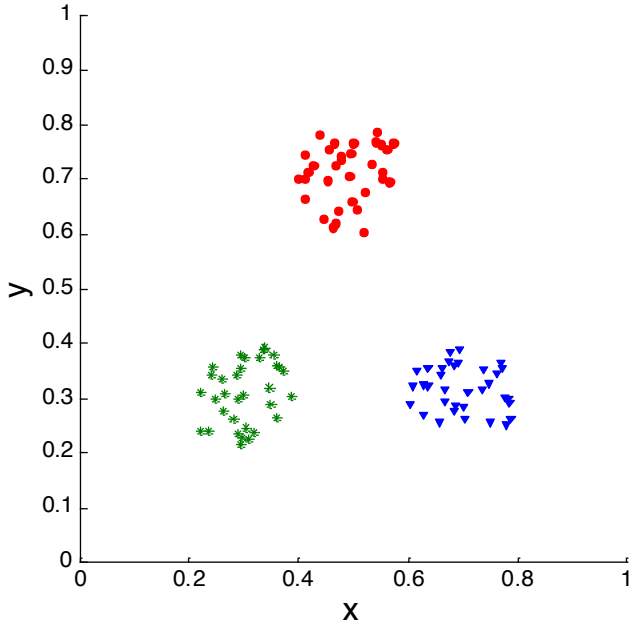
Corr = -0.9235

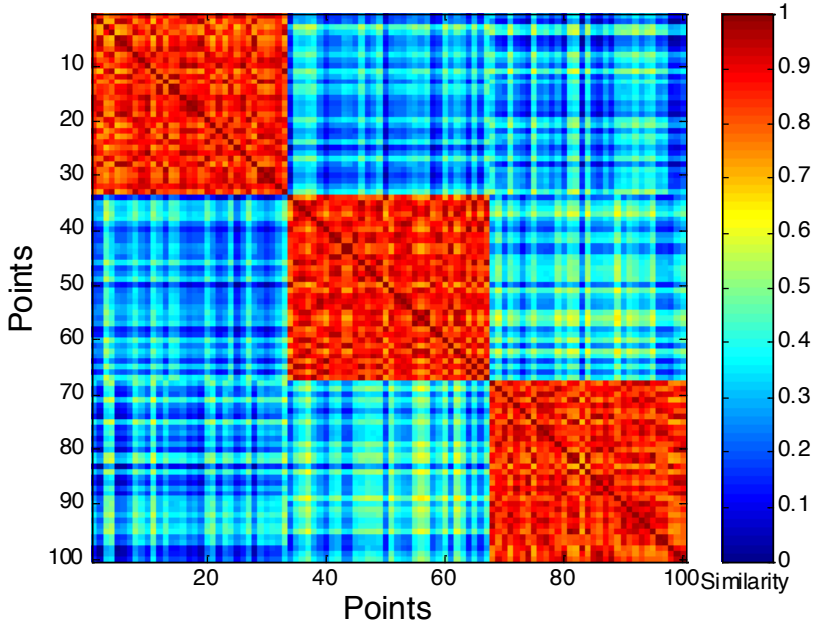


Corr = -0.5810

USING SIMILARITY MATRIX FOR CLUSTER VALIDATION

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

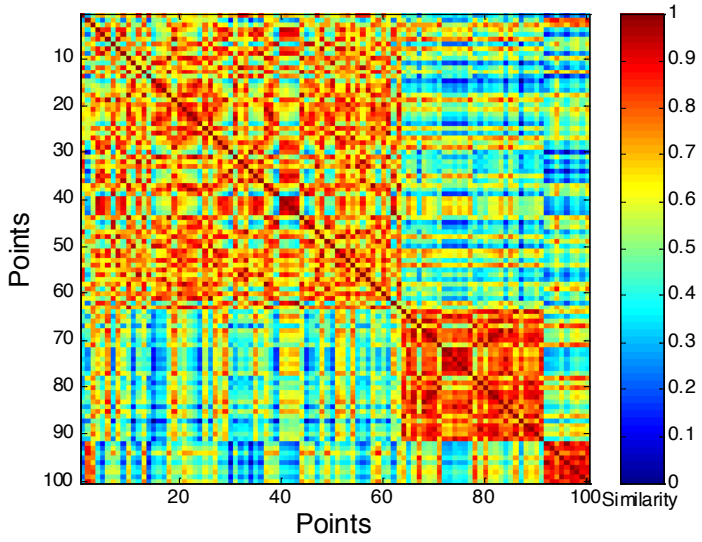


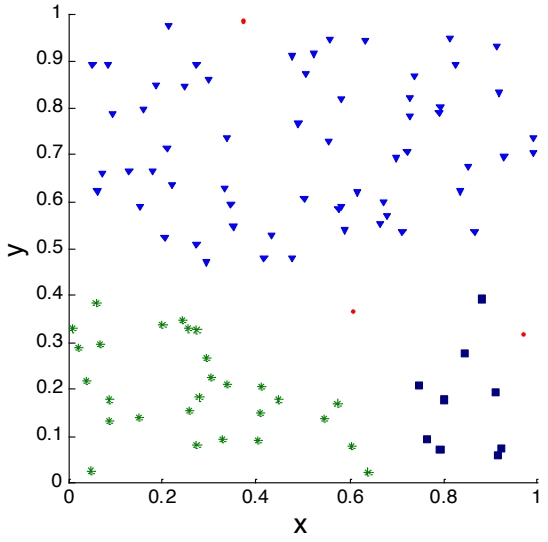


