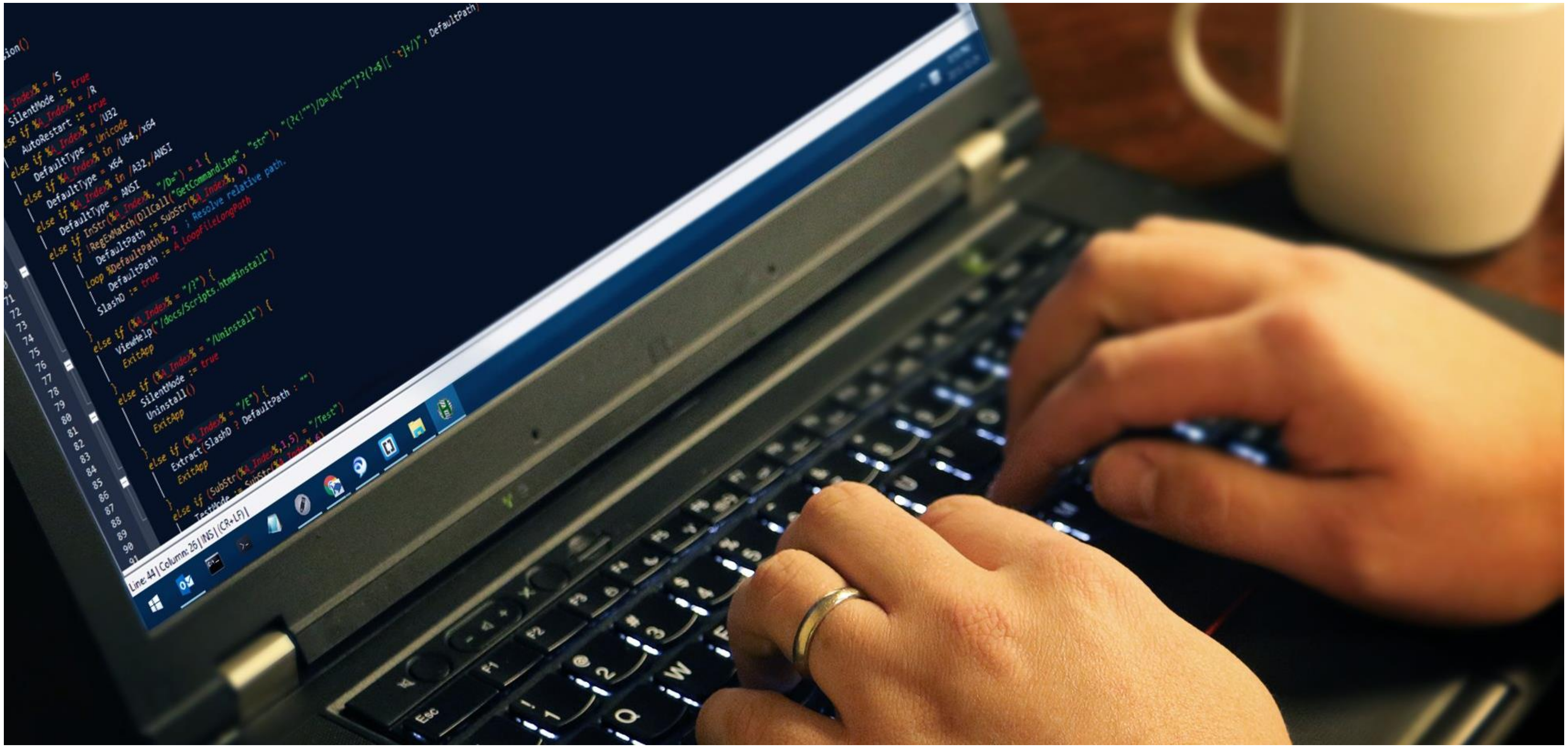


AI-Data Mining

Cluster & association analysis by python

DataMining-*Cluster Analysis*

Statistical method of partitioning a sample into homogeneous classes.

Purpose

1. Sort observations into groups (or clusters) such that the degree of association is:

- Strong between members of the *same cluster*
- Weak between members of *different clusters*

Example From Practice Exercise Week 10

4. Iris Data Set: This database widely used for pattern recognition literature. The data set include 5 columns:

i. sepal length in cm

ii. sepal width in cm

iii. petal length in cm

iv. petal width in cm

v. class:

-- Iris Setosa

-- Iris Versicolour

-- Iris Virginica

DataMining - *About Fisher's Iris data set*

Relevant Information:

------This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. (See Duba & hart, for example.) The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant.

5.1,3.8,1.9,0.4,Iris-setosa
4.8,3.0,1.4,0.3,Iris-setosa
5.1,3.8,1.6,0.2,Iris-setosa
4.6,3.2,1.4,0.2,Iris-setosa
5.3,3.7,1.5,0.2,Iris-setosa
5.0,3.3,1.4,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
5.5,2.3,4.0,1.3,Iris-versicolor
6.0,2.5,Iris-virginica
5.1,1.9,Iris-virginica
5.9,2.1,Iris-virginica
5.6,1.8,Iris-virginica
5.8,2.2,Iris-virginica

Iris flower data set
Anderson's Iris data set
Fisher's iris data set

Iris setosa

Iris versicolor

Iris virginica

DataMining - *About Fisher's Iris data set*

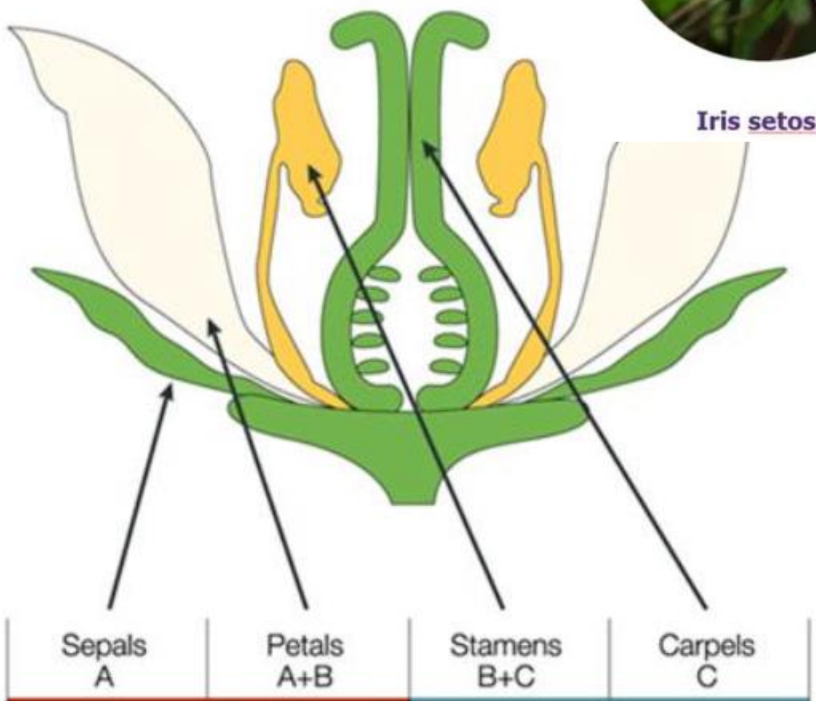
Attribute Information: ***Sepal length***, ***Sepal width***, ***Petal length***, ***Petal width***. (cm)

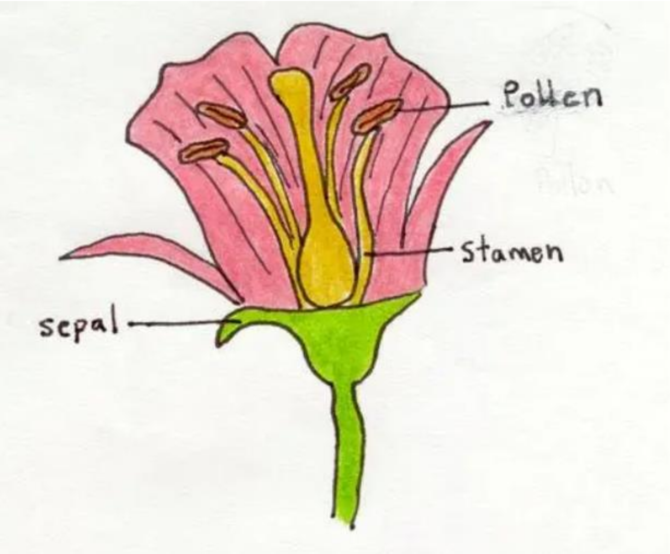
class: ***Iris Setosa***, ***Iris Versicolour***, ***Iris Virginica***.


