Clustering
Lecture 1: Basics

Jing Gao
SUNY Buffalo

Outline

- **Basics**
  - Motivation, definition, evaluation
- **Methods**
  - Partitional
  - Hierarchical
  - Density-based
  - Mixture model
  - Spectral methods
- **Advanced topics**
  - Clustering in MapReduce
  - Clustering ensemble
  - Semi-supervised clustering, subspace clustering, co-clustering, etc.

Readings

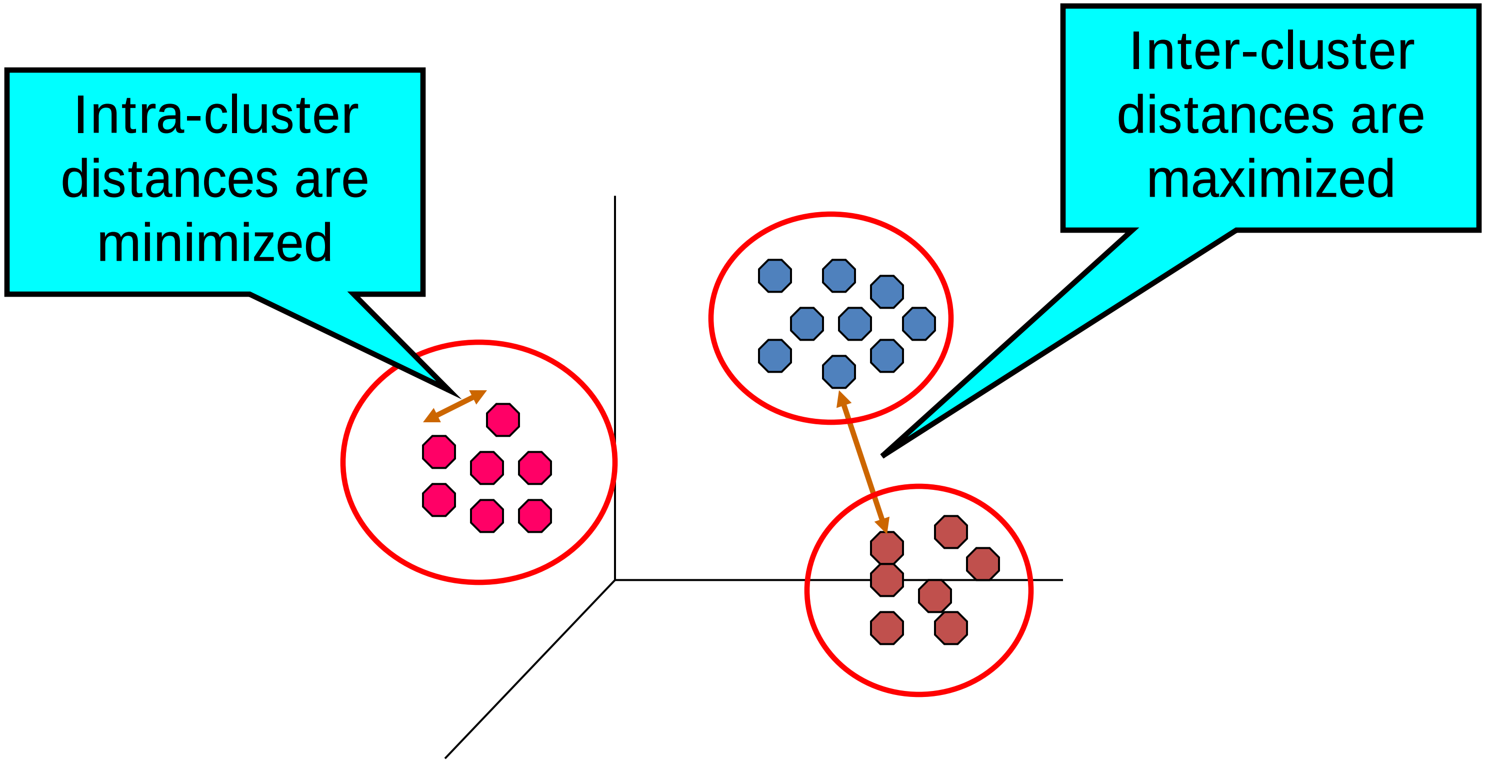
- Tan, Steinbach, Kumar, Chapters 8 and 9.
- Han, Kamber, Pei. Data Mining: Concepts and Techniques. Chapters 10 and 11.
- Additional readings posted on website

Clustering Basics

- Definition and Motivation
- Data Preprocessing and Similarity Computation
- Objective of Clustering
- Clustering Evaluation

Clustering

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



Intra-cluster distances are minimized

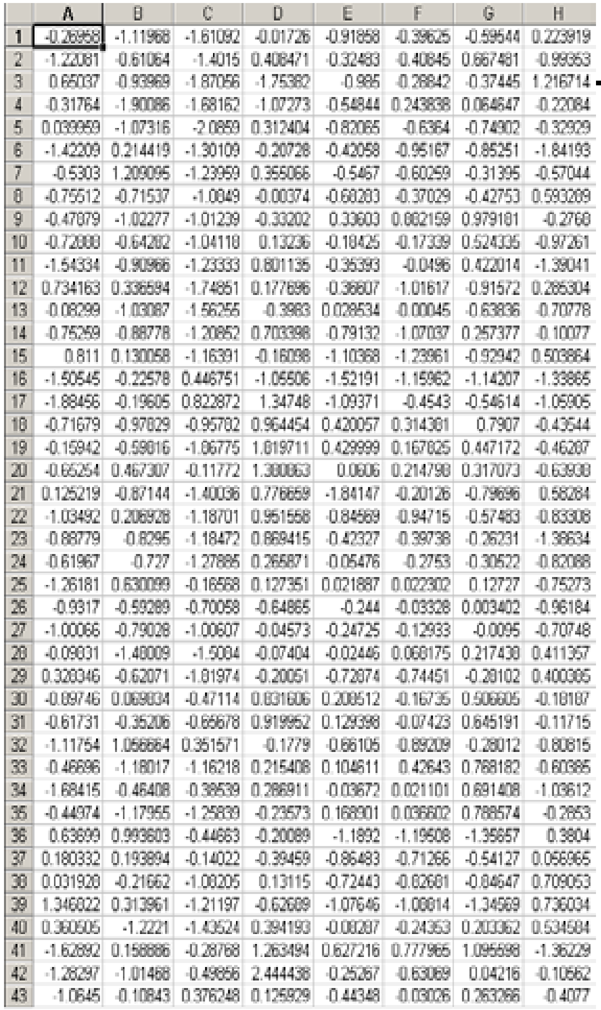
Inter-cluster distances are maximized

Application Examples

- A stand-alone tool: explore data distribution
- A preprocessing step for other algorithms
- Pattern recognition, spatial data analysis, image processing, market research, WWW, …
  - Cluster documents
  - Cluster web log data to discover groups of similar access patterns