$$sim(i,j) = 1 - \frac{d_{ij} - d_{min}}{d_{max} - d_{min}}$$

USING SIMILARITY MATRIX FOR CLUSTER VALIDATION

Clusters in random data are not so crisp

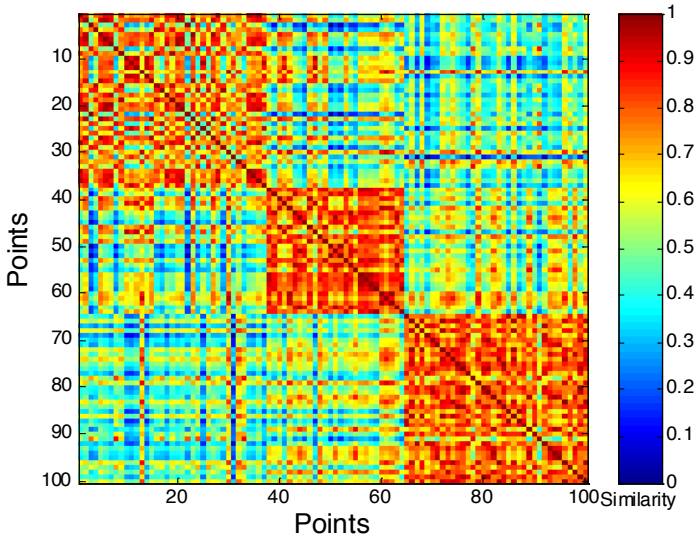


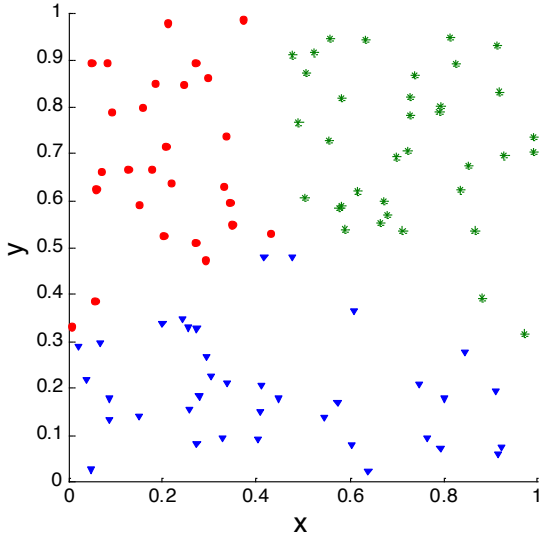


DBSCAN

USING SIMILARITY MATRIX FOR CLUSTER VALIDATION