Iris setosa


Iris versicolor

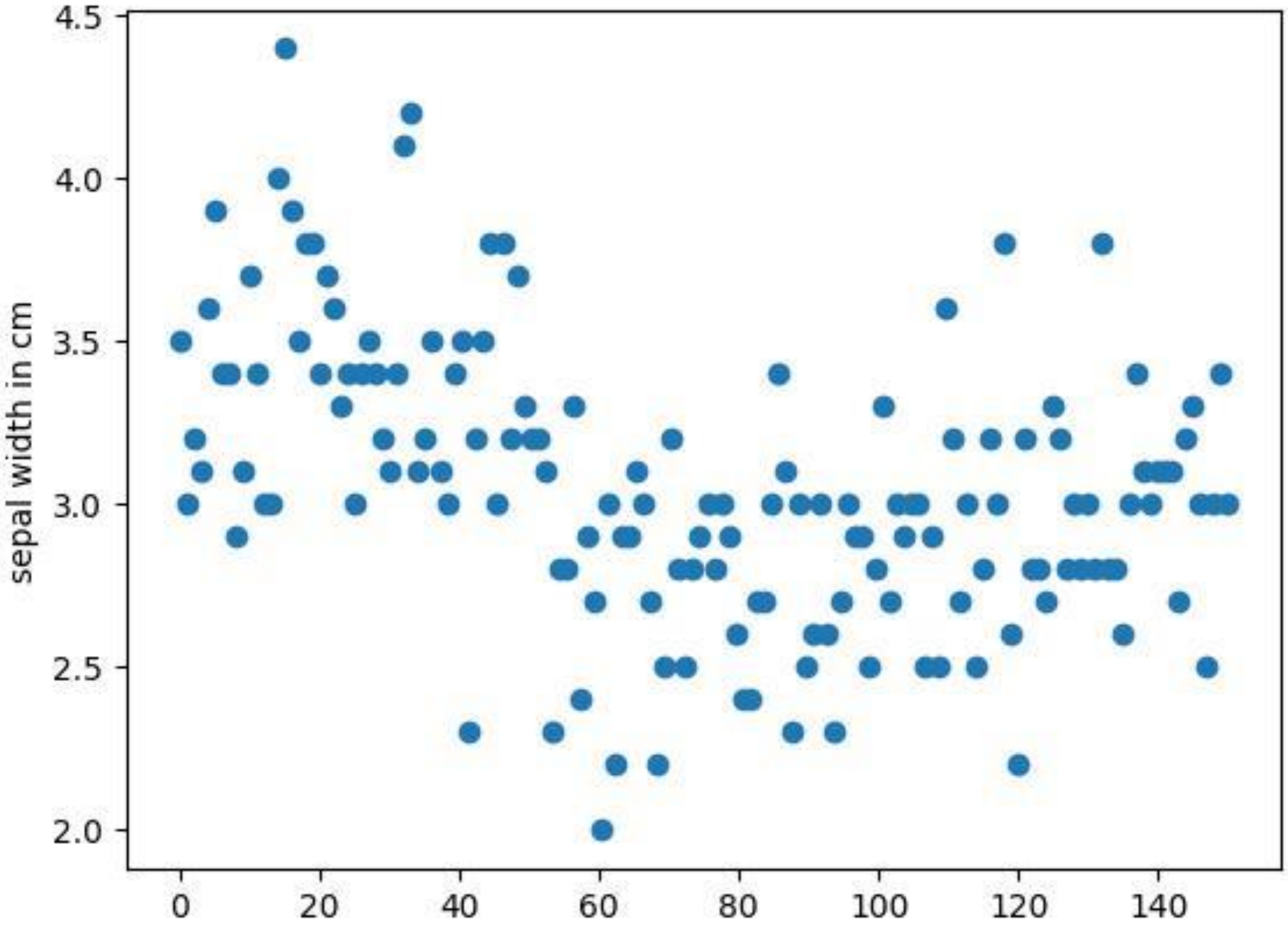

Iris virginica





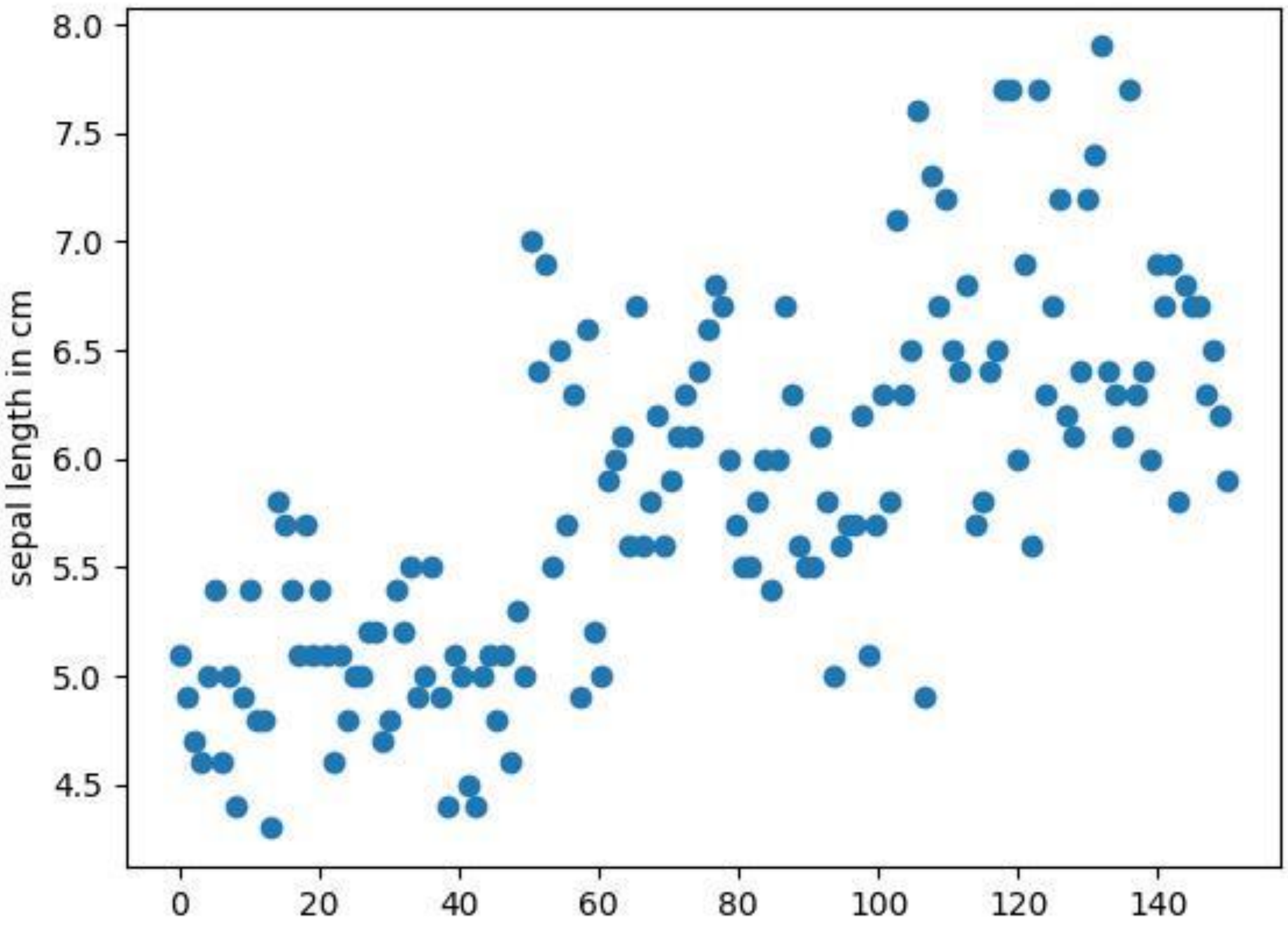
DataMining-*Cluster Analysis*

Example from Practice Exercise week 10



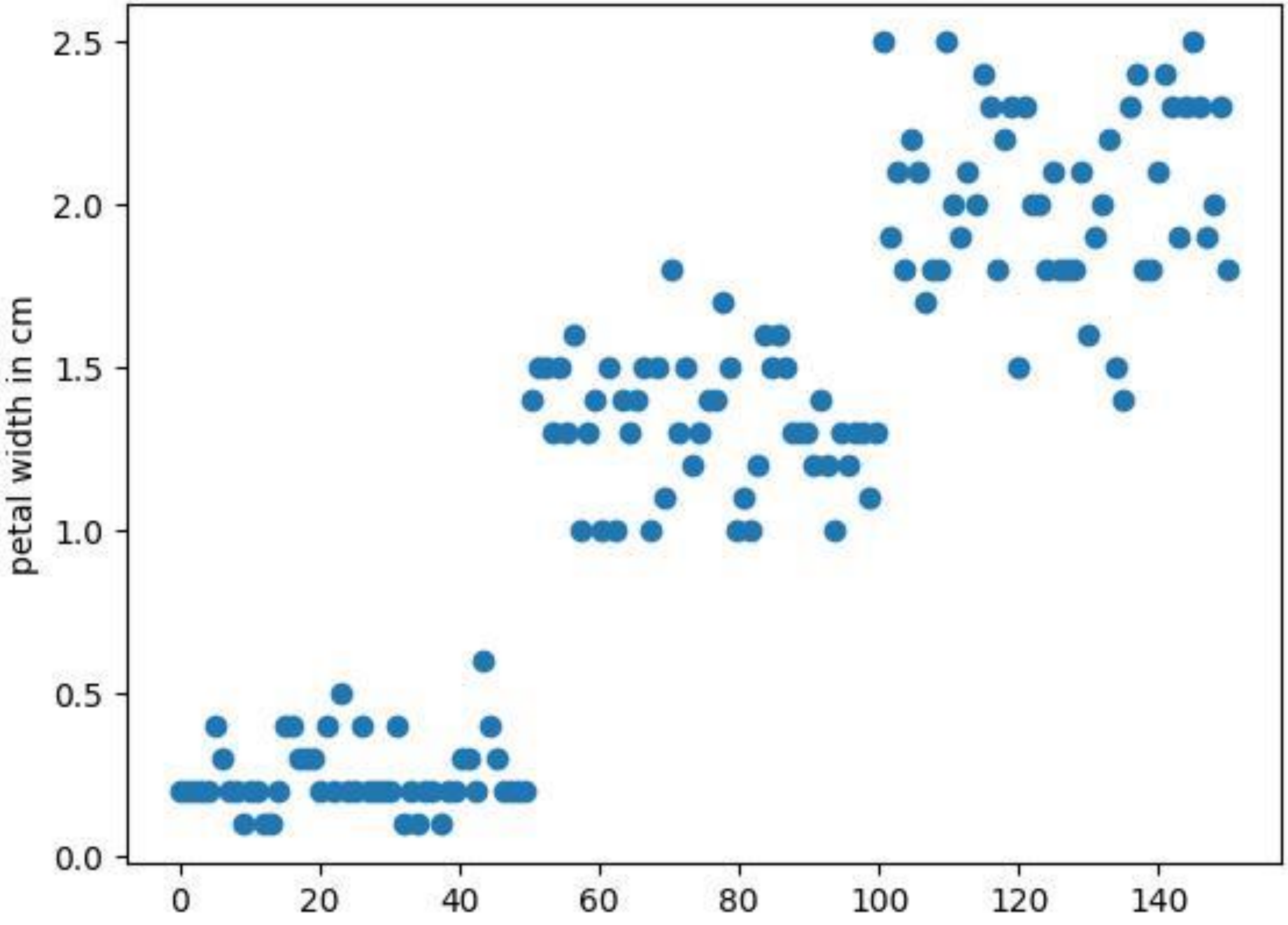
DataMining-*Cluster Analysis*

Example from Practice Exercise week 10

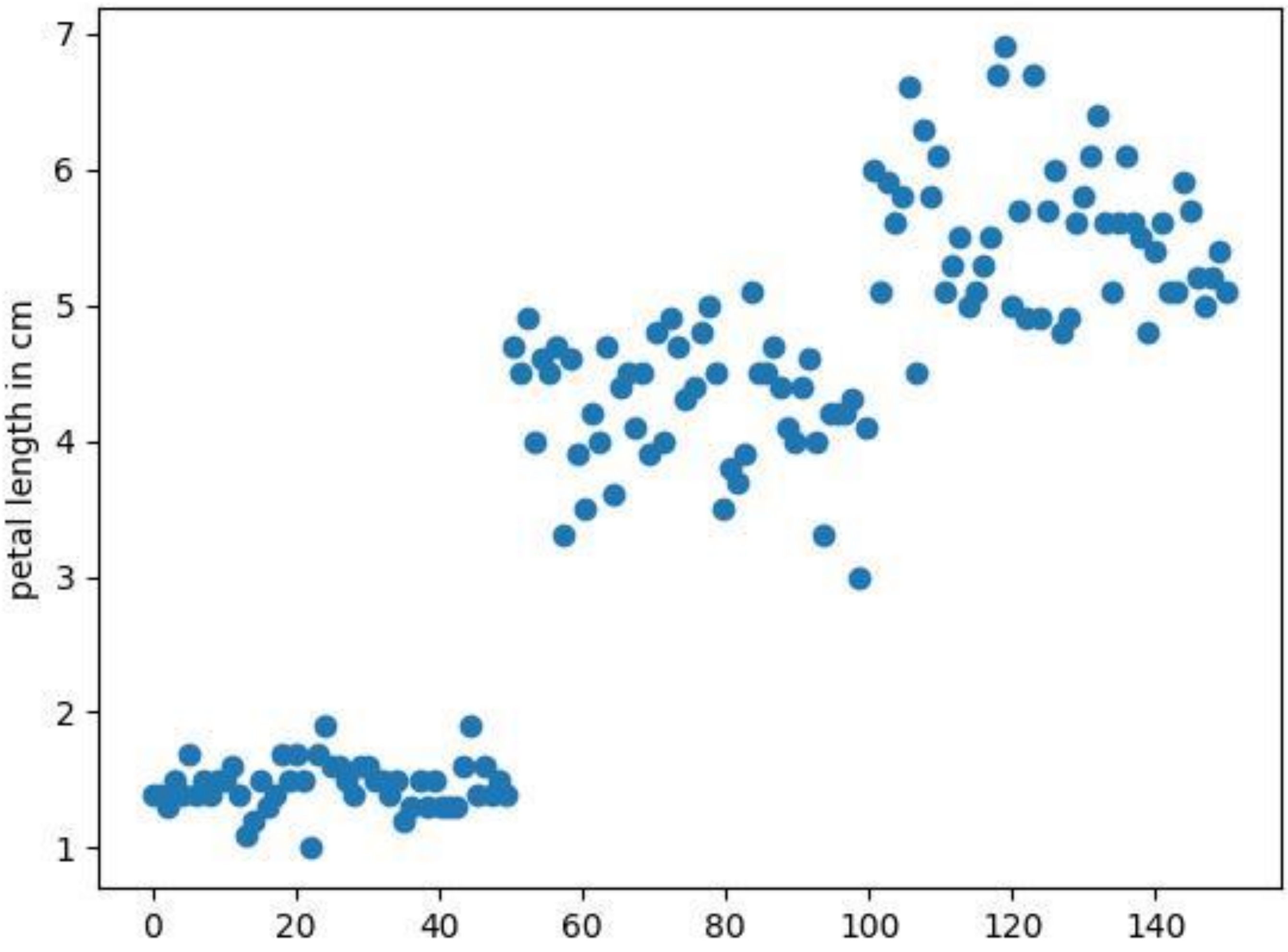


DataMining-*Cluster Analysis*

Example from Practice Exercise week 10

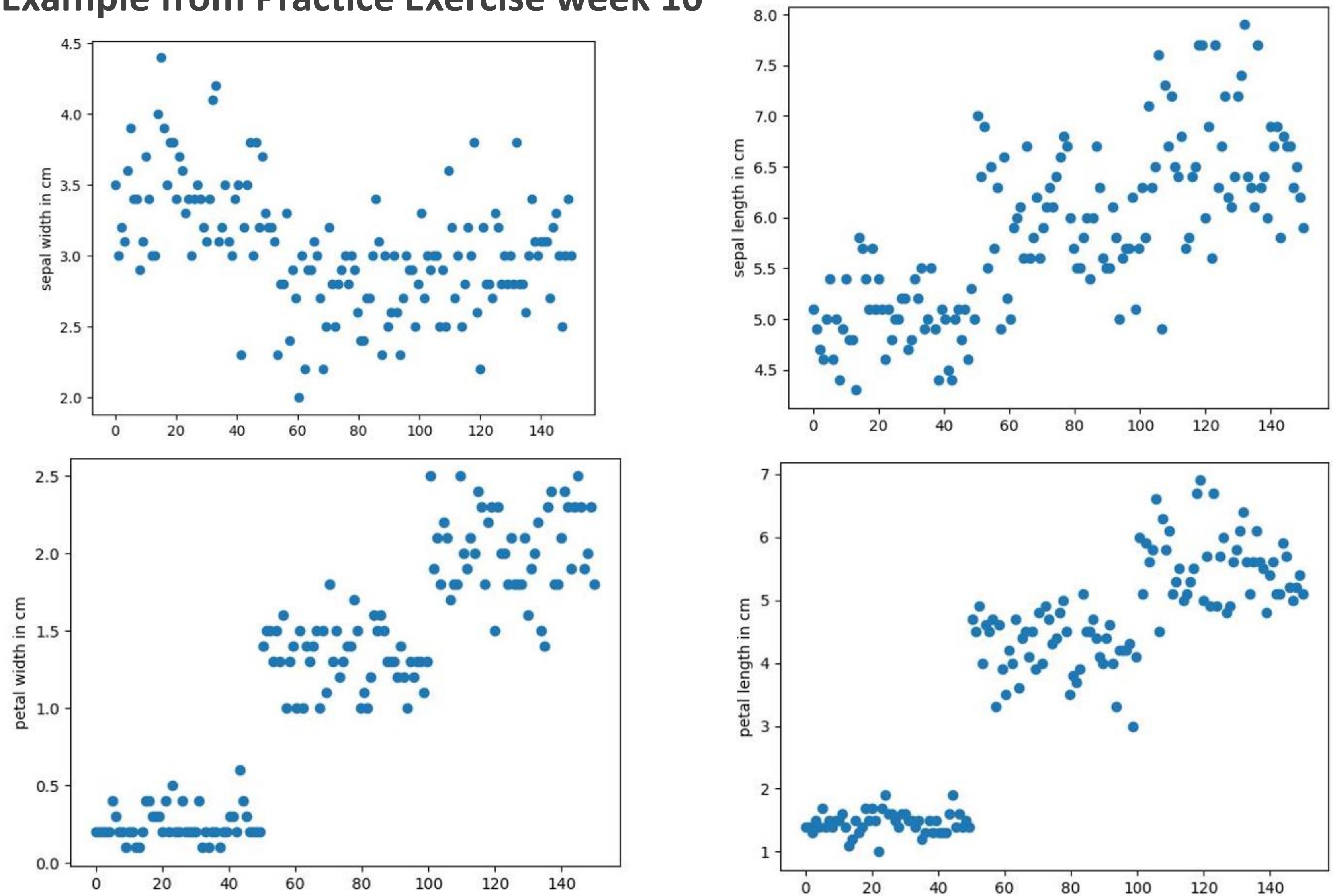


Example from Practice Exercise week 10



DataMining-*Cluster Analysis*

Example from Practice Exercise week 10



DataMining-*Cluster Analysis*

Statistical method of partitioning a sample into homogeneous classes.

Purpose

1. Sort observations into groups (or clusters) such that the degree of association is:

- Strong between members of the *same cluster*
- Weak between members of *different clusters*