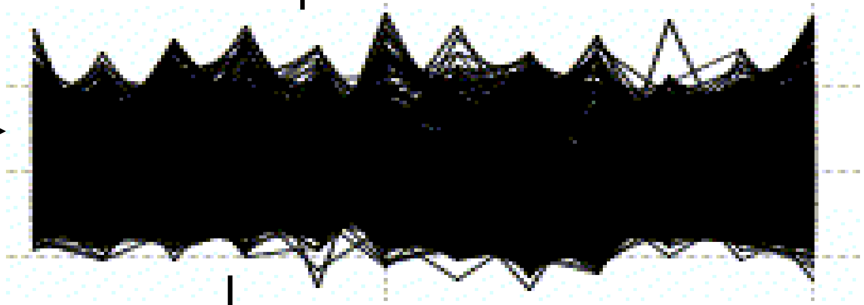
Clustering Co-expressed Genes

Gene Expression Data Matrix

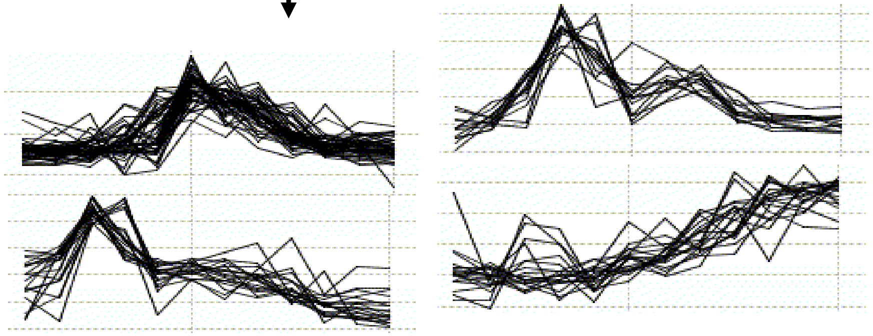


	A	B	C	D	E	F	G	H
1	-0.26958	-1.11968	-1.61092	-0.01726	-0.91858	-0.39625	-0.59544	0.223919
2	-1.22081	-0.61064	-1.4015	0.408471	-0.32483	-0.40845	0.667481	-0.99363
3	0.65037	-0.93969	-1.87056	-1.75382	-0.985	-0.28842	-0.37445	1.216714
4	-0.31764	-1.90086	-1.68162	-1.07273	-0.54844	0.243838	0.064647	-0.22084
5	0.039969	-1.07316	-2.0859	0.312404	-0.82065	-0.6364	-0.74902	-0.32929
6	-1.42209	0.214419	-1.30109	-0.20728	-0.42058	-0.95167	-0.85251	-1.84193
7	-0.5303	1.209095	-1.23959	0.355066	-0.5467	-0.60259	-0.31395	-0.57044
8	-0.75512	-0.71537	-1.0049	-0.00374	-0.68283	-0.37029	-0.42753	0.593289
9	-0.47079	-1.02277	-1.01239	-0.33202	0.33603	0.082159	0.979181	-0.2768
10	-0.72088	-0.64282	-1.04118	0.13236	-0.18425	-0.17339	0.524335	-0.97261
11	-1.54334	-0.90966	-1.23333	0.801135	-0.35393	-0.0496	0.422014	-1.39041
12	0.734163	0.336594	-1.74851	0.177696	-0.36807	-1.01617	-0.91572	0.285304
13	-0.08299	-1.03087	-1.56255	-0.3983	0.028534	-0.00045	-0.63836	-0.70778
14	-0.75259	-0.88778	-1.20852	0.703398	-0.79132	-1.07037	0.257377	-0.10077
15	0.811	0.130058	-1.16391	-0.16098	-1.10368	-1.23961	-0.92942	0.503864
16	-1.50545	-0.22578	0.446751	-1.05506	-1.52191	-1.15962	-1.14207	-1.33865
17	-1.88456	-0.19606	0.822872	1.34748	-1.09371	-0.4543	-0.54614	-1.05905
18	-0.71679	-0.97029	-0.95782	0.964454	0.420057	0.314381	0.7907	-0.43544
19	-0.15942	-0.59016	-1.06775	1.019711	0.429999	0.167025	0.447172	-0.46207
20	-0.65254	0.467307	-0.11772	1.380863	0.0606	0.214798	0.317073	-0.63938
21	0.125219	-0.87144	-1.40036	0.776659	-1.84147	-0.20126	-0.79696	0.58284
22	-1.03492	0.206928	-1.18701	0.951558	-0.84569	-0.94715	-0.57483	-0.83308
23	-0.88779	-0.8295	-1.18472	0.869415	-0.42327	-0.39738	-0.26231	-1.38634
24	-0.61967	-0.727	-1.27885	0.265871	-0.05476	-0.2753	-0.30522	-0.82088
25	-1.26181	0.630099	-0.16568	0.127351	0.021887	0.022302	0.12727	-0.75273
26	-0.9317	-0.59289	-0.70058	-0.64865	-0.244	-0.03328	0.003402	-0.96184
27	-1.00066	-0.79028	-1.00607	-0.04573	-0.24725	-0.12933	-0.0095	-0.70748
28	-0.09031	-1.40009	-1.5004	-0.07404	-0.02446	0.068175	0.217438	0.411357
29	0.320346	-0.62071	-1.01974	-0.20051	-0.72874	-0.74451	-0.28102	0.400305
30	-0.09746	0.069034	-0.47114	0.831606	0.208512	-0.16735	0.506605	-0.18187
31	-0.61731	-0.35206	-0.65678	0.919952	0.129398	-0.07423	0.645191	-0.11715
32	-1.11754	1.056664	0.351571	-0.1779	-0.66105	-0.89209	-0.28012	-0.80815
33	-0.46696	-1.18017	-1.16218	0.215408	0.104611	0.42643	0.768182	-0.60385
34	-1.68415	-0.46408	-0.38539	0.286911	-0.03672	0.021101	0.691408	-1.03612
35	-0.44974	-1.17955	-1.25839	-0.23573	0.168901	0.036602	0.788574	-0.2853
36	0.63899	0.993603	-0.44663	-0.20089	-1.1892	-1.19508	-1.35857	0.3804
37	0.180332	0.193894	-0.14022	-0.39459	-0.86483	-0.71266	-0.54127	0.056965
38	0.031928	-0.21662	-1.00205	0.13115	-0.72443	-0.02681	-0.04647	0.709053
39	1.346022	0.313961	-1.21197	-0.62609	-1.07646	-1.00814	-1.34569	0.736034
40	0.360505	-1.2221	-1.43524	0.394193	-0.08287	-0.24353	0.203362	0.534584
41	-1.62892	0.158886	-0.28768	1.263494	0.627216	0.777965	1.095698	-1.36229
42	-1.28297	-1.01468	-0.49856	2.444438	-0.25267	-0.63069	0.04216	-0.10562
43	-1.0645	-0.10843	0.376248	0.125929	-0.44348	-0.03026	0.263266	-0.4077

Gene Expression Patterns



Co-expressed Genes

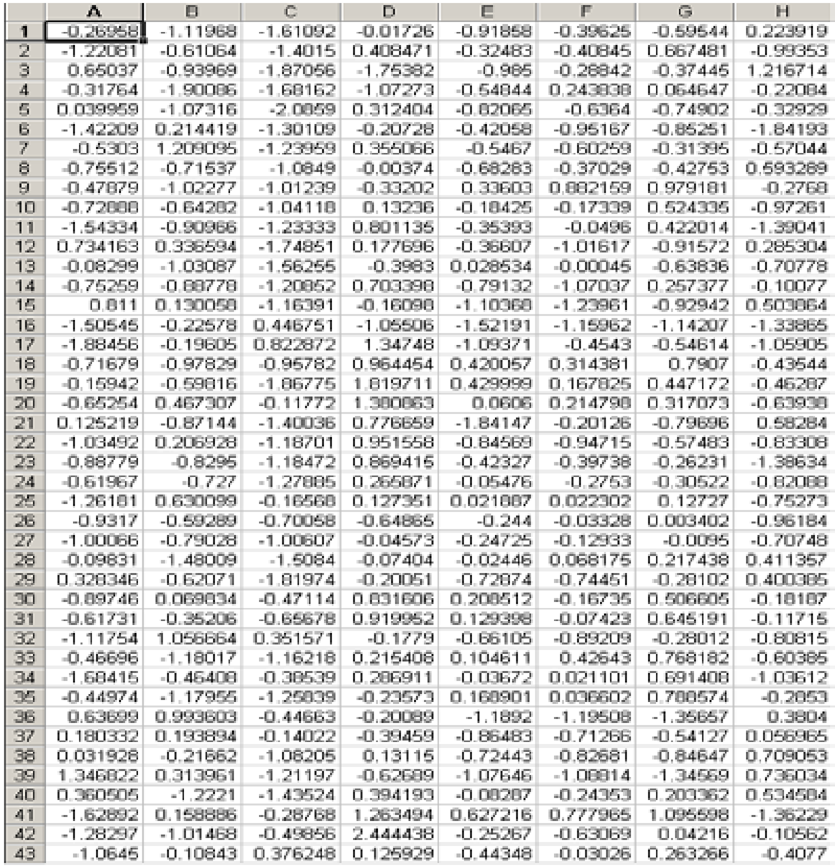


Why looking for co-expressed genes?

¾ *Co-expression indicates co-function;*

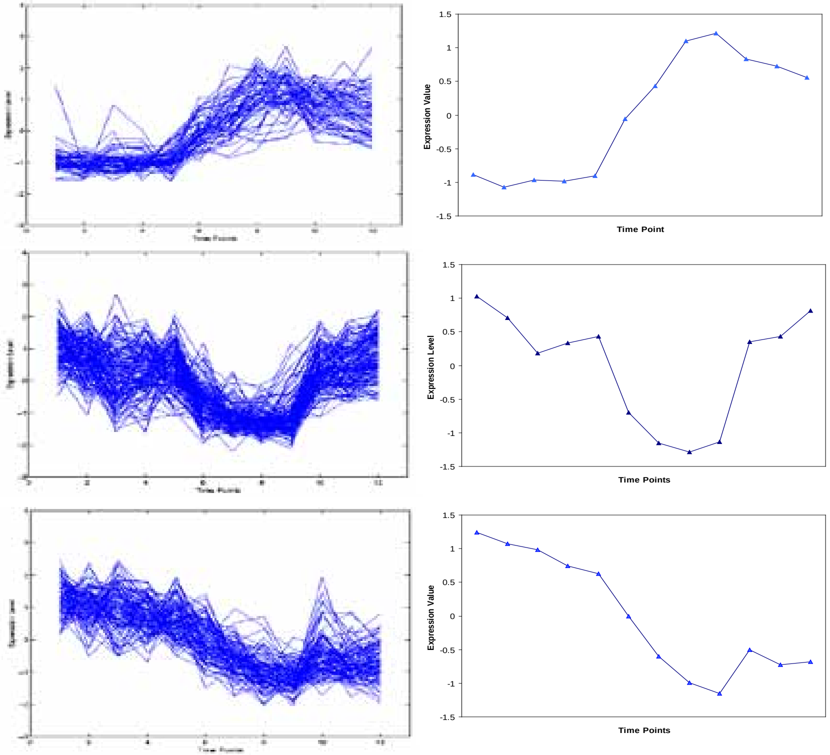
¾ *Co-expression also indicates co-regulation.*