Clusters in random data are not so crisp

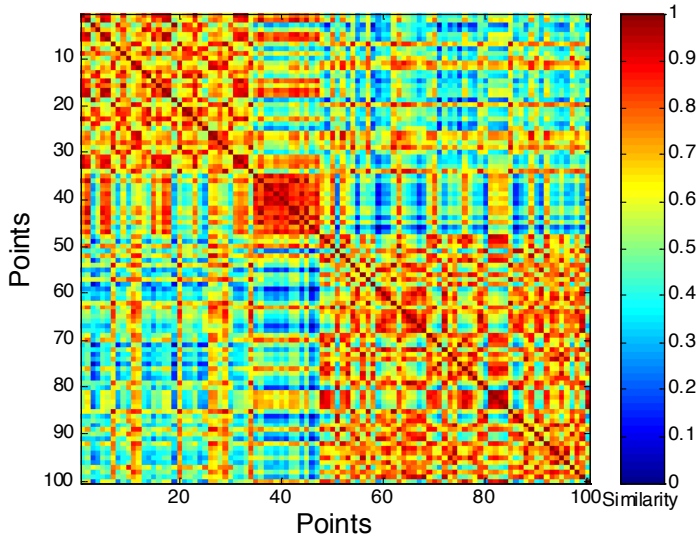


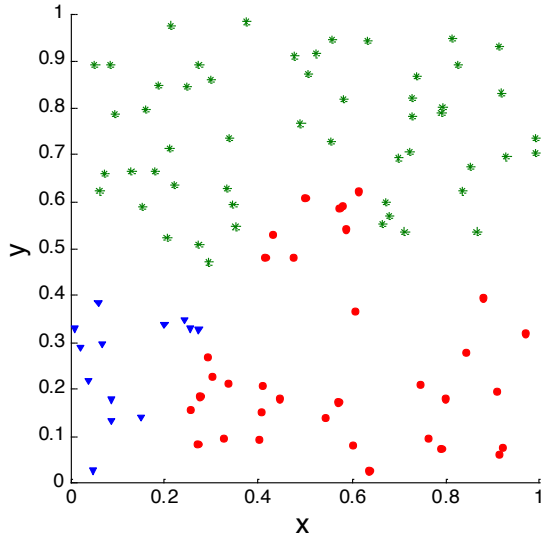


K-means

USING SIMILARITY MATRIX FOR CLUSTER VALIDATION

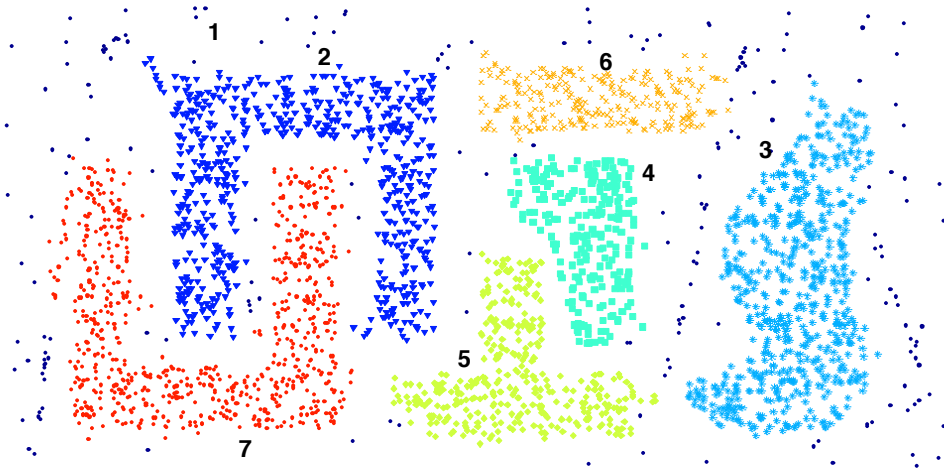
Clusters in random data are not so crisp