2. Define a formal classification scheme that was not previously evident

Supervised vs unsupervised learning

1. **Supervised**

Can train your model and use it for “new” data with some accuracy

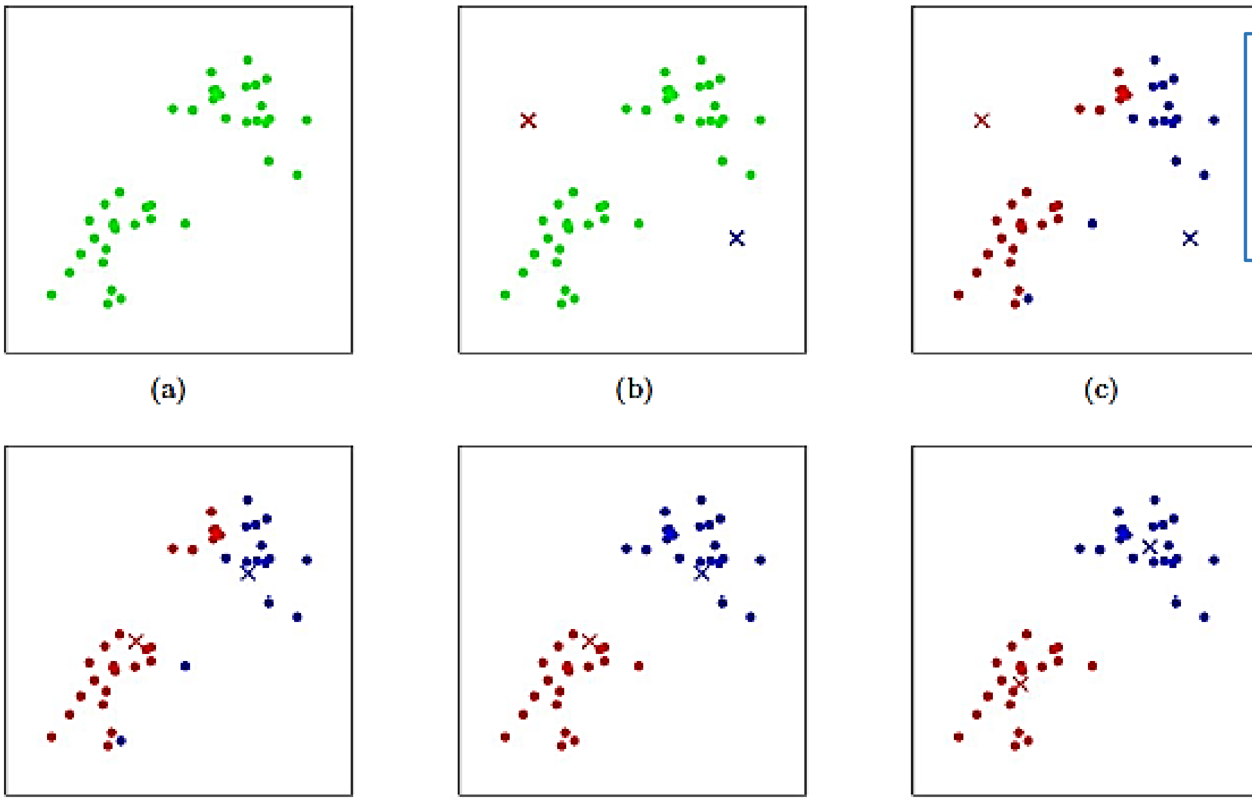
- Initial model: Use a portion of the data to “train” your data and ”test” using the remaining portion
- *e.g.*, Linear and logistic regression, classification.

2. **Unsupervised**

- Does not use output data for further learning
- *e.g.*, Cluster analysis

K-means cluster

- Randomly assign k centroids
- Assign all data points to their closest centroids
- Update centroid assignments
- Repeat the previous two steps until centroids are stable



(a) (b) (c) (d) (e) (f)

Euclidean distance:

$$d_{euclidean}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Manhattan:

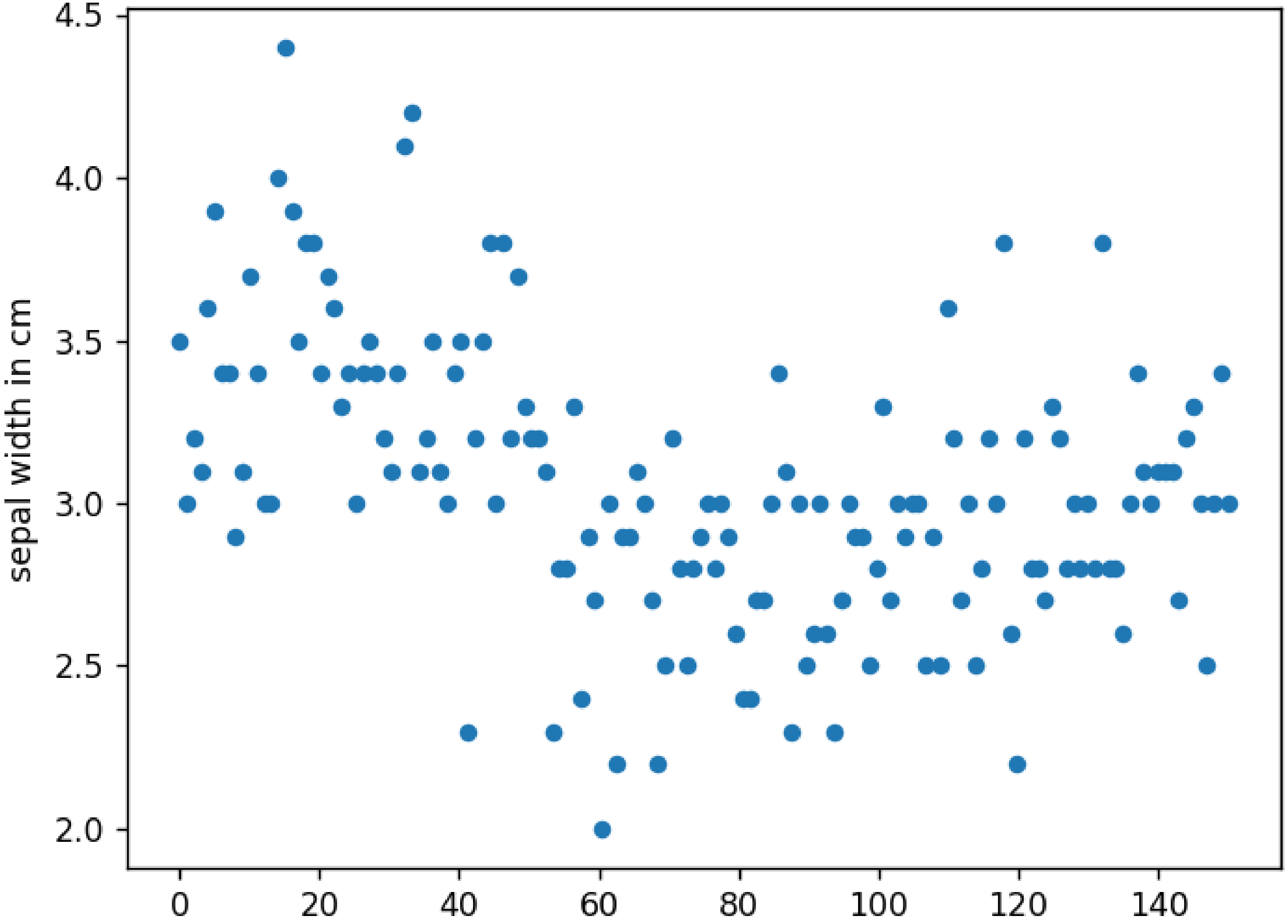
$$d_{manhattan} = \sum_{i=1}^{n} |(x_i - y_i)|$$