Gene-based Clustering



	A	B	C	D	E	F	G	H
1	-0.26958	-1.11968	-1.61092	-0.01726	-0.91858	-0.39625	-0.59544	0.223919
2	-1.22081	-0.61064	-1.4015	0.408471	-0.32483	-0.40845	0.667481	-0.99353
3	0.65037	-0.93969	-1.87056	-1.75382	-0.985	-0.28842	-0.37445	1.216714
4	-0.31764	-1.90086	-1.68162	-1.07273	-0.54844	0.243838	0.064647	-0.22084
5	0.039959	-1.07316	-2.0859	0.312404	-0.82065	-0.6364	-0.74902	-0.32929
6	-1.42209	0.214419	-1.30109	-0.20728	-0.42058	-0.95167	-0.85251	-1.84193
7	-0.5303	1.209095	-1.23959	0.355066	-0.5467	-0.60259	-0.31395	-0.57044
8	-0.75512	-0.71537	-1.0849	-0.00374	-0.68283	-0.37029	-0.42753	0.593289
9	-0.47879	-1.02277	-1.01239	-0.33202	0.33603	0.882159	0.979181	-0.2768
10	-0.72888	-0.64282	-1.04118	0.13236	-0.18425	-0.17339	0.524335	-0.97261
11	-1.54334	-0.90966	-1.23333	0.801135	-0.35393	-0.0496	0.422014	-1.39041
12	0.734163	0.336594	-1.74851	0.177696	-0.36607	-1.01617	-0.91572	0.285304
13	-0.08299	-1.03087	-1.56255	-0.3983	0.028534	-0.00045	-0.63836	-0.70778
14	-0.75259	-0.88778	-1.20852	0.703398	-0.79132	-1.07037	0.257377	-0.10077
15	0.811	0.130058	-1.16391	-0.16098	-1.10368	-1.23961	-0.92942	0.503864
16	-1.50545	-0.22578	0.446751	-1.05506	-1.52191	-1.15962	-1.14207	-1.33865
17	-1.88456	-0.19605	0.822872	1.34748	-1.09371	-0.4543	-0.54614	-1.05905
18	-0.71679	-0.97829	-0.95782	0.964454	0.420057	0.314381	0.7907	-0.43544
19	-0.15942	-0.59816	-1.86775	1.819711	0.429999	0.167825	0.447172	-0.46287
20	-0.65254	0.467307	-0.11772	1.380863	0.0606	0.214798	0.317073	-0.63938
21	0.125219	-0.87144	-1.40036	0.776659	-1.84147	-0.20126	-0.79696	0.58284
22	-1.03492	0.206928	-1.18701	0.951558	-0.84569	-0.94715	-0.57483	-0.83308
23	-0.88779	-0.8295	-1.18472	0.869415	-0.42327	-0.39738	-0.26231	-1.38634
24	-0.61967	-0.727	-1.27885	0.265871	-0.05476	-0.2753	-0.30522	-0.82088
25	-1.26181	0.630099	-0.16568	0.127351	0.021887	0.022302	0.12727	-0.75273
26	-0.9317	-0.59289	-0.70058	-0.64865	-0.244	-0.03328	0.003402	-0.96184
27	-1.00066	-0.79028	-1.00607	-0.04573	-0.24725	-0.12933	-0.0095	-0.70748
28	-0.09831	-1.48009	-1.5084	-0.07404	-0.02446	0.068175	0.217438	0.411357
29	0.328346	-0.62071	-1.81974	-0.20051	-0.72874	-0.74451	-0.28102	0.400385
30	-0.89746	0.069834	-0.47114	0.831606	0.208512	-0.16735	0.506605	-0.18187
31	-0.61731	-0.35206	-0.65678	0.919952	0.129398	-0.07423	0.645191	-0.11715
32	-1.11754	1.056664	0.351571	-0.1779	-0.66105	-0.89209	-0.28012	-0.80815
33	-0.46696	-1.18017	-1.16218	0.215408	0.104611	0.42643	0.768182	-0.60385
34	-1.68415	-0.46408	-0.38539	0.286911	-0.03672	0.021101	0.691408	-1.03612
35	-0.44974	-1.17955	-1.25839	-0.23573	0.168901	0.036602	0.788574	-0.2853
36	0.63699	0.993603	-0.44663	-0.20089	-1.1892	-1.19508	-1.35657	0.3804
37	0.180332	0.193894	-0.14022	-0.39459	-0.86483	-0.71266	-0.54127	0.056965
38	0.031928	-0.21662	-1.08205	0.13115	-0.72443	-0.82681	-0.84647	0.709053
39	1.346822	0.313961	-1.21197	-0.62689	-1.07646	-1.08814	-1.34569	0.736034
40	0.360505	-1.2221	-1.43524	0.394193	-0.08287	-0.24353	0.203362	0.534584
41	-1.62892	0.158886	-0.28768	1.263494	0.627216	0.777965	1.095598	-1.36229
42	-1.28297	-1.01468	-0.49856	2.444438	-0.25267	-0.63069	0.04216	-0.10562
43	-1.0645	-0.10843	0.376248	0.125929	-0.44348	-0.03026	0.263266	-0.4077

Iyer’s data [2]



Examples of co-expressed genes and coherent patterns in gene expression data

¢ [2] Iyer, V.R. et al. The transcriptional program in the response of human fibroblasts to serum. *Science*, 283:83–87, 1999.

Other Applications

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Climate: understanding earth climate, find patterns of atmosphere and ocean

Two Important Aspects

- **Properties of input data**
  - Define the similarity or dissimilarity between points
- **Requirement of clustering**
  - Define the objective and methodology

Clustering Basics

- Definition and Motivation
- Data Preprocessing and Distance computation
- Objective of Clustering
- Clustering Evaluation

Data Representation

- Data: Collection of data objects and their attributes

- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as dimension, variable, field, characteristic, or feature

- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Data Matrix

- **Represents *n* objects with *p* attributes**
  - An *n* by *p* matrix

The value of the *i*-th object on the *f*-th attribute

$$\begin{array}{cc} & \textbf{Attributes} \\ \textbf{Objects} & \begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix} \end{array}$$

Gene Expression Data

	condition 1	condition 2	condition 3	condition 4	• condition…
gene 1	0.13	0.72	0.1	0.57	
gene 2	0.34	1.58	1.05	1.15	
gene 3	0.43	1.1	0.97	1	
gene 4	1.22	0.97	1	0.85	
gene 5	-0.89	1.21	1.29	1.08	
gene 6	1.1	1.45	1.44	1.12	
gene 7	0.83	1.15	1.1	1	
gene 8	0.87	1.32	1.35	1.13	
gene 9	-0.33	1.01	1.38	1.21	
gene 10	0.10	0.85	1.03	1	
gene …					

Clustering genes

- Genes are objects
- Experiment conditions are attributes
- Find genes with similar behavior

Similarity and Dissimilarity

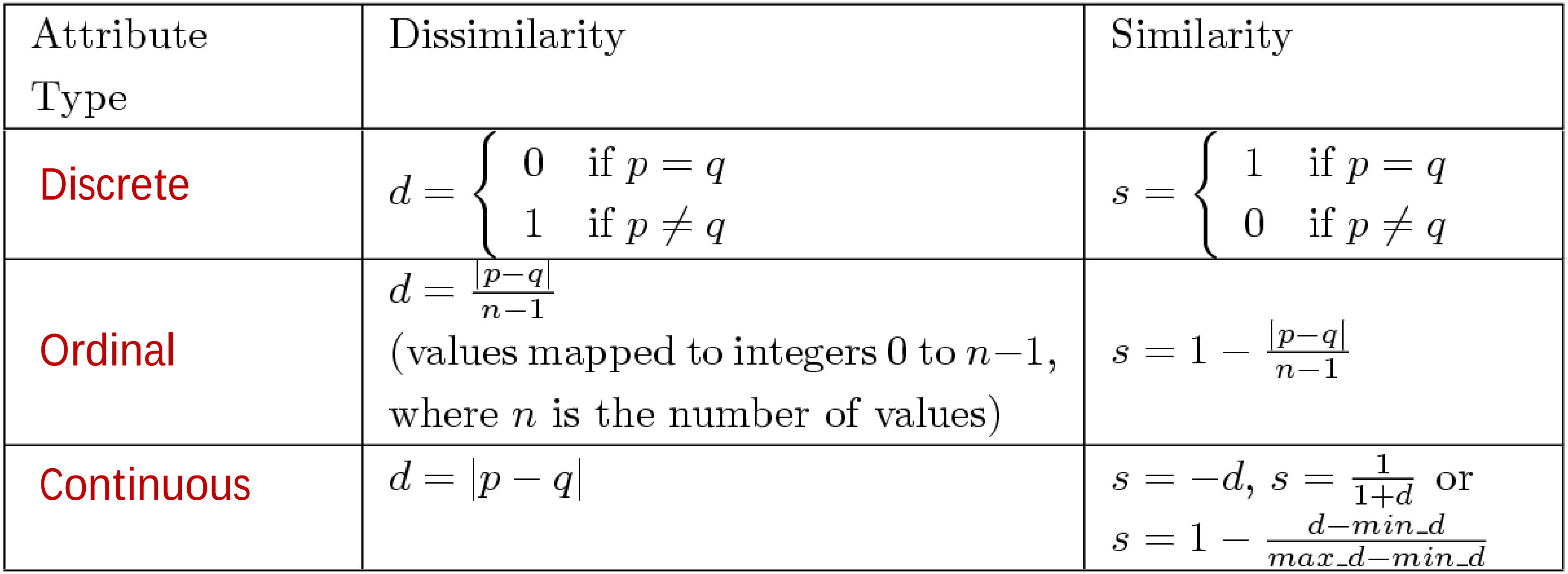
- **Similarity**
  - Numerical measure of how alike two data objects are
  - Is higher when objects are more alike
  - Often falls in the range [0,1]
- **Dissimilarity**
  - Numerical measure of how different are two data objects
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies

Types of Attributes

- **Discrete**
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Note: binary attributes are a special case of discrete attributes
- **Ordinal**
  - Has only a finite or countably infinite set of values
  - Order of values is important
  - Examples: rankings (e.g., pain level 1-10), grades (A, B, C, D)
- **Continuous**
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight
  - Continuous attributes are typically represented as floating-point variables

Similarity/Dissimilarity for Simple Attributes

p and *q* are the attribute values for two data objects.