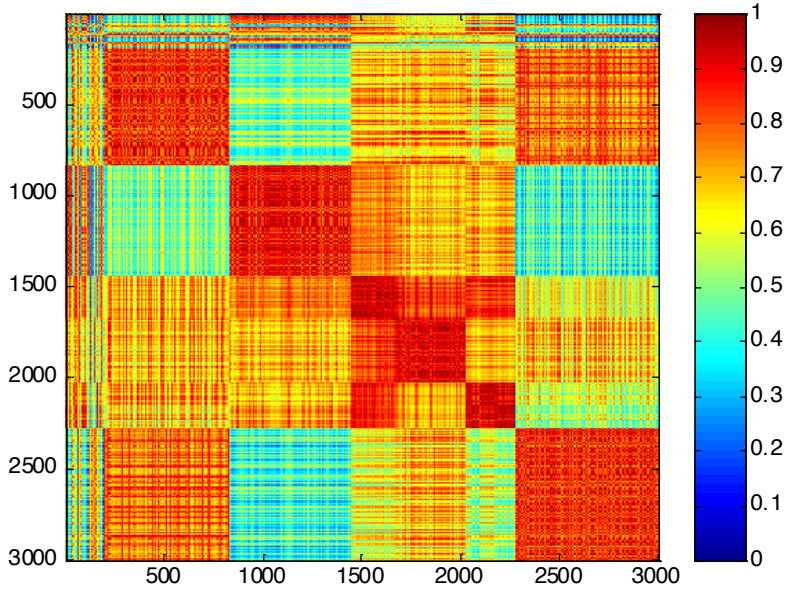




Complete Link

USING SIMILARITY MATRIX FOR CLUSTER VALIDATION





DBSCAN

- Clusters in more complicated figures aren't well separated

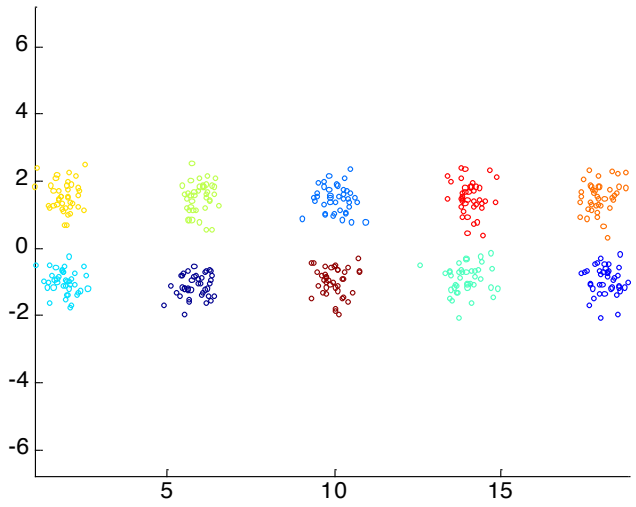
INTERNAL MEASURES: SSE

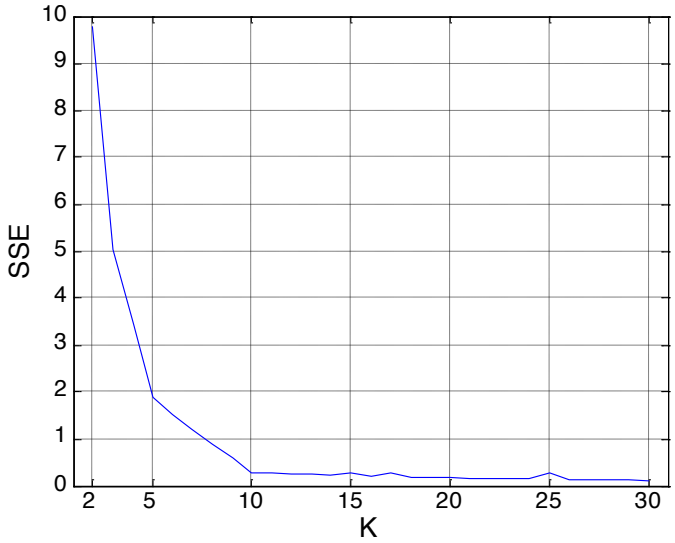
Internal Index: Used to measure the goodness of a clustering structure without reference to external information