Minkowski:

$$d_{minkowski} = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$$

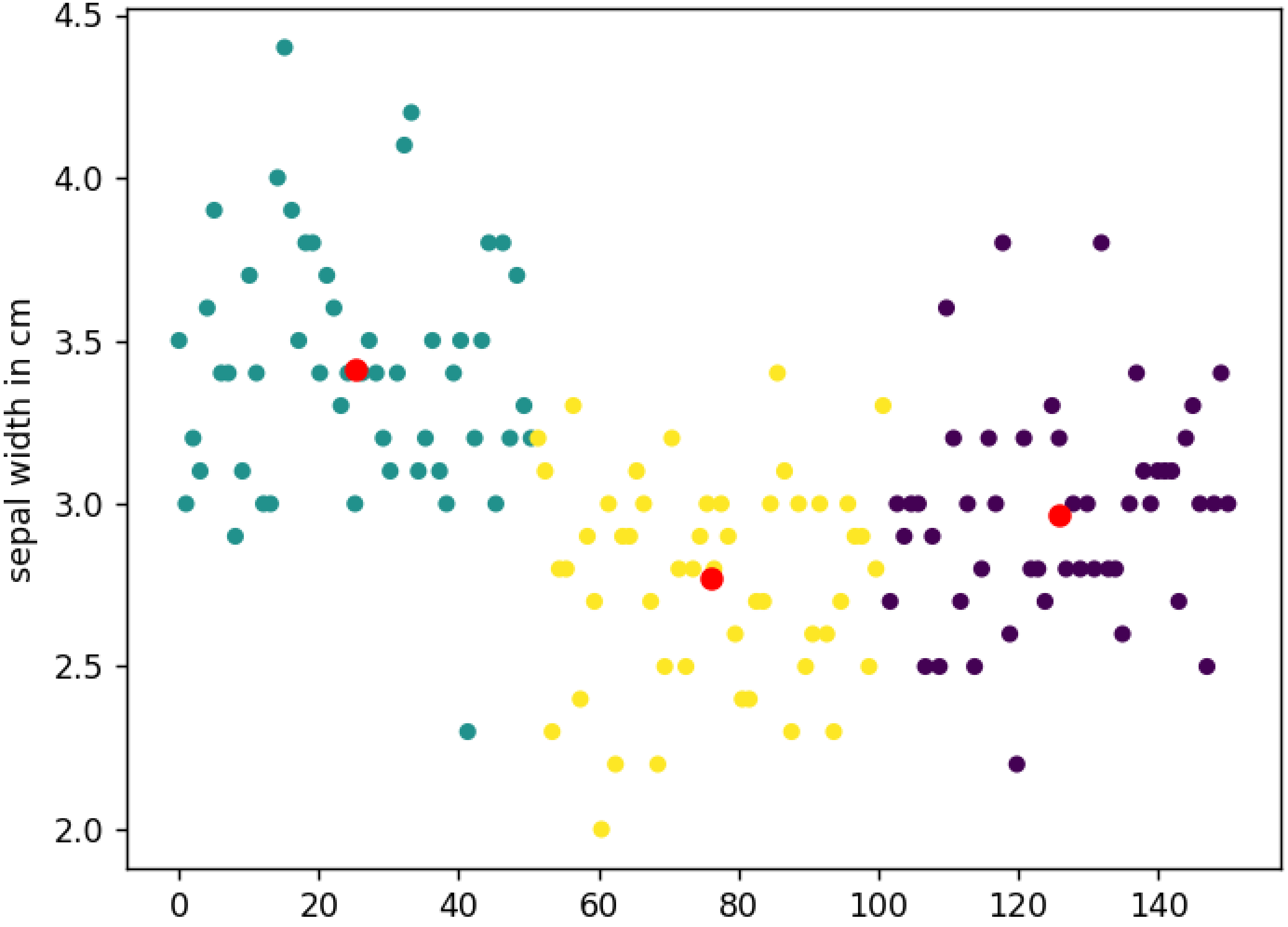
Cluster K-means *Iris data set*

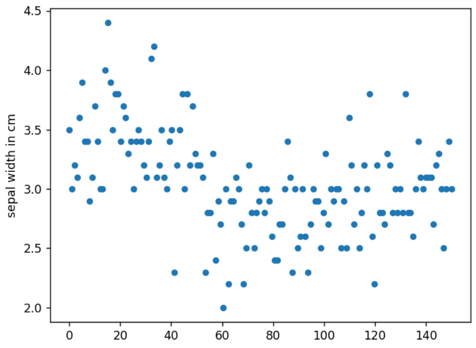
samples=150



Cluster K-means *Iris data set*

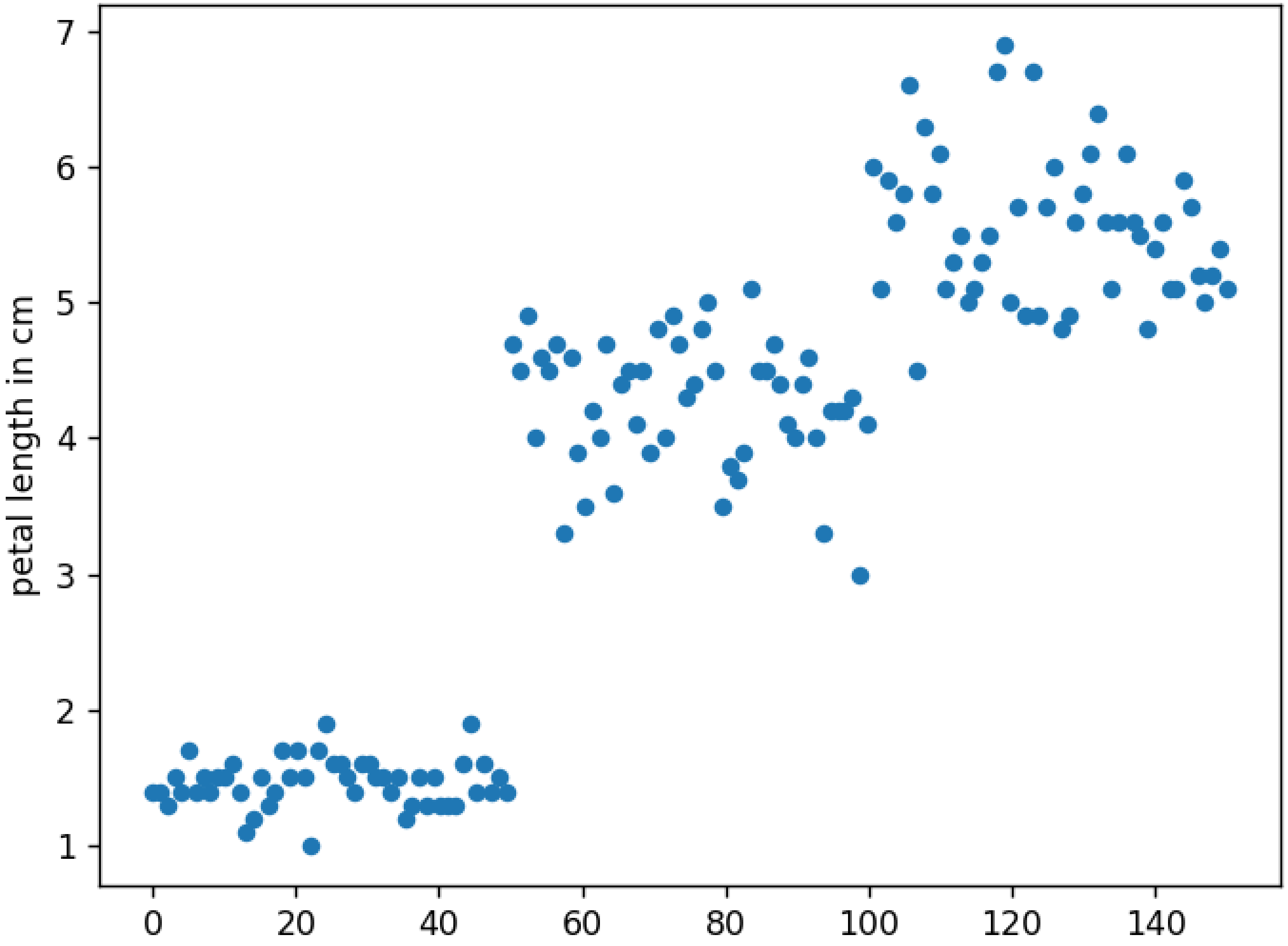
samples=150





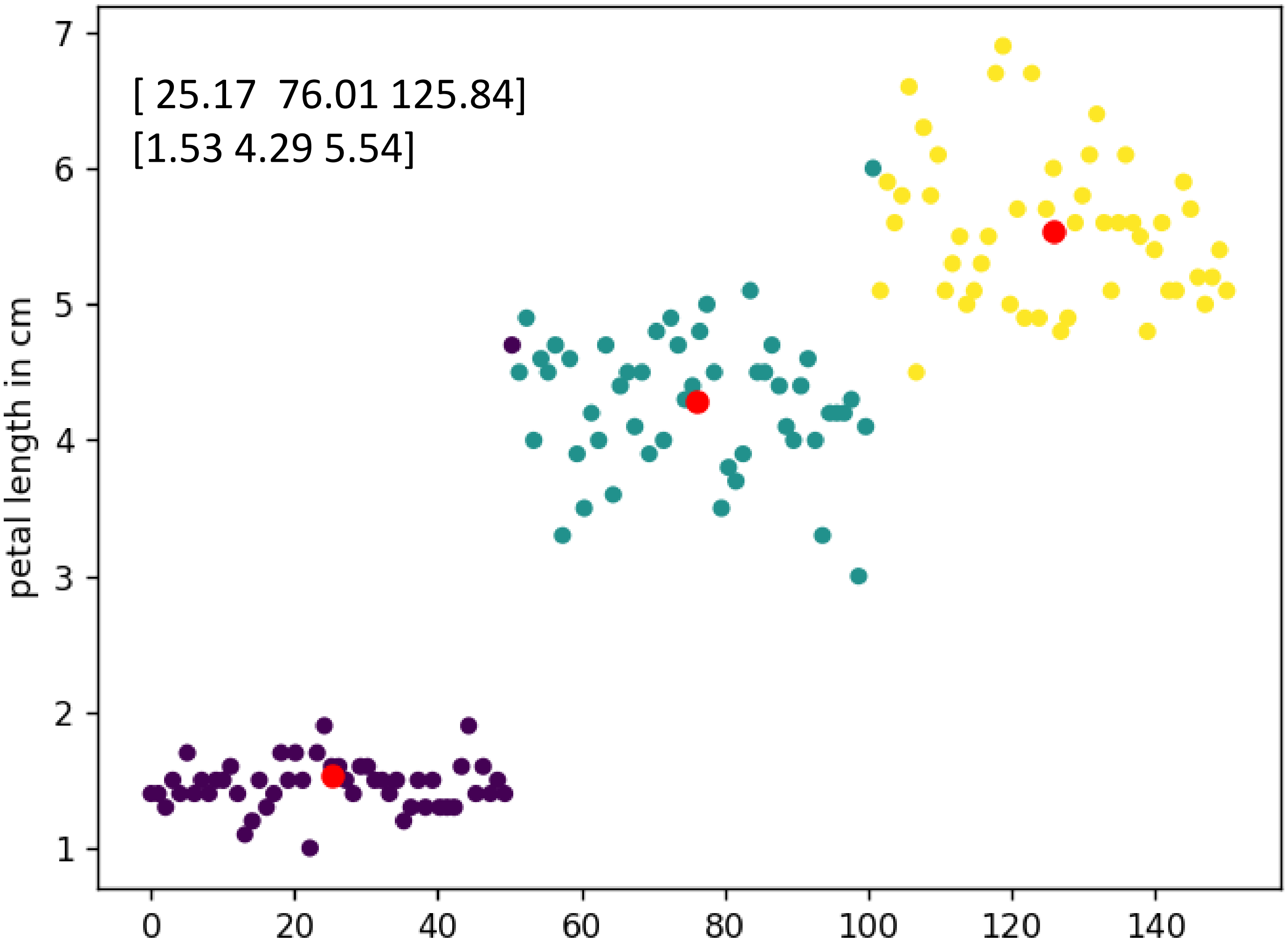