Attribute Type	Dissimilarity	Similarity
Discrete	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Continuous	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Dissimilarity and similarity between *p* and *q*

Distance Matrix

- **Represents pairwise distance in *n* objects**
  - An *n* by *n* matrix
  - *d*(*i*,*j*): distance or dissimilarity between objects *i* and *j*
  - Nonnegative
  - Close to 0: similar

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Data Matrix -> Distance Matrix

	s 1	s 2	s 3	s 4	…
g 1	0.13	0.72	0.1	0.57	
g 2	0.34	1.58	1.05	1.15	
g 3	0.43	1.1	0.97	1	
g 4	1.22	0.97	1	0.85	
g 5	-0.89	1.21	1.29	1.08	
g 6	1.1	1.45	1.44	1.12	
g 7	0.83	1.15	1.1	1	
g 8	0.87	1.32	1.35	1.13	
g 9	-0.33	1.01	1.38	1.21	
g 10	0.10	0.85	1.03	1	
…					

Original Data Matrix

	g 1	g 2	g 3	g 4	…
g 1	0	*d*(1,2)	*d*(1,3)	*d*(1,4)	
g 2		0	*d*(2,3)	*d*(2,4)	
g 3			0	*d*(3,4)	
g 4				0	
…					

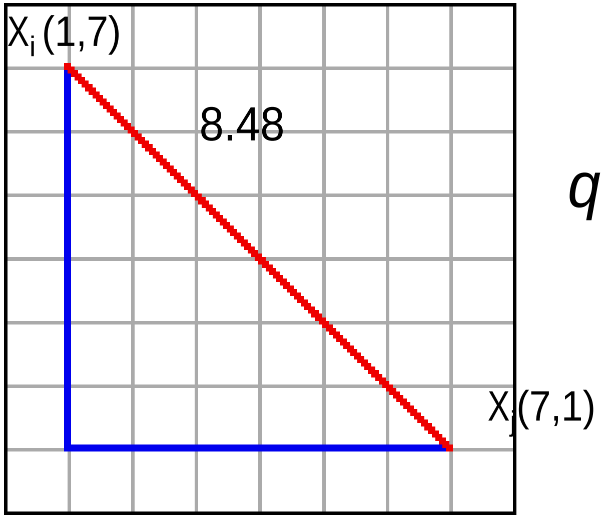
Distance Matrix

Minkowski Distance—Continuous Attribute

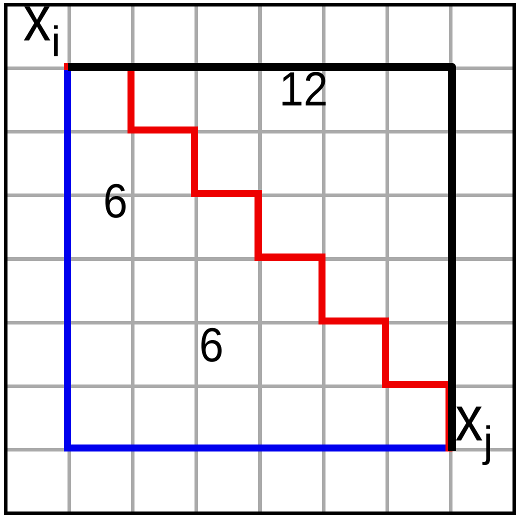
- Minkowski distance: a generalization

$$d(i,j) = \sqrt[q]{|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + \ldots + |x_{i_p} - x_{j_p}|^q} \quad (q > 0)$$

- If q = 2, d is Euclidean distance
- If q = 1, d is Manhattan distance



q=2



q=1

Standardization

- **Calculate the mean absolute deviation**

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \dots + x_{nf}).$$

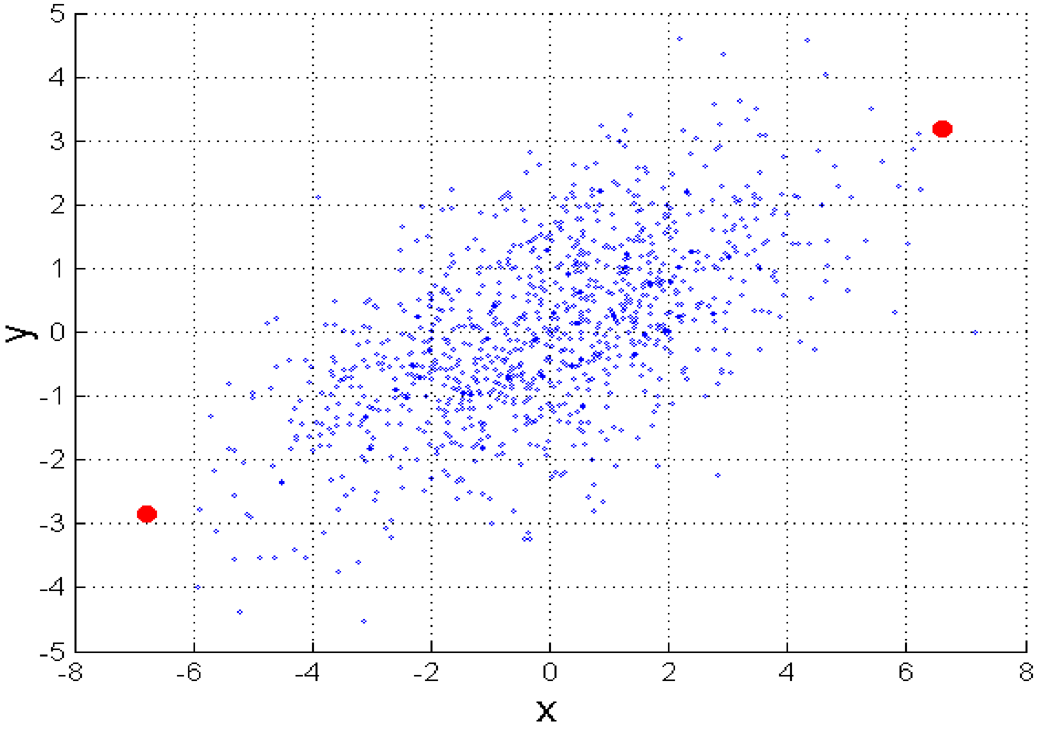
$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f|)$$

- **Calculate the standardized measurement (z-score)**

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

Mahalanobis Distance

$$d(p,q) = (p - q)\,\text{å}^{-1}(p - q)^T$$

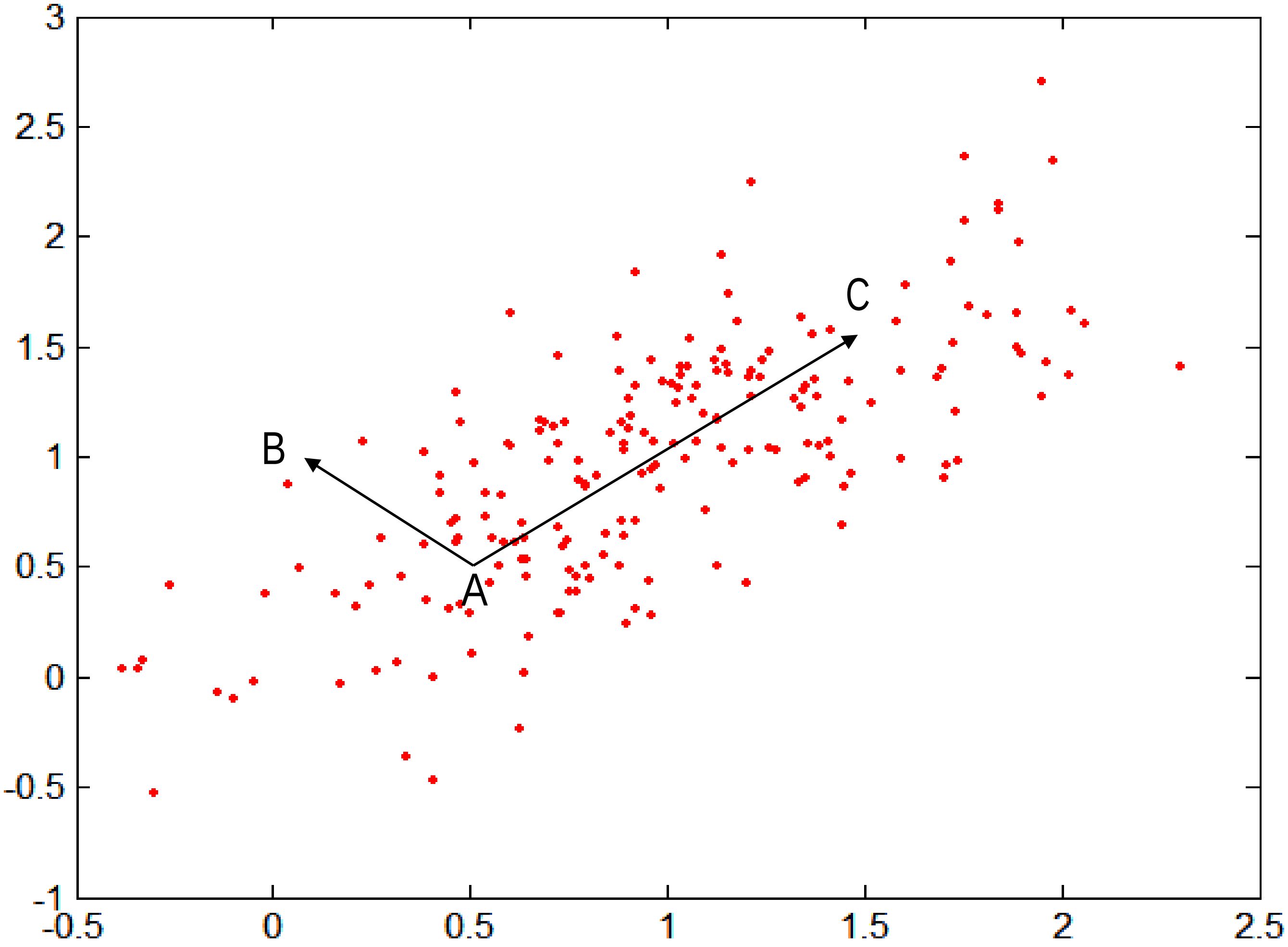


S is the covariance matrix of the input data *X*

$$S_{j,k} = \frac{1}{n - 1} \overset{n}{\underset{i=1}{\text{å}}} (X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k)$$

For red points, the Euclidean distance is 14.7, Mahalanobis distance is 6.

Mahalanobis Distance



Covariance Matrix:

$$S = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4

Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well known properties

1. *d(p, q) ³ 0* for all *p* and *q* and *d(p, q) = 0* only if *p* = *q*. (Positive definiteness)
2. *d(p, q) = d(q, p)* for all *p* and *q*. (Symmetry)
3. d*(p, r) £ d(p, q) + d(q, r)* for all points *p*, *q*, and *r*. (Triangle Inequality)

where *d(p, q)* is the distance (dissimilarity) between points (data objects), *p* and *q*.