- Example: SSE

SSE is good for comparing two clusterings or two clusters (average SSE).

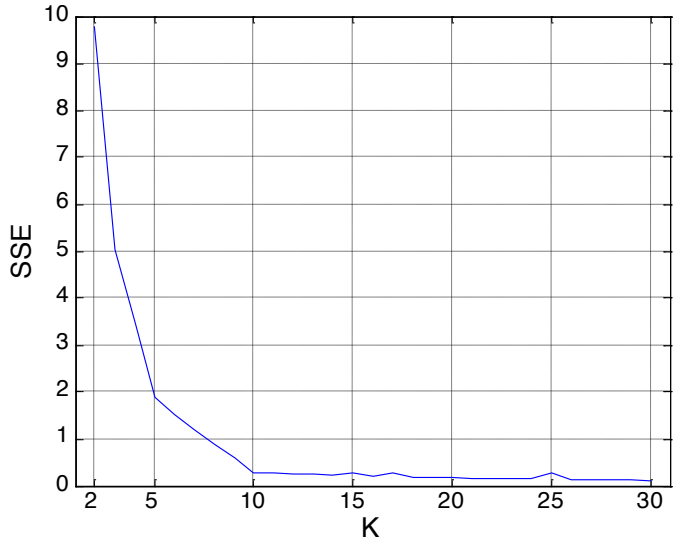
Can also be used to estimate the number of clusters





ESTIMATING THE "RIGHT" NUMBER OF CLUSTERS

Typical approach: find a "knee" in an internal measure curve.



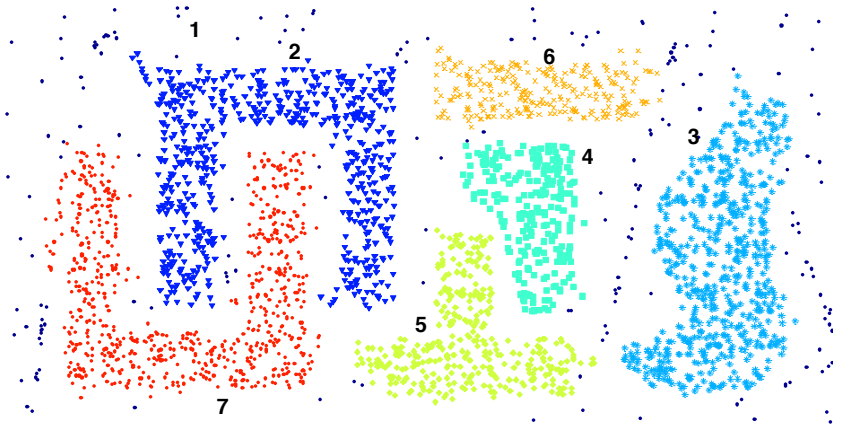
Question: why not the k that minimizes the SSE?