Cluster K-means *Iris data set*

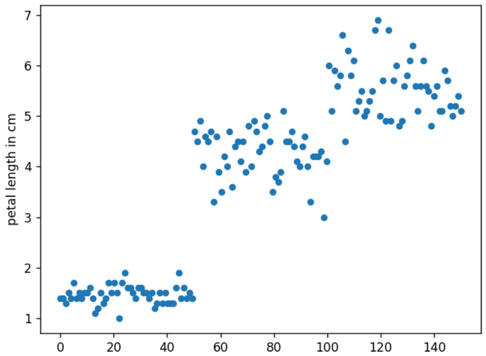
samples=150



Cluster K-means *Iris data set*

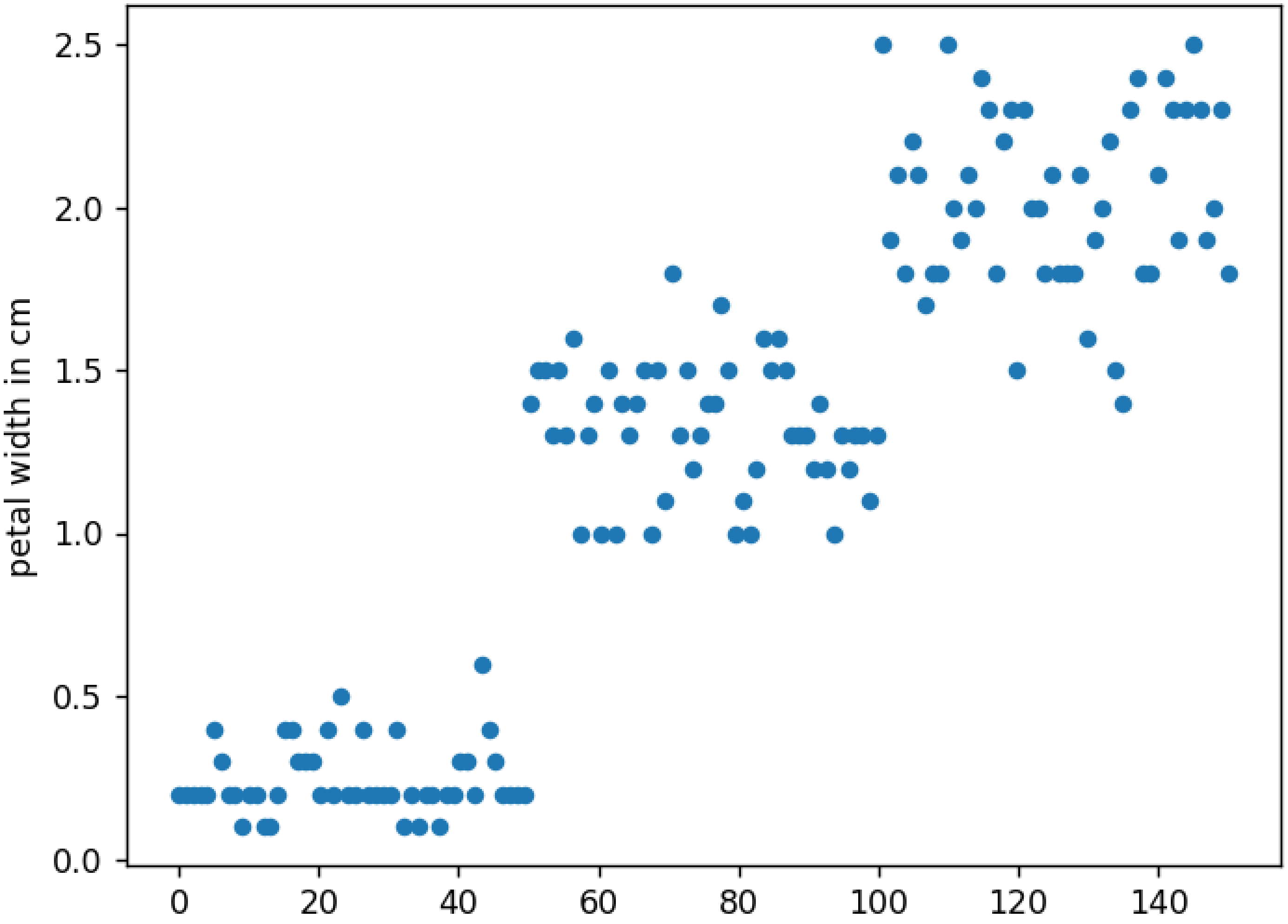
samples=150





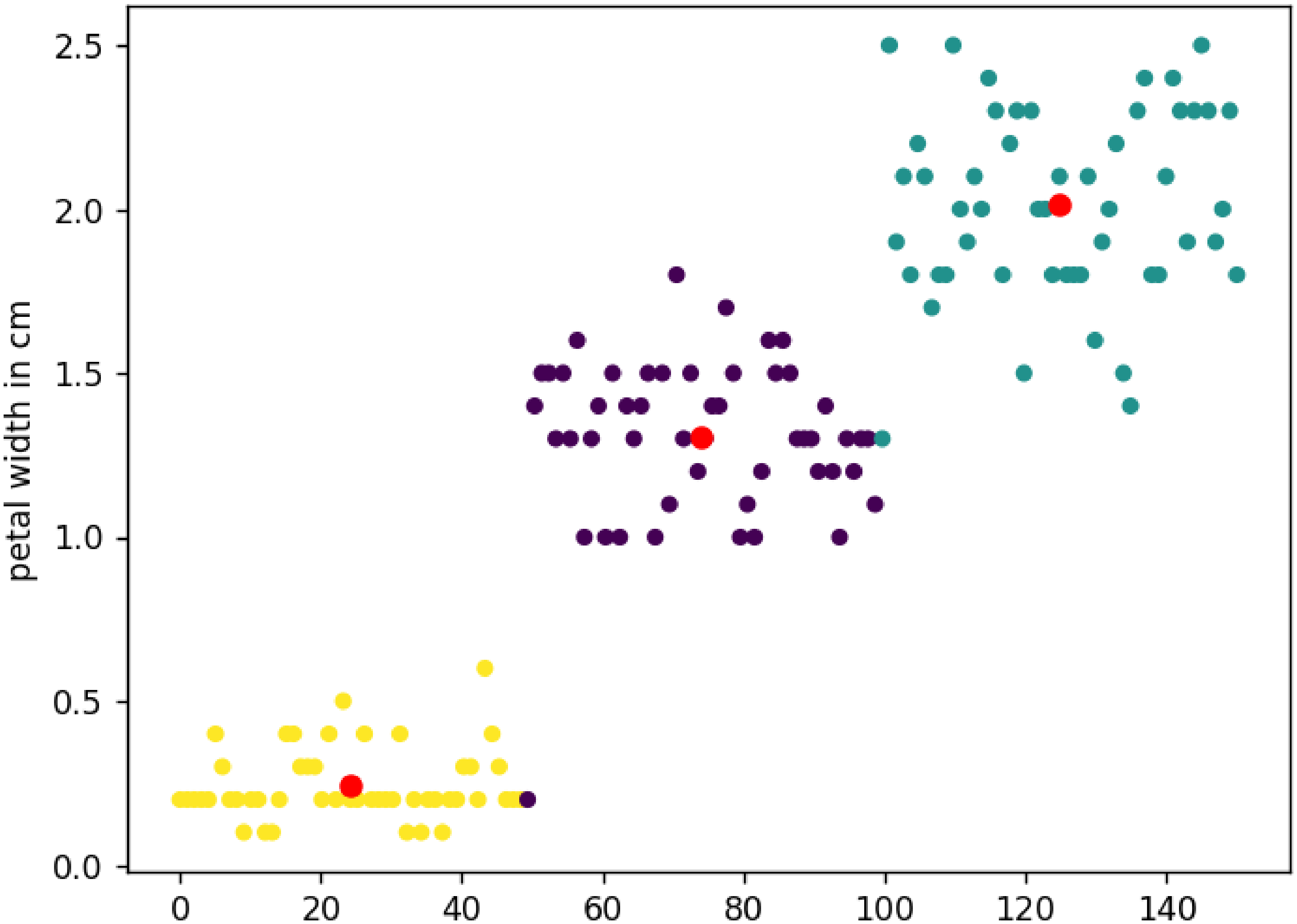
Cluster K-means *Iris data set*

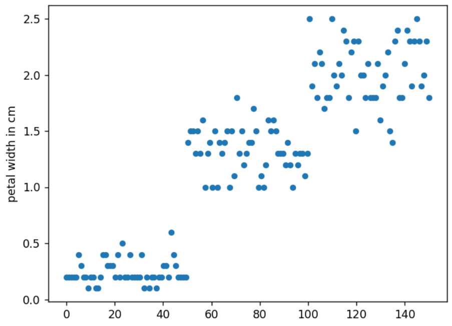
samples=150



Cluster K-means *Iris data set*