- A distance that satisfies these properties is a metric

Similarity for Binary Attributes

- Common situation is that objects, *p* and *q*, have only binary attributes

- Compute similarities using the following quantities

  M_{01} = the number of attributes where p was 0 and q was 1
  M_{10} = the number of attributes where p was 1 and q was 0
  M_{00} = the number of attributes where p was 0 and q was 0
  M_{11} = the number of attributes where p was 1 and q was 1

- Simple Matching and Jaccard Coefficients

SMC = number of matches / total number of attributes
= $(M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

J = number of matches / number of not-both-zero attributes values
= $(M_{11}) / (M_{01} + M_{10} + M_{11})$

SMC versus Jaccard: Example

p = 1 0 0 0 0 0 0 0 0 0

q = 0 0 0 0 0 0 1 0 0 1

M_{01} = 2 (the number of attributes where p was 0 and q was 1)
M_{10} = 1 (the number of attributes where p was 1 and q was 0)
M_{00} = 7 (the number of attributes where p was 0 and q was 0)
M_{11} = 0 (the number of attributes where p was 1 and q was 1)

$SMC = (M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00}) = (0+7) / (2+1+0+7) = 0.7$

$J = (M_{11}) / (M_{01} + M_{10} + M_{11}) = 0 / (2 + 1 + 0) = 0$

Document Data

- Each document becomes a `term' vector,
  - each term is a component (attribute) of the vector,
  - the value of each component is the number of times the corresponding term occurs in the document.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Cosine Similarity

- If d_1 and d_2 are two document vectors, then

$$\cos(d_1, d_2) = (d_1 \cdot d_2) / ||d_1||\ ||d_2|| ,$$

where $\cdot$ indicates vector dot product and $|| d ||$ is the length of vector d.

- Example:

d_1 = **3 2 0 5 0 0 0 2 0 0**

d_2 = **1 0 0 0 0 0 0 1 0 2**

$d_1 \cdot d_2$= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5

$||d_1||$ = $(3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5}$ = $(42)^{0.5}$ = 6.481

$||d_2||$ = $(1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5}$ = $(6)^{0.5}$ = 2.245

$\cos(d_1, d_2) = .3150$

Correlation

- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, p and q, and then take their dot product (continuous attributes)

$$p'_k = (p_k - mean(p)) / std(p)$$

$$q'_k = (q_k - mean(q)) / std(q)$$

$$s(p,q) = p' \cdot q'$$

Common Properties of a Similarity

- Similarities, also have some well known properties.

  1. $s(p, q) = 1$ (or maximum similarity) only if $p = q$.

  2. $s(p, q) = s(q, p)$ for all p and q. (Symmetry)

  where $s(p, q)$ is the similarity between points (data objects), p and q.

Characteristics of the Input Data Are Important

- Sparseness
- Attribute type
- Type of Data
- Dimensionality
- Noise and Outliers
- Type of Distribution
- => Conduct preprocessing and select the appropriate dissimilarity or similarity measure
- => Determine the objective of clustering and choose the appropriate method

Clustering Basics

- Definition and Motivation
- Data Preprocessing and Distance computation
- Objective of Clustering
- Clustering Evaluation

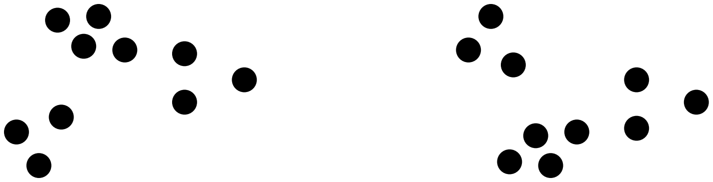
Considerations for Cluster Analysis

- **Partitioning criteria**
  - Single level vs. hierarchical partitioning (often, multi-level hierarchical partitioning is desirable)
- **Separation of clusters**
  - Exclusive (e.g., one customer belongs to only one region) vs. overlapping (e.g., one document may belong to more than one topic)
- **Hard versus fuzzy**
  - In fuzzy clustering, a point belongs to every cluster with some weight between 0 and 1
  - Weights must sum to 1
  - Probabilistic clustering has similar characteristics
- **Similarity measure and data types**
- **Heterogeneous versus homogeneous**
  - Cluster of widely different sizes, shapes, and densities

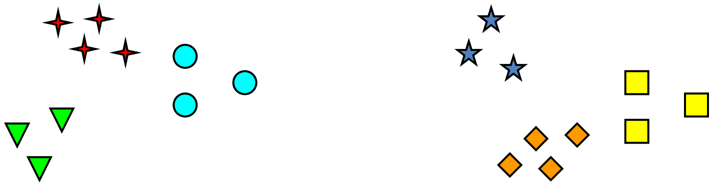
Requirements of Clustering

- Scalability
- Ability to deal with different types of attributes
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Discovery of clusters with arbitrary shape
- Insensitive to order of input records
- High dimensionality
- Incorporation of user-specified constraints
- Interpretability and usability
- **What clustering results we want to get?**

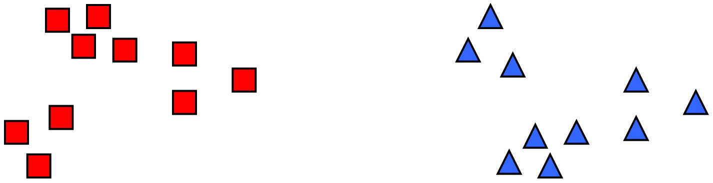
Notion of a Cluster can be Ambiguous



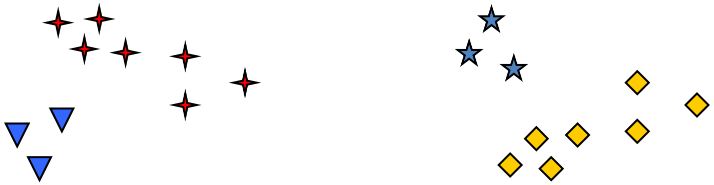
How many clusters?