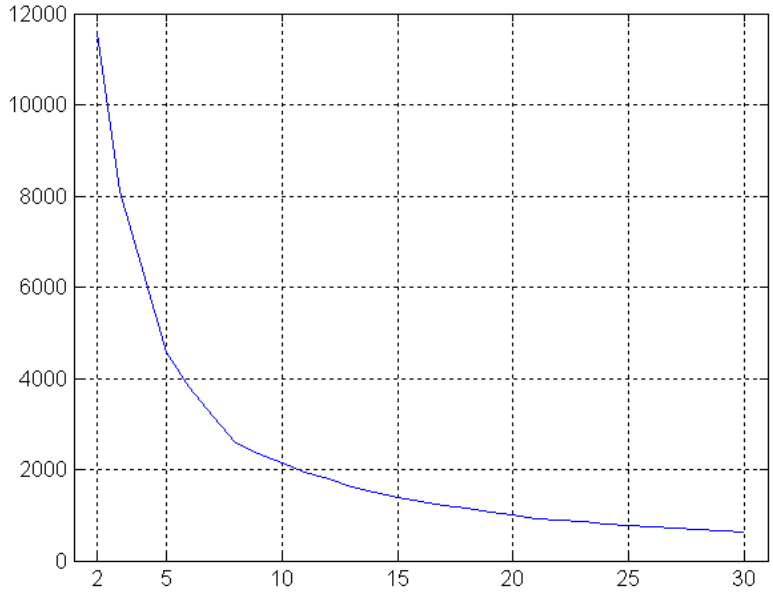
- Forward reference: minimize a measure, but with a "simple" clustering

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

INTERNAL MEASURES: SSE

SSE curve for a more complicated data set





SSE of clusters found using K-means

INTERNAL MEASURES: 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

Example: Squared Error

- Cohesion is measured by the within cluster sum of squares (SSE)

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

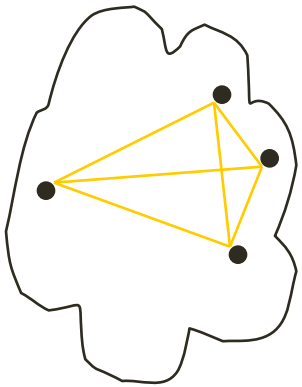
- Separation is measured by the between cluster sum of squares
  - Where m_i is the size of cluster i

$$BSS = \sum_{i} m_i (c - c_i)^2$$

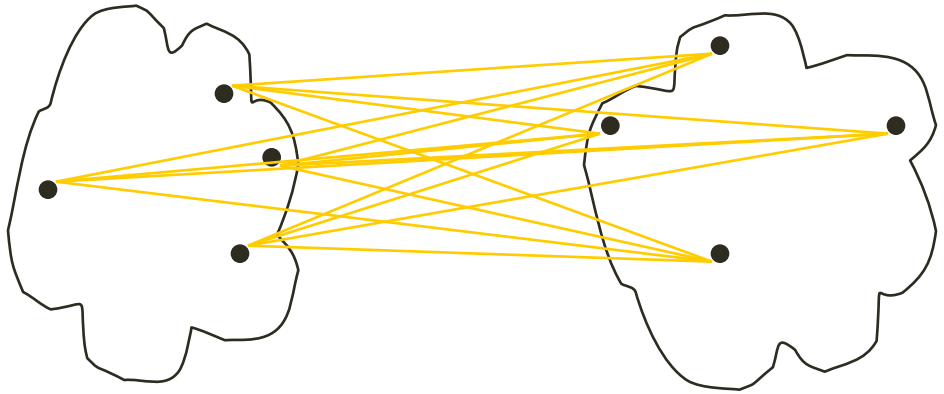
INTERNAL MEASURES: COHESION AND SEPARATION

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

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

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

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, senators classified as republican or democrat.