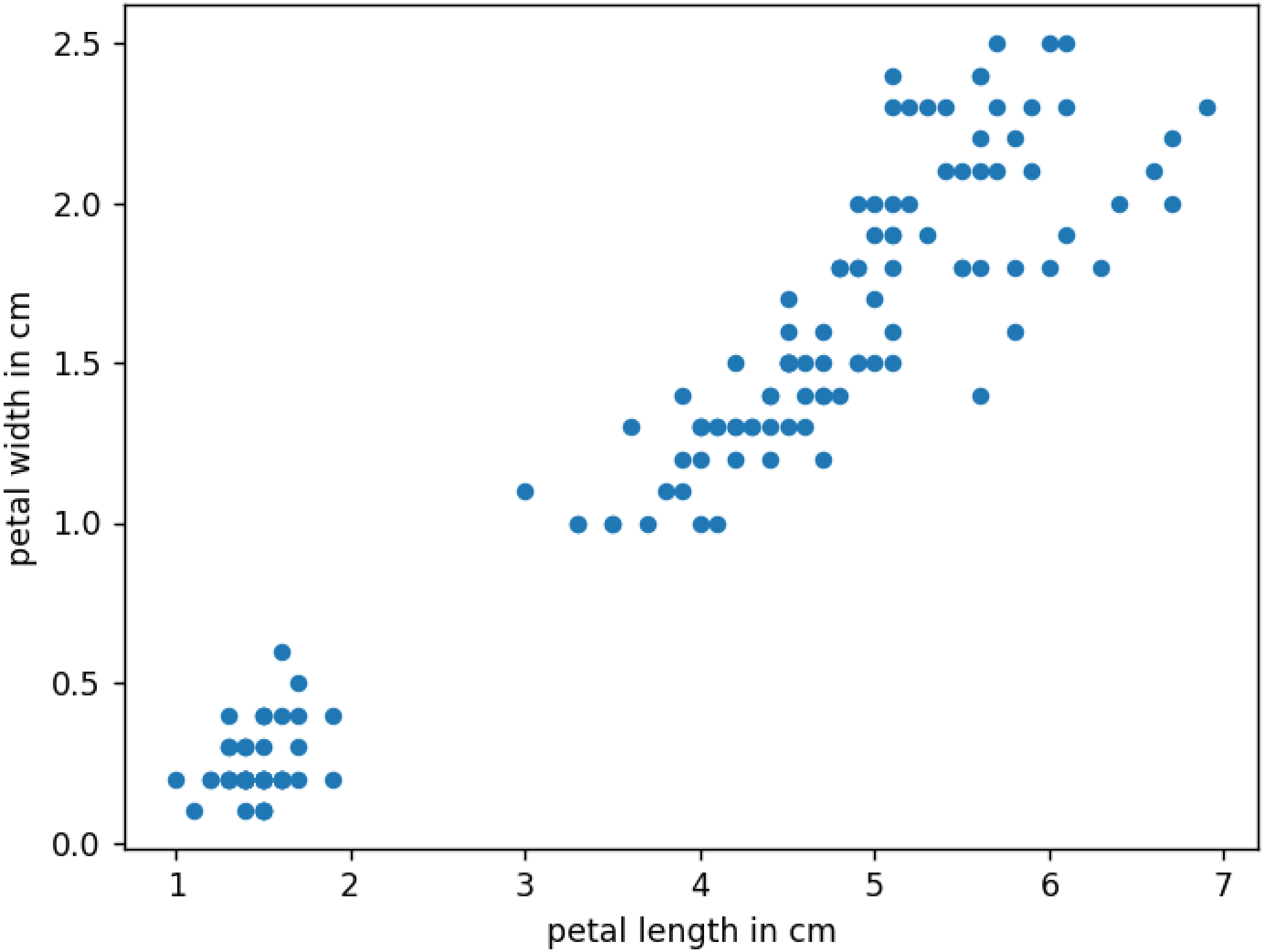
samples=150





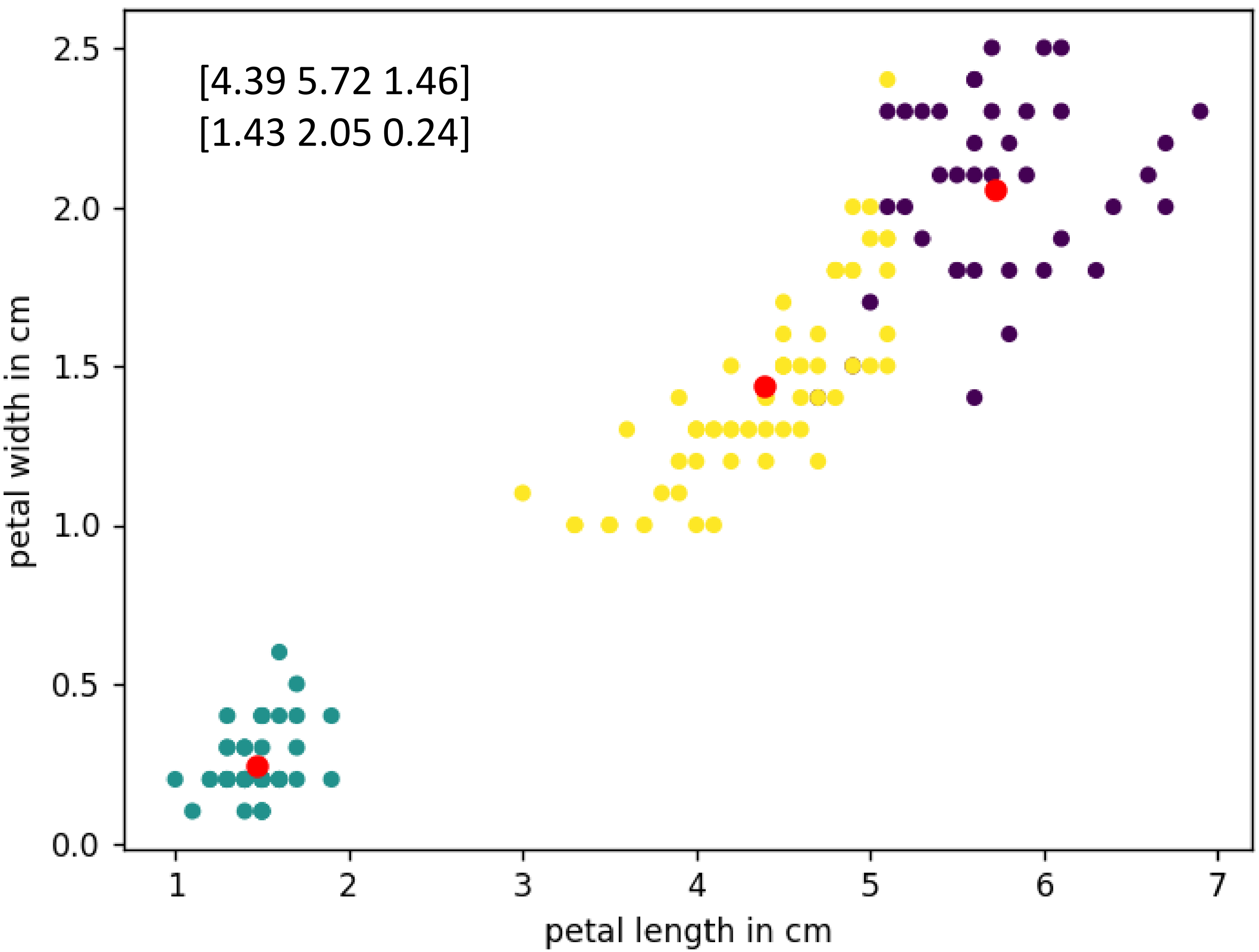
Cluster K-means *Iris data set*

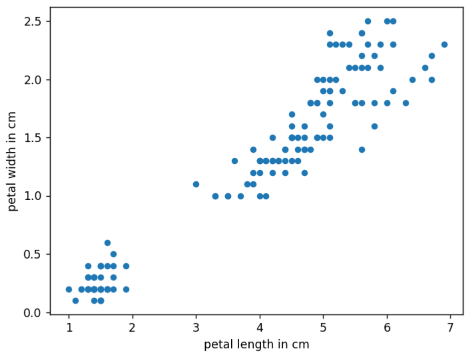
samples=150



Cluster K-means *Iris data set*

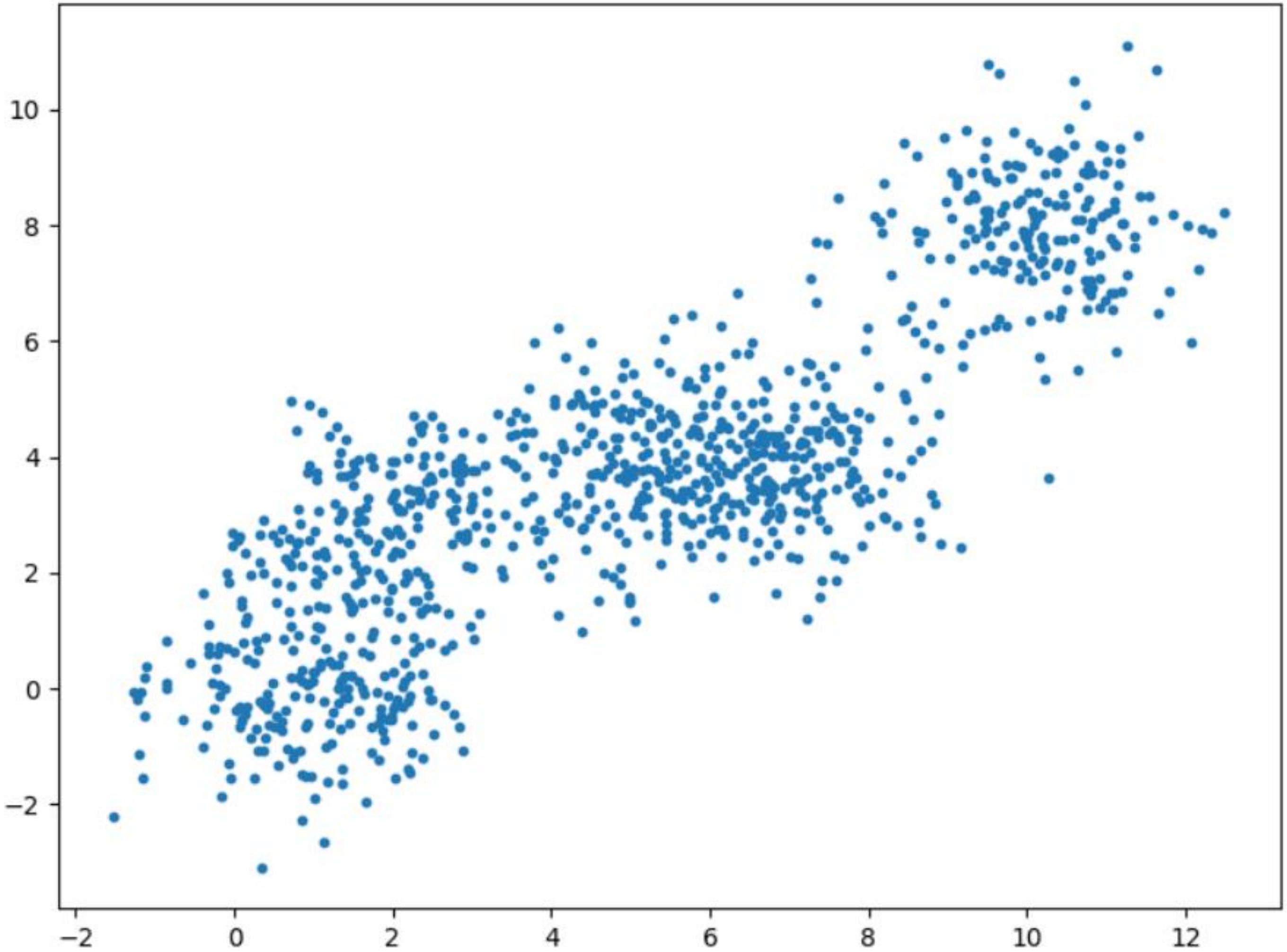
samples=150





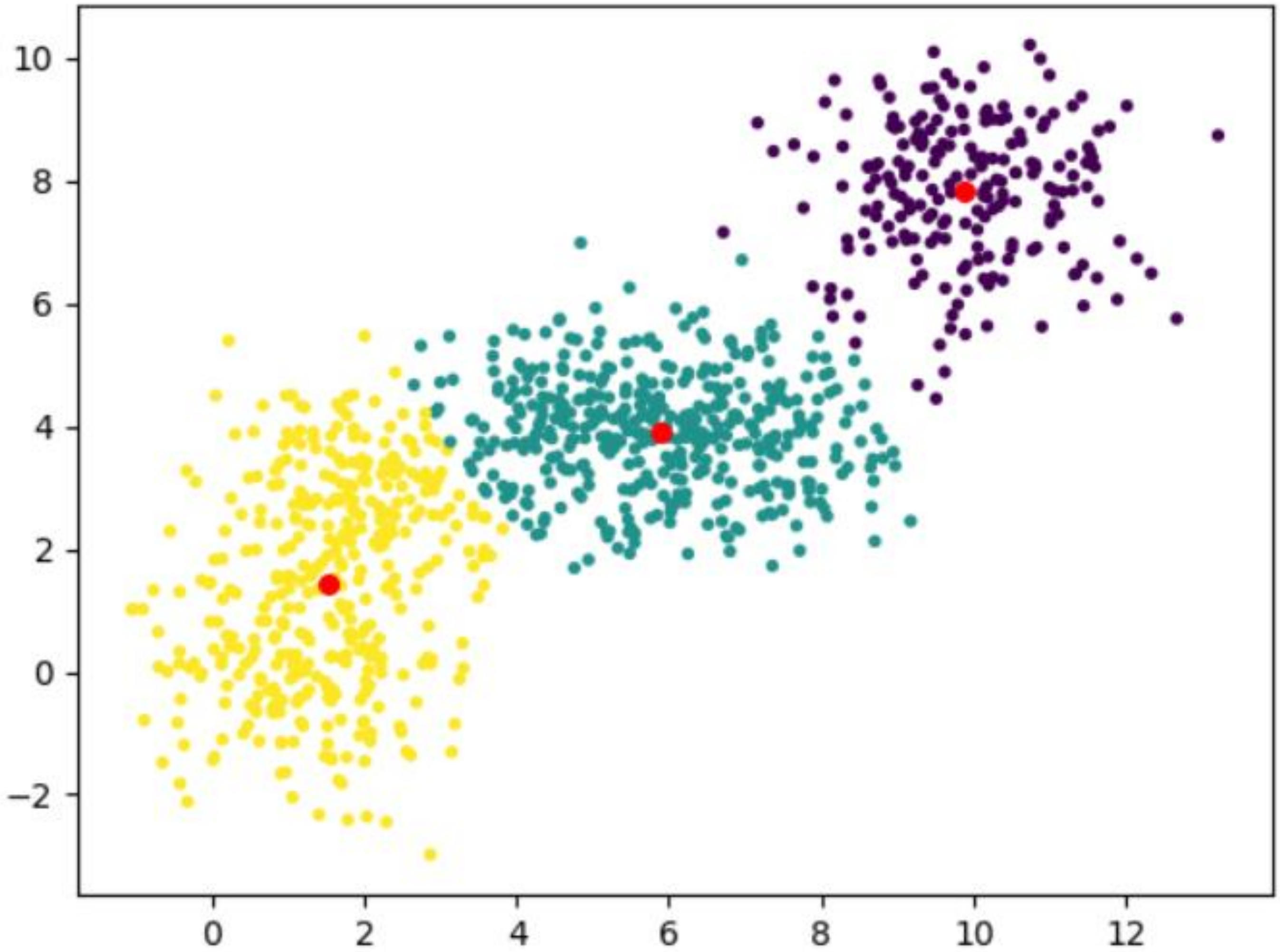
Cluster K-means example_2