Six Clusters

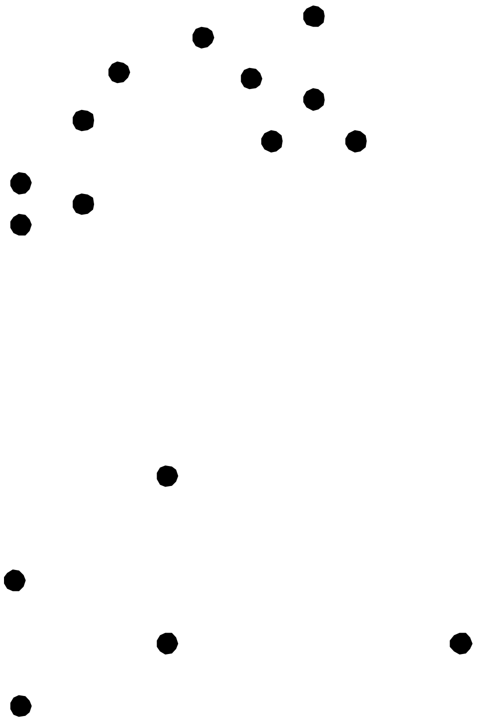


Two Clusters

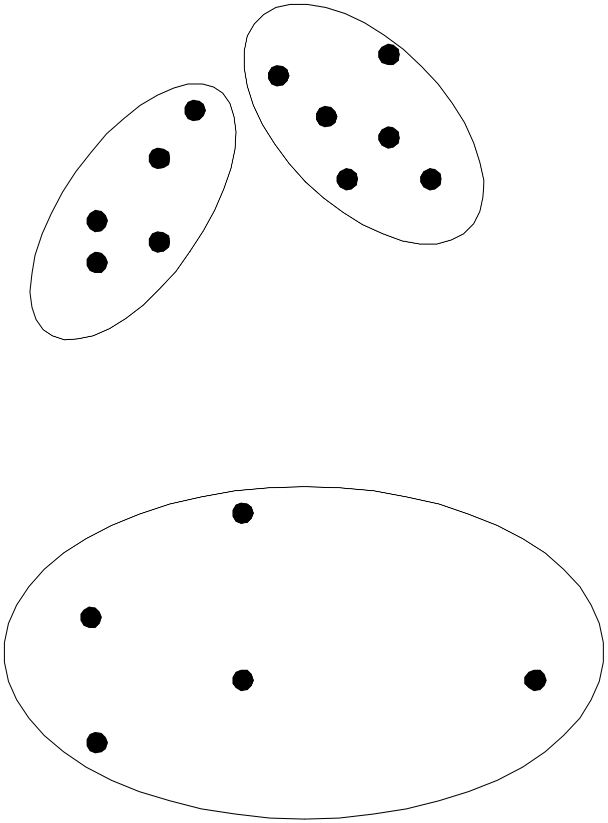


Four Clusters

Partitional Clustering

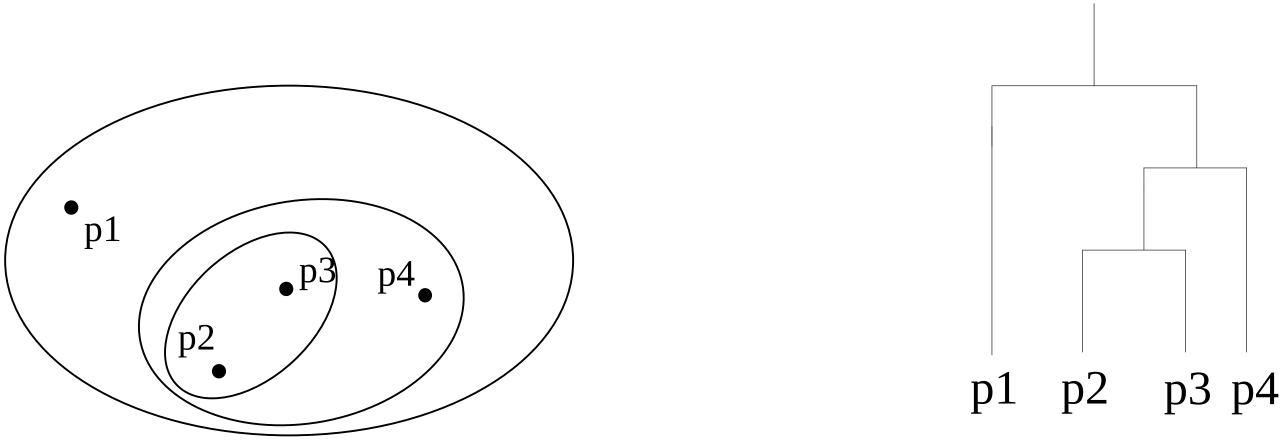


Input Data

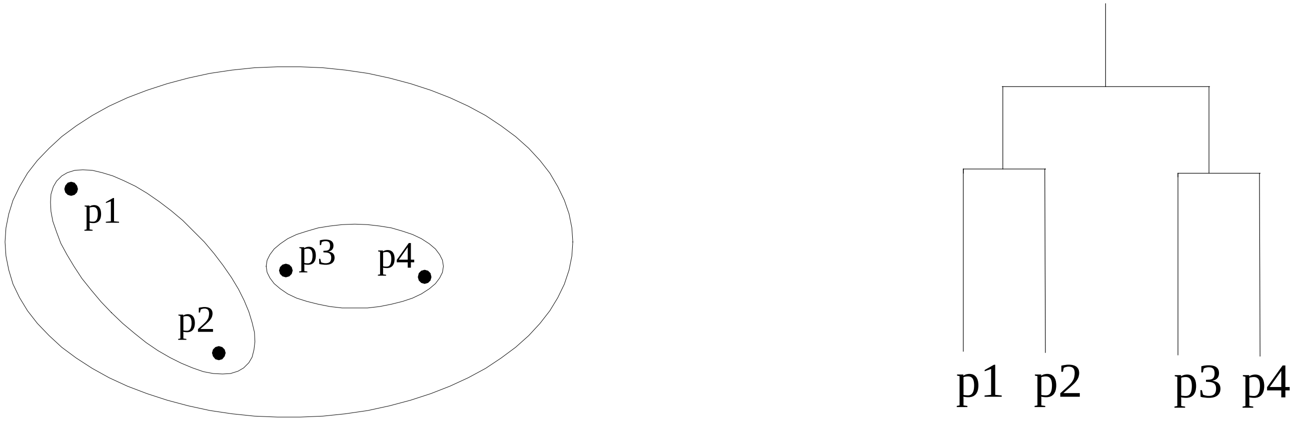


A Partitional Clustering

Hierarchical Clustering



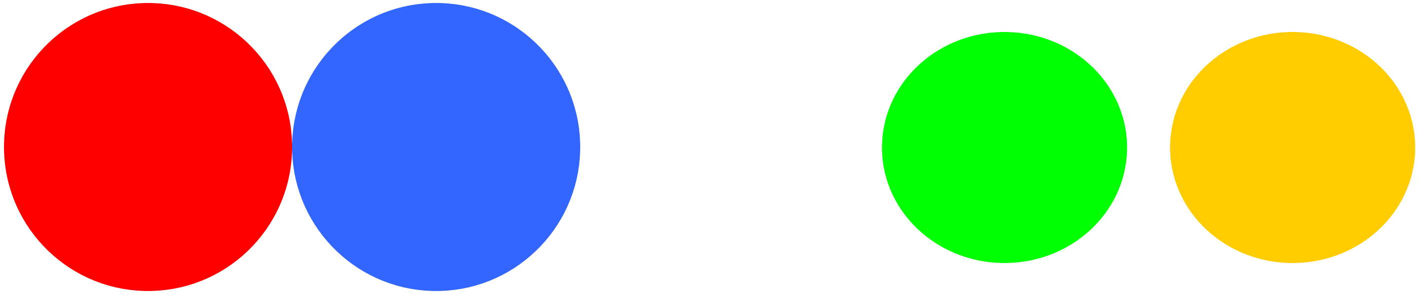
Clustering Solution 1



Clustering Solution 2

Types of Clusters: Center-Based

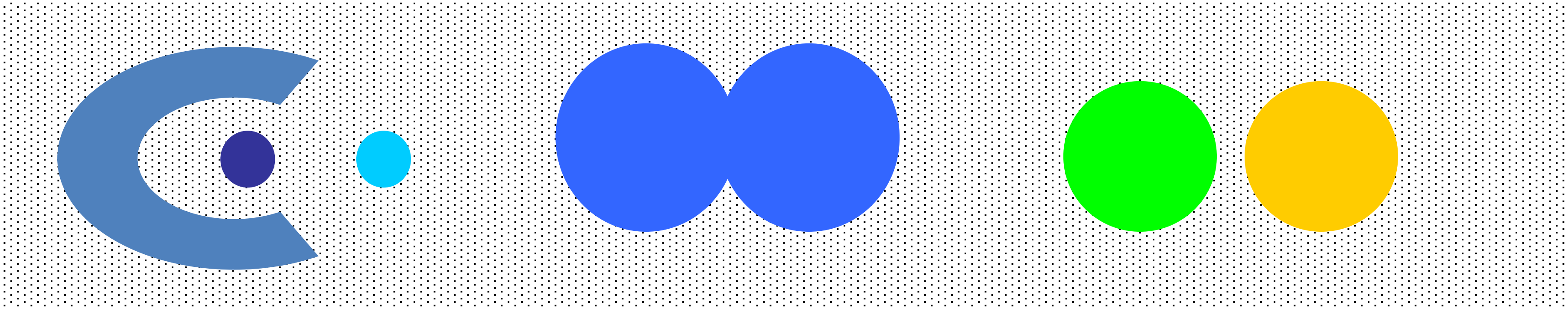
- **Center-based**
  - A cluster is a set of objects such that an object in a cluster is closer (more similar) to the "center" of a cluster, than to the center of any other cluster
  - The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most "representative" point of a cluster



4 center-based clusters

Types of Clusters: Density-Based

- **Density-based**
  - A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
  - Used when the clusters are irregular or intertwined, and when noise and outliers are present.



6 density-based clusters

Clustering Basics

- Definition and Motivation
- Data Preprocessing and Distance computation
- Objective of Clustering
- Clustering Evaluation

Cluster Validation

- **Cluster validation**
  - Quality: “goodness” of clusters
  - Assess the quality and reliability of clustering results

- **Why validation?**
  - To avoid finding clusters formed by chance
  - To compare clustering algorithms
  - To choose clustering parameters
    - e.g., the number of clusters

Aspects of Cluster Validation

- Comparing the clustering results to *ground truth* (externally known results)
  - External Index
- Evaluating the quality of clusters *without* reference to external information
  - Use only the data
  - Internal Index
- Determining the *reliability* of clusters
  - To what confidence level, the clusters are not formed by chance
  - Statistical framework

Comparing to Ground Truth

- **Notation**
  - N: number of objects in the data set
  - $P=\{P_1,\ldots,P_s\}$: the set of "ground truth" clusters
  - $C=\{C_1,\ldots,C_t\}$: the set of clusters reported by a clustering algorithm
- **The "incidence matrix"**
  - N ´ N (both rows and columns correspond to objects)
  - $P_{ij} = 1$ if O_i and O_j belong to the same "ground truth" cluster in P; $P_{ij}=0$ otherwise
  - $C_{ij} = 1$ if O_i and O_j belong to the same cluster in C; $C_{ij}=0$ otherwise

Rand Index and Jaccard Coefficient

- A pair of data object (O_i,O_j) falls into one of the following categories
  - SS: C_{ij}=1 and P_{ij}=1; (agree)
  - DD: C_{ij}=0 and P_{ij}=0; (agree)
  - SD: C_{ij}=1 and P_{ij}=0; (disagree)
  - DS: C_{ij}=0 and P_{ij}=1; (disagree)
- **Rand index** $$Rand = \frac{|Agree|}{|Agree| + |Disagree|} = \frac{|SS| + |DD|}{|SS| + |SD| + |DS| + |DD|}$$
  - may be dominated by DD
- **Jaccard Coefficient** $$Jaccard\ \ coefficient = \frac{|SS|}{|SS| + |SD| + |DS|}$$

Clustering

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

Groundtruth

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

Clustering

Ground truth

	Same Cluster	Different Cluster
Same Cluster	9	4
Different Cluster	4	8

$$Rand = \frac{|SS| + |DD|}{|SS| + |SD| + |DS| + |DD|} = \frac{17}{25}$$

$$Jaccard = \frac{|SS|}{|SS| + |SD| + |DS|} = \frac{9}{17}$$

Entropy and Purity

- **Notation**
  - $|C_k \cap P_j|$ the number of objects in both the *k*-th cluster of the clustering solution and *j*-th cluster of the groundtruth
  - $|C_k|$ the number of objects in the *k*-th cluster of the clustering solution
  - $|P_j|$ the number of objects in the *j*-th cluster of the groundtruth
- **Purity** $Purity = \frac{1}{N}\sum_k \max_j |C_k \cap P_j|$
- **Normalized Mutual Information**

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

$$H(C) = \sum_k \frac{|C_k|}{N}\log\frac{|C_k|}{N} \qquad H(P) = \sum_j \frac{|P_j|}{N}\log\frac{|P_j|}{N}$$

Example

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

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

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

$$= 0.7203$$

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

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

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

Internal Index

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

Cohesion and Separation

- **Cohesion** is measured by the within cluster sum of squares

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

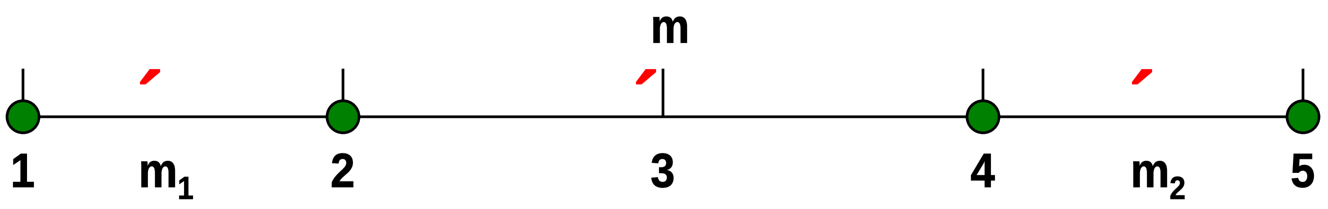
- **Separation** is measured by the between cluster sum of squares