In this case we want the clusters to be homogeneous with respect to classes

- Each cluster should contain elements of mostly one class
- Also each class should ideally be assigned to a single cluster

This does not always make sense

- Clustering is not the same as classification

But this is what people use most of the time

MEASURES

n = number of points

m_i = points in cluster i

c_j = points in class j

m_{ij}= points in cluster i coming from class j

$p_{ij} = m_{ij}/m_i$= prob of element from class j in cluster i

Entropy:

- Of a cluster i: $e_i = -\sum_{j=1}^{L} p_{ij} \log p_{ij}$
  - Highest when uniform, zero when single class
- Of a clustering: $e = \sum_{i=1}^{K} \frac{m_i}{n} e_i$

Purity:

- Of a cluster i: $p_i = \max_j p_{ij}$
- Of a clustering: $purity = \sum_{i=1}^{K} \frac{m_i}{n} p_i$

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

MEASURES

Precision:

- Of cluster i with respect to class j: $Prec(i,j) = p_{ij}$
  - For the precision of a clustering you can take the maximum

Recall:

- Of cluster i with respect to class j: $Rec(i,j) = \frac{m_{ij}}{c_j}$
  - For the precision of a clustering you can take the maximum

F-measure:

- Harmonic Mean of Precision and Recall:

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

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

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

	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

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

ANOTHER BAD 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:

Purity: 1

Precision: 1

Recall: 0.35

EXTERNAL MEASURES OF CLUSTER VALIDITY: ENTROPY AND PURITY

Table 5.9. K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

entropy For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij}, the 'probability' that a member of cluster j belongs to class i as follows: $p_{ij} = m_{ij}/m_j$, where m_j is the number of values in cluster j and m_{ij} is the number of values of class i in cluster j. Then using this class distribution, the entropy of each cluster j is calculated using the standard formula $e_j = \sum_{i=1}^{L} p_{ij} \log_2 p_{ij}$, where the L is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster, i.e., $e = \sum_{i=1}^{K} \frac{m_i}{m} e_j$, where m_j is the size of cluster j, K is the number of clusters, and m is the total number of data points.

purity Using the terminology derived for entropy, the purity of cluster j, is given by $purity_j = \max p_{ij}$ and the overall purity of a clustering by $purity = \sum_{i=1}^{K} \frac{m_i}{m} purity_j$.

FINAL COMMENT ON CLUSTER VALIDITY

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis.

Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage."

Algorithms for Clustering Data, Jain and Dubes

MINIMUM DESCRIPTION LENGTH

OCCAM'S RAZOR

Most data mining tasks can be described as creating a model for the data

- E.g., the EM algorithm models the data as a mixture of Gaussians, the K-means models the data as a set of centroids.

What is the right model?

Occam's razor: All other things being equal, the simplest model is the best.

- A good principle for life as well

OCCAM'S RAZOR AND MDL

What is a simple model?

Minimum Description Length Principle: Every model provides a (lossless) encoding of our data. The model that gives the shortest encoding (best compression) of the data is the best.

- Related: Kolmogorov complexity. Find the shortest program that produces the data (uncomputable).
- MDL restricts the family of models considered

- Encoding cost: cost of party A to transmit to party B the data.

MINIMUM DESCRIPTION LENGTH (MDL)

The description length consists of two terms

- The cost of describing the model (model cost)
- The cost of describing the data given the model (data cost).
- $L(D) = L(M) + L(D|M)$

There is a tradeoff between the two costs

- Very complex models describe the data in a lot of detail but are expensive to describe the model
- Very simple models are cheap to describe but it is expensive to describe the data given the model

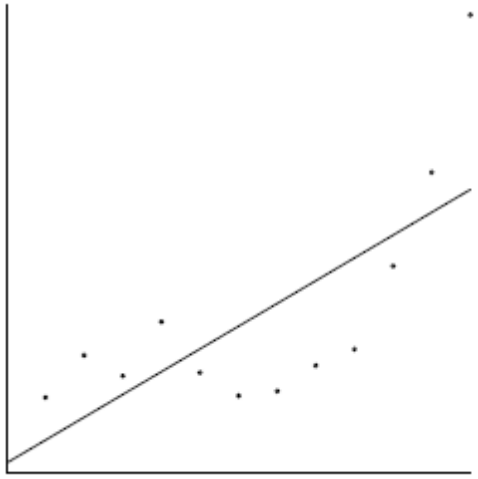
This is generic idea for finding the right model

- We use MDL as a blanket name.