samples=1000



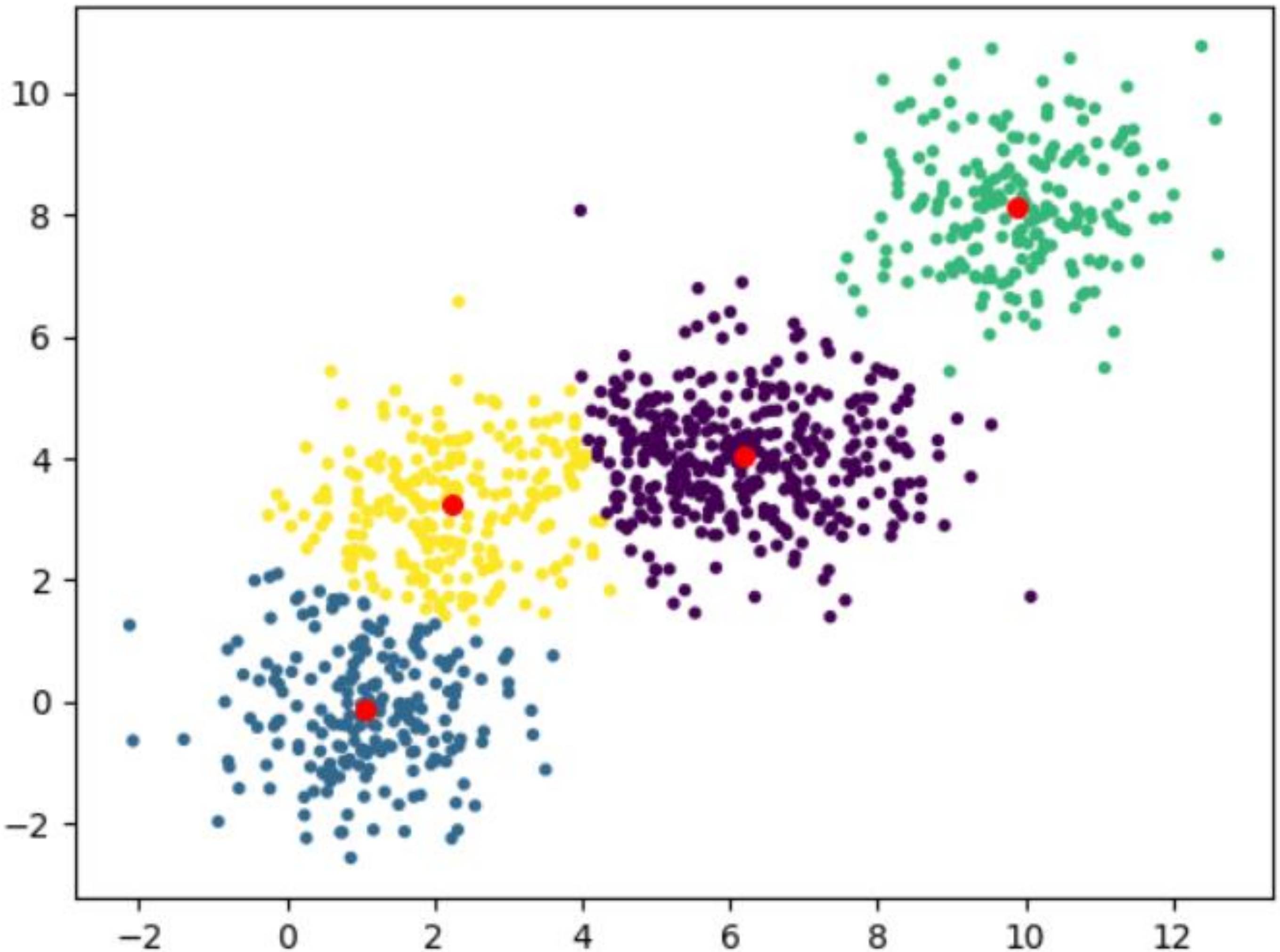
Cluster K-means example

- K=3



Cluster K-means example

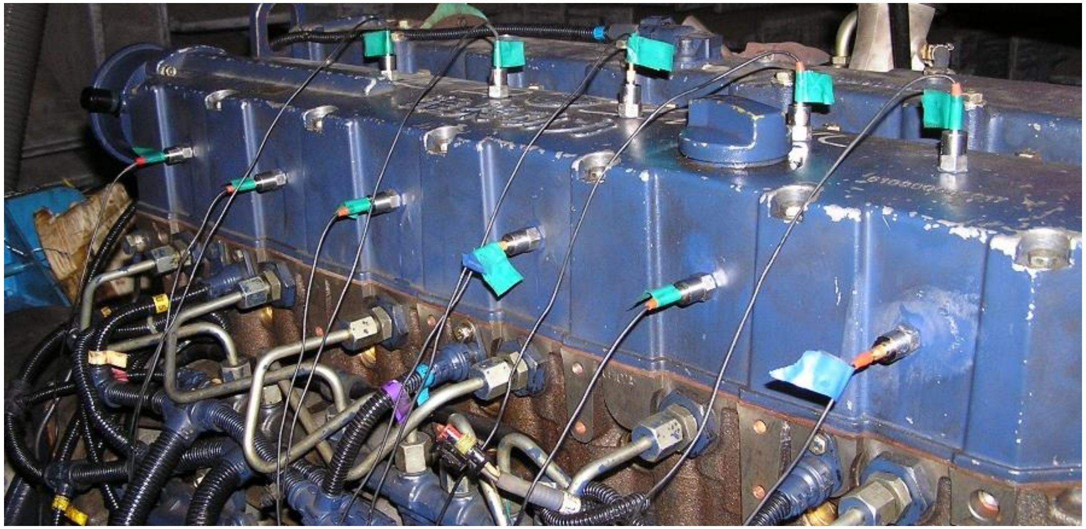
- K=4



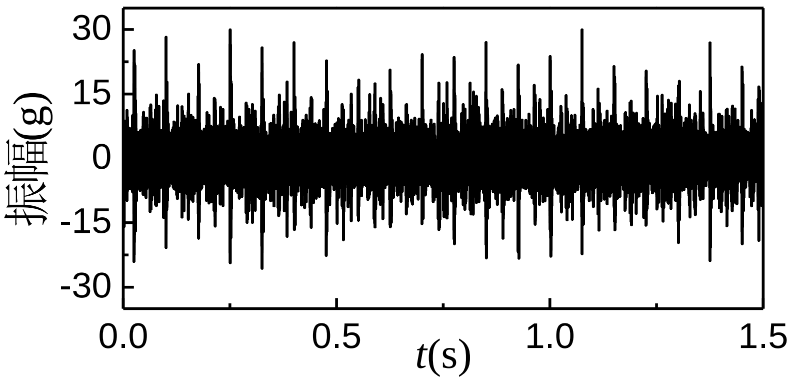
Cluster application in my research

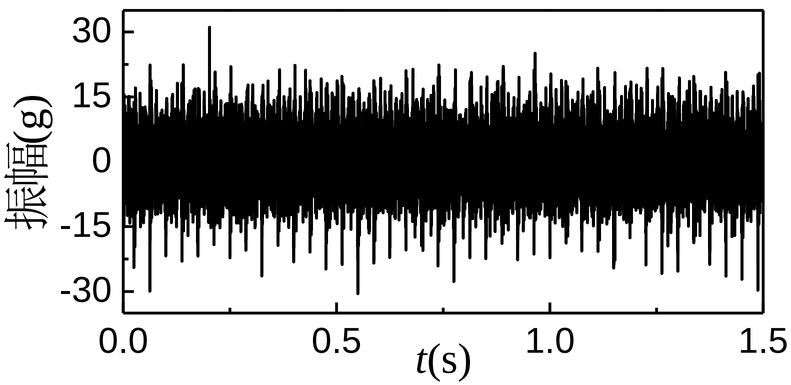
Fault Diagnose for Disel Engine Base on Vibration Signal