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

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

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

Example



K=1 :

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

$$BSS = 4´\ (3-3)^2 = 0$$

$$Total = 10 + 0 = 10$$

K=2 :

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

$$BSS = 2´\ (3-1.5)^2 + 2´\ (4.5-3)^2 = 9$$

$$Total = 1 + 9 = 10$$

K=4:

$$WSS = (1-1)^2 + (2-2)^2 + (4-4)^2 + (5-5)^2 = 0$$

$$BSS = 1´\ (1-3)^2 + 1´\ (2-3)^2 + 1´\ (4-3)^2 + 1´\ (5-3)^2 = 10$$

$$Total = 0 + 10 = 10$$

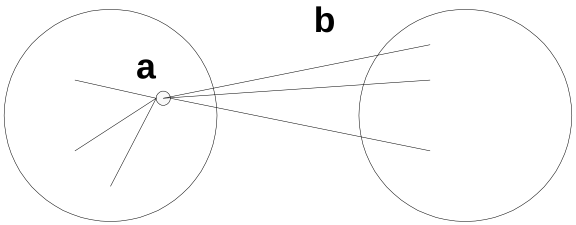
Silhouette Coefficient

- Silhouette Coefficient combines ideas of both cohesion and separation

- For an individual point, *i*
  - Calculate ***a*** = average distance of *i* to the points in its cluster
  - Calculate ***b*** = min (average distance of *i* to points in another cluster)
  - The **silhouette coefficient** for a point is then given by

    s = 1 – a/b   if a < b,   (s = b/a - 1   if a ³ b, not the usual case)

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



- Can calculate the Average Silhouette width for a cluster or a clustering

Correlation with Distance Matrix

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

Correlation with Distance Matrix

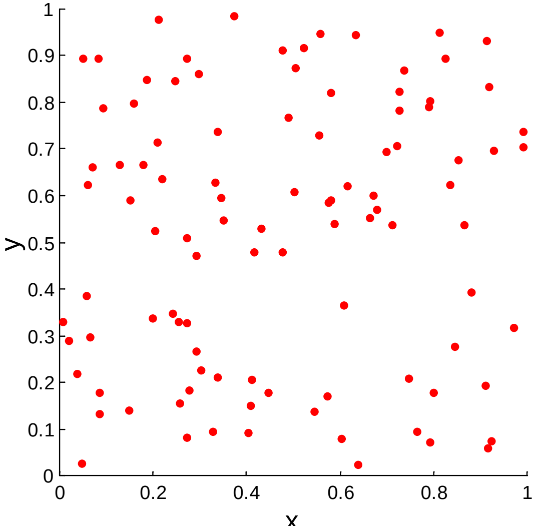
- Given Distance Matrix D = *{d_{11},d_{12}, …, d_{nn}}* and Incidence Matrix *C*= *{ c_{11}, c_{12},…, c_{nn}}* .

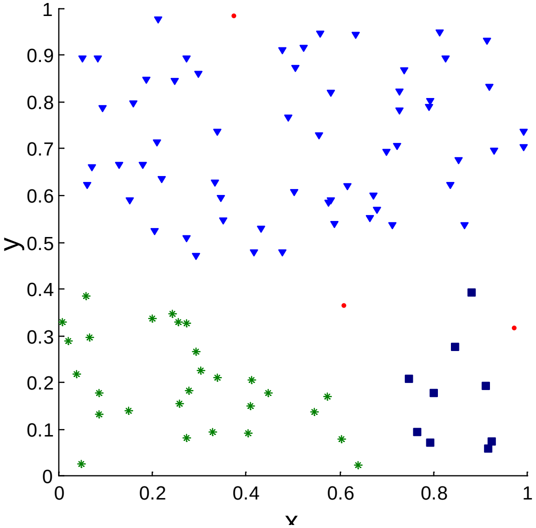
- Correlation *r* between *D* and *C* is given by

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

Are There Clusters in the Data?