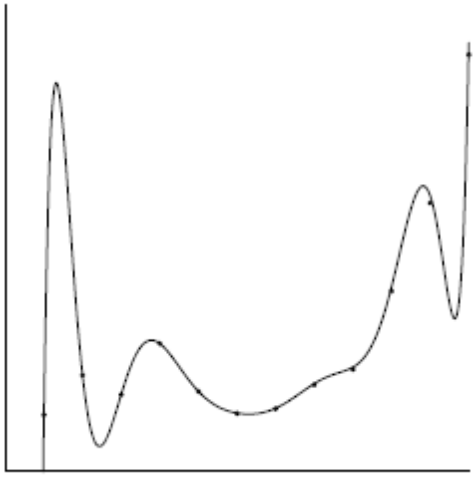
EXAMPLE

Regression: find a polynomial for describing a set of values

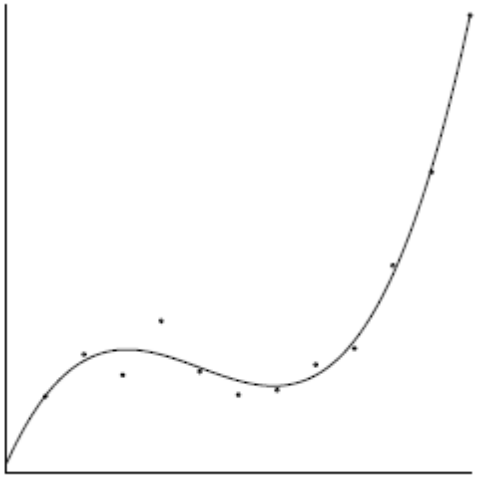
- Model complexity (model cost): polynomial coefficients
- Goodness of fit (data cost): difference between real value and the polynomial value



Minimum model cost
High data cost



High model cost
Minimum data cost



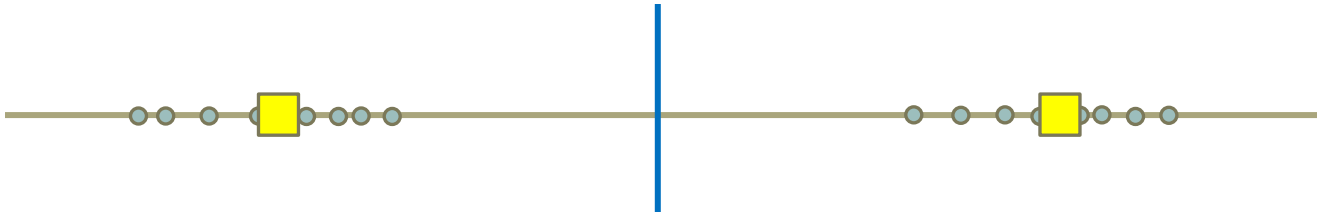
Low model cost
Low data cost

MDL avoids overfitting automatically!

EXAMPLE

Suppose you want to describe a set of integer numbers

- Cost of describing a single number is proportional to the value of the number x (e.g., logx).
- How can we get an efficient description?



Cluster integers into two clusters and describe the cluster by the centroid and the points by their distance from the centroid

- Model cost: cost of the centroids
- Data cost: cost of cluster membership and distance from centroid

What are the two extreme cases?

MDL AND DATA MINING

Why does the shorter encoding make sense?

- Shorter encoding implies regularities in the data
- Regularities in the data imply patterns
- Patterns are interesting

Example

000010000100001000010000100001000010000100001000010000100001

- Short description length, just repeat 12 times 00001

010011100101001101101010000111010111101101101010111001001110 0

- Random sequence, no patterns, no compression

ISSUES WITH MDL

What is the right model family?

- This determines the kind of solutions that we can have
  - E.g., polynomials
  - Clusterings

What is the encoding cost?

- Determines the function that we optimize
- Information theory