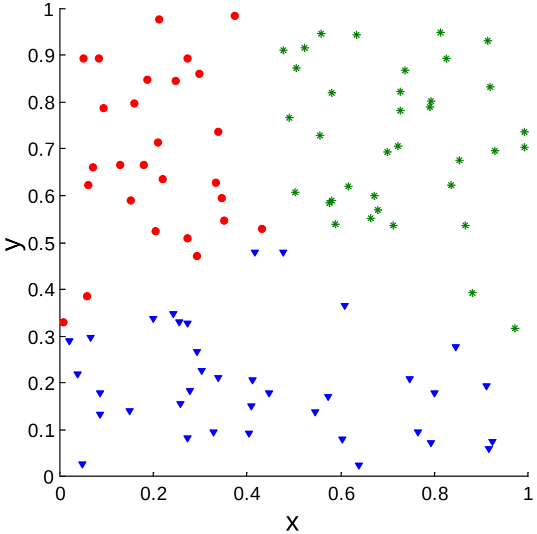
Random Points

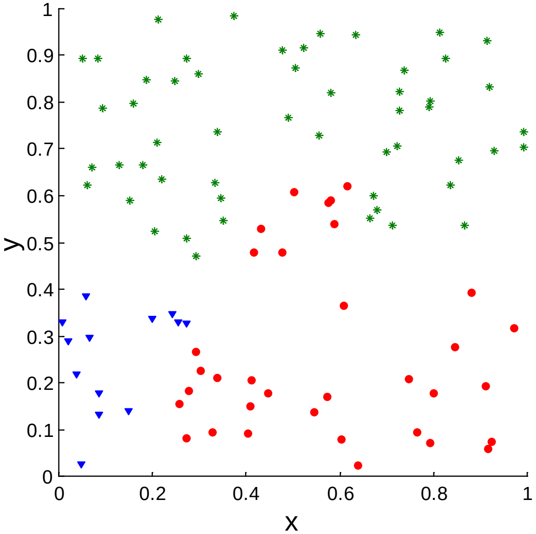




DBSCAN

K-means

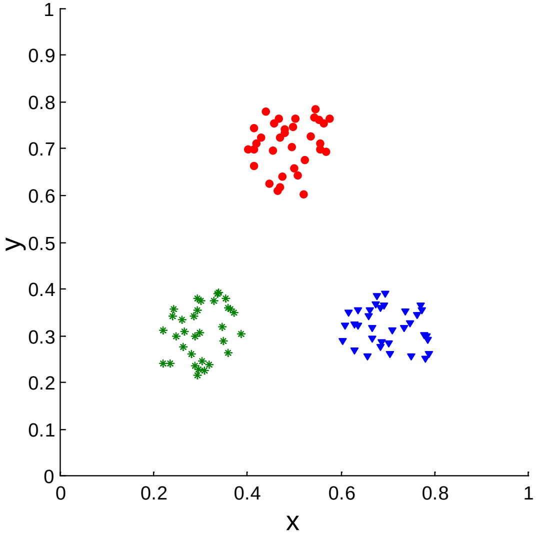




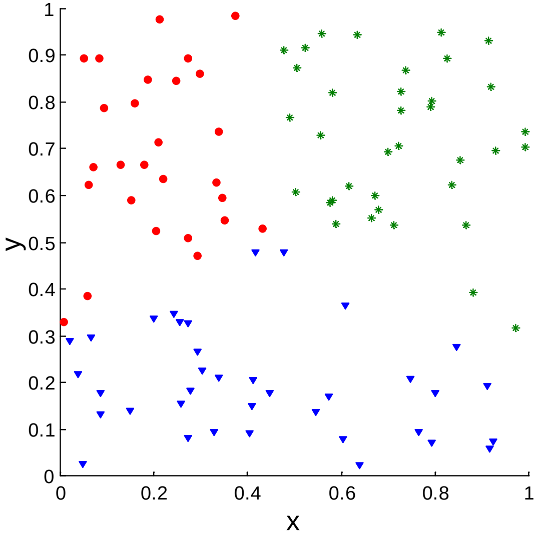
Complete Link

Measuring Cluster Validity Via Correlation

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



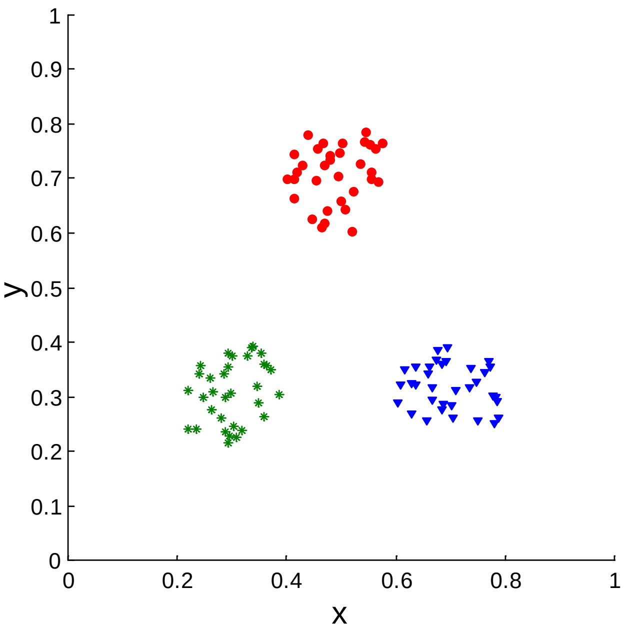
Corr = -0.9235

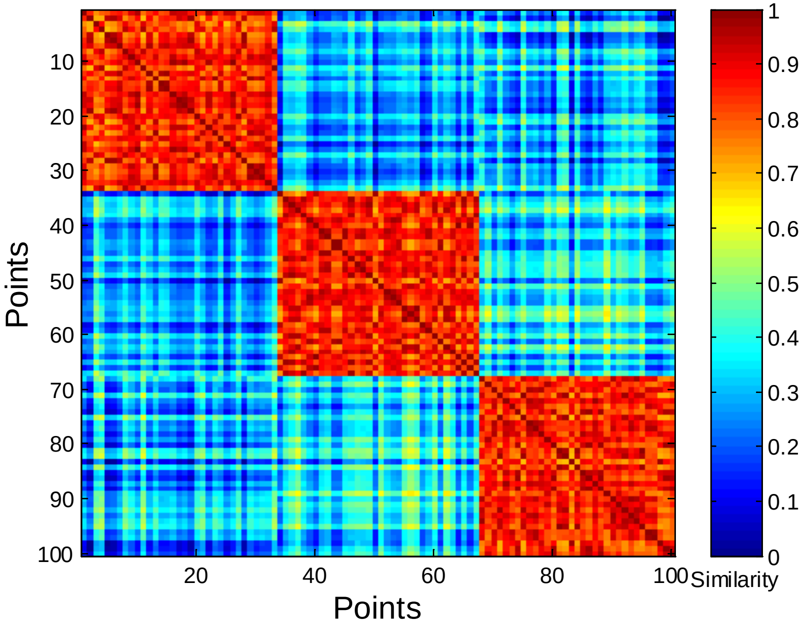


Corr = -0.5810

Using Similarity Matrix for Cluster Validation

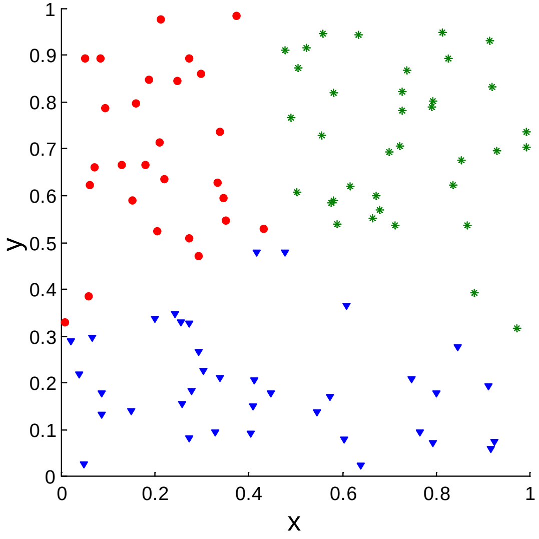
- Order the similarity matrix with respect to cluster labels and inspect visually.

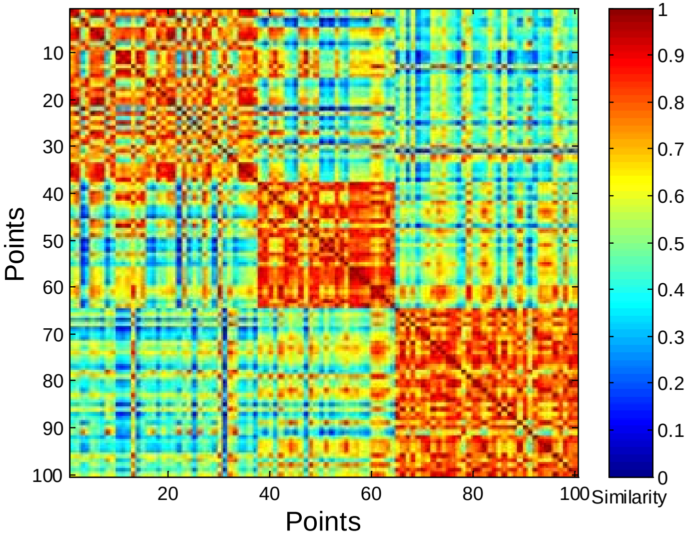




Using Similarity Matrix for Cluster Validation

- Clusters in random data are not so crisp



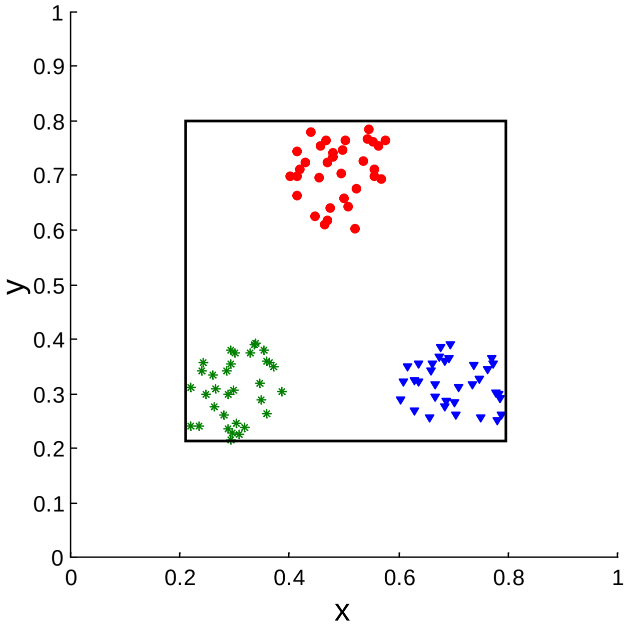


Reliability of Clusters

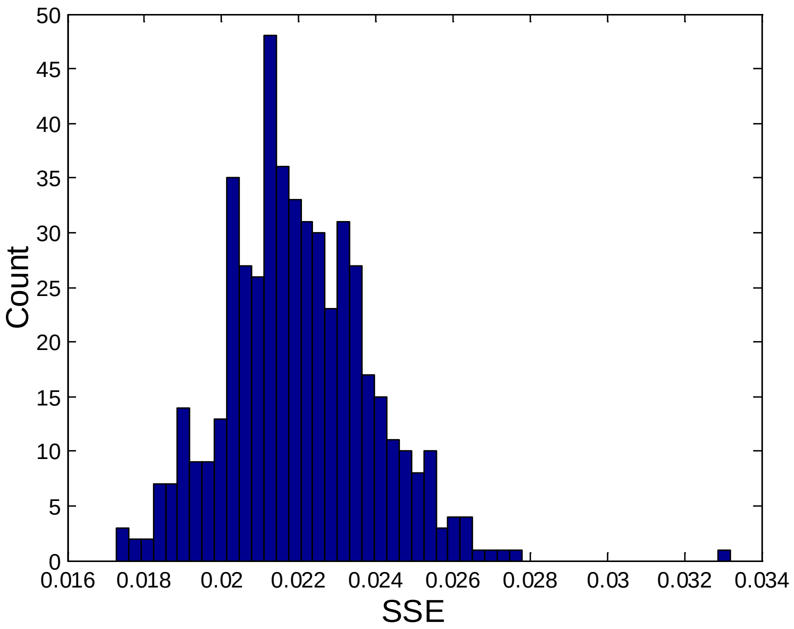
- Need a framework to interpret any measure
  - For example, if our measure of evaluation has the value, 10, is that good, fair, or poor?
- Statistics provide a framework for cluster validity
  - The more “atypical” a clustering result is, the more likely it represents valid structure in the data

Statistical Framework for SSE

- **Example**
  - Compare SSE of 0.005 against three clusters in random data
  - SSE Histogram of 500 sets of random data points of size 100—lowest SSE is 0.0173

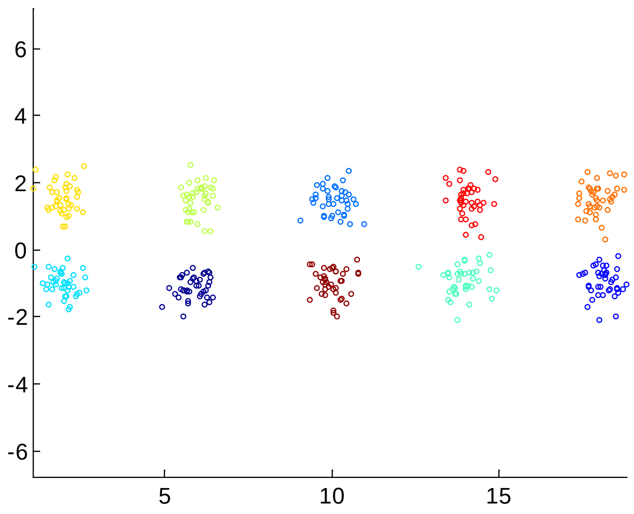


SSE = 0.005

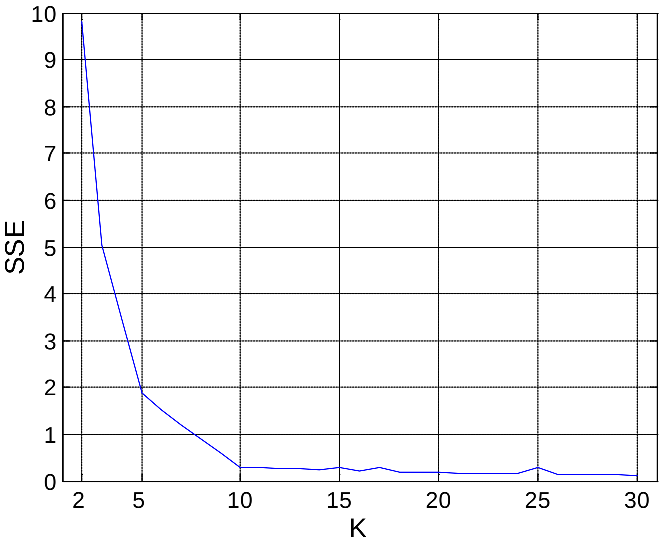


Determine the Number of Clusters Using SSE

- SSE curve



Clustering of Input Data



SSE wrt K

Take-away Message

- What’s clustering?
- Why clustering is important?
- How to preprocess data and compute dissimilarity/similarity from data?
- What’s a good clustering solution?
- How to evaluate the clustering results?