Cluster application in my research



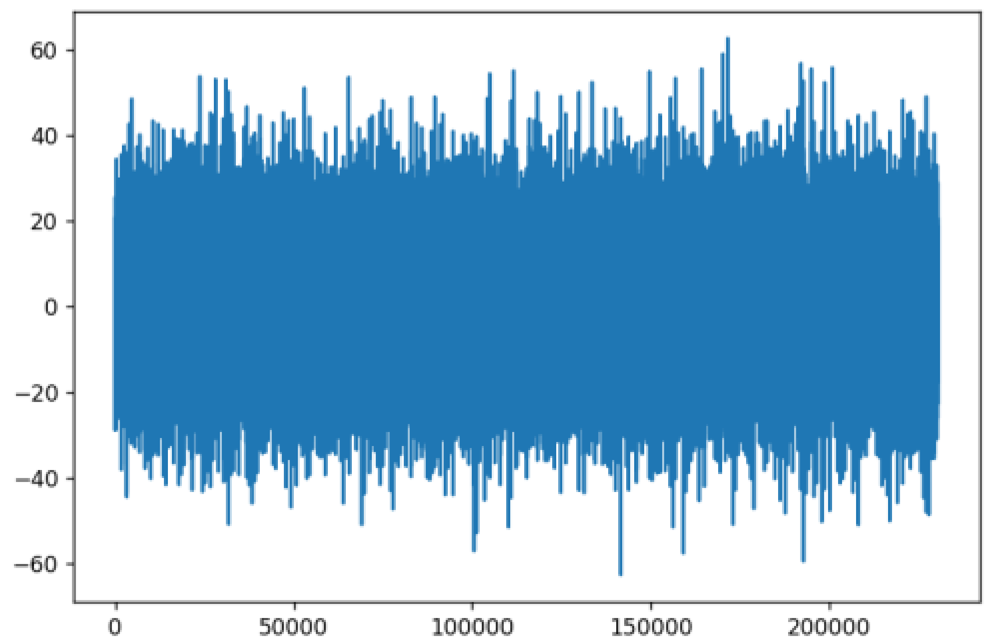


sampling rate:： 25.6kHz

25600/second

3. The "**VibrationData.csv**" contains vibration acceleration signals in three directions (XYZ) at a certain position on the diesel engine, shown as follows:

	A	B	C	D	E	F
1	Time	DirectionX	Time	DirectionY	Time	DirectionZ
2	1.74E-05	-2.464599609	1.74E-05	6.405395508	1.74E-05	-2.416381836
3	5.64E-05	-2.174804688	5.64E-05	-13.42712402	5.64E-05	11.70776367
4	9.55E-05	-3.50402832	9.55E-05	-10.92236328	9.55E-05	-18.2668457
5	0.000134543	3.293945313	0.000134543	5.922241211	0.000134543	-7.789916992
6	0.000173606	12.06115723	0.000173606	-0.742797852	0.000173606	18.25866699
7	0.000212668	17.54418945	0.000212668	5.217407227	0.000212668	3.065185547
8	0.000251731	9.780883789	0.000251731	24.3605957	0.000251731	-15.39233398
9	0.000290793	-13.80981445	0.000290793	10.234375	0.000290793	22.29553223

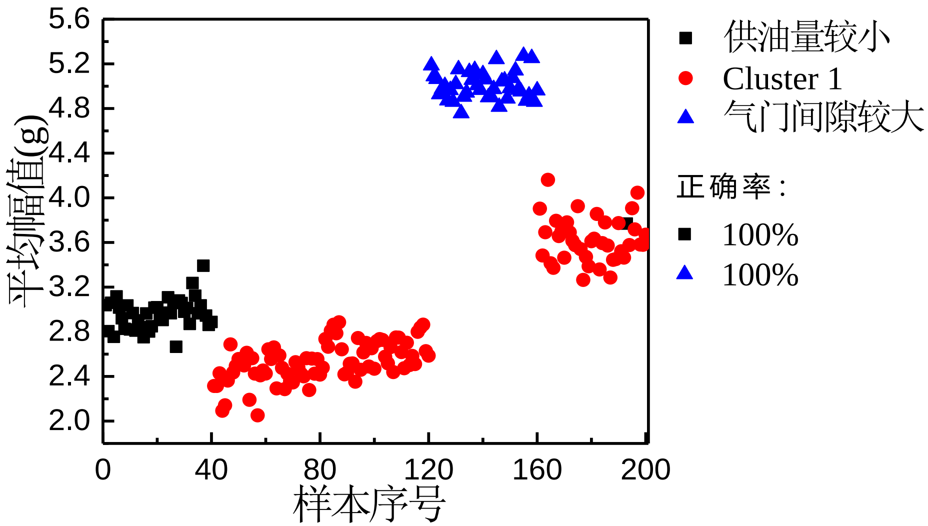


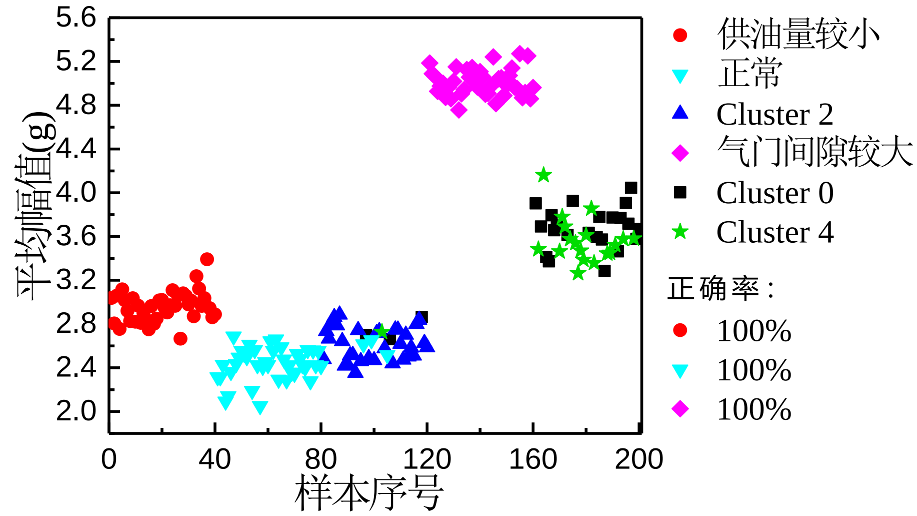
Cluster application in my research

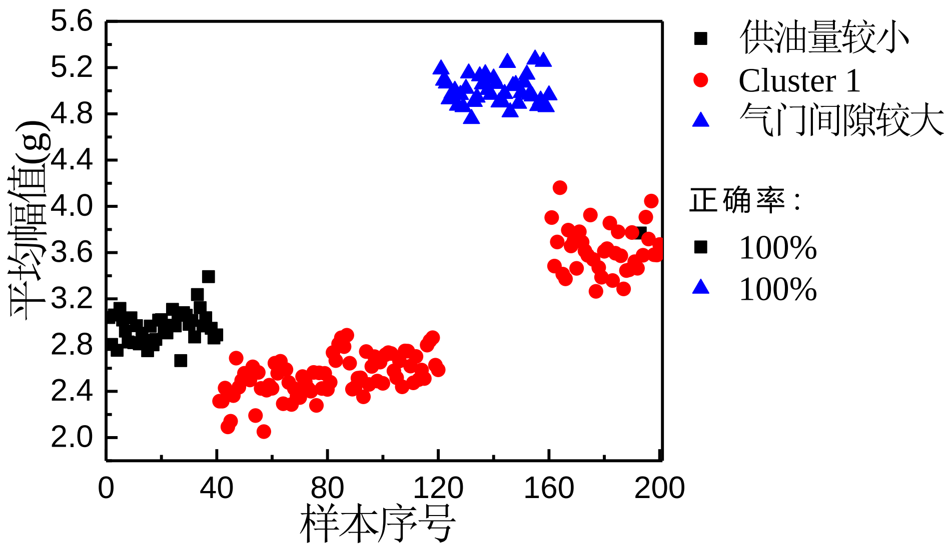
Extract Features from Vibration Signal

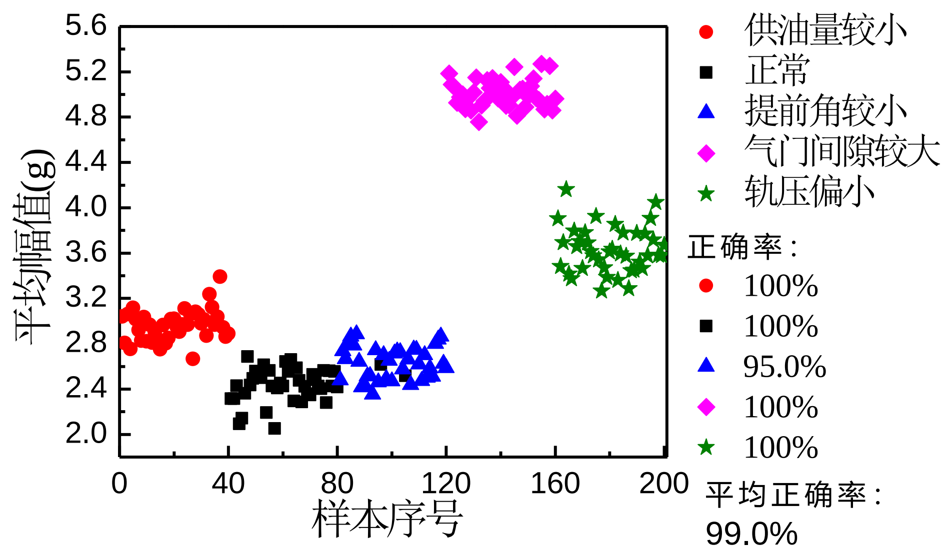
(1) $x_{\max} = \max(x_i)$	(2) $x_{\min} = \min(x_i)$	(3) $x_{pp} = x_{\max} - x_{\min}$
(4) $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$	(5) $\lvert\bar{x}\rvert = \frac{1}{N}\sum_{i=1}^{N} \lvert x_i \rvert$	(6) $\psi_x{}^2 = \frac{1}{N}\sum_{i=1}^{N} x_i{}^2$
(7) $\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$	(8) $\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$	(9) $C_f = \frac{x_p}{x_{rms}}$
(10) $C_e = \frac{x_p}{x_r}$	(11) $S_f = \frac{x_{rms}}{\lvert \bar{x} \rvert}$	(12) $K = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^4}{\sigma^4}$

Cluster application in my research









平均正确率：
99.0%

Data mining – Association Analysis

the "true story" about using data mining to identify a relation between sales of beer and diapers

Core:

So what are the facts? In 1992, Thomas Blischok, manager of a retail consulting group at Teradata, and his staff prepared an analysis of 1.2 million market baskets from about 25 Osco Drug stores. Database queries were developed to identify affinities. The analysis "did discover that between 5:00 and 7:00 p.m. that consumers bought beer and diapers". Osco managers did NOT exploit the beer and diapers relationship by moving the products closer together on the shelves. This decision support study was conducted using query tools to find an association. The true story is very bland compared to the legend ……

Original text: http://www.dssresources.com/newsletters/66.php

Data mining – Association Analysis

the "true story" about using data mining to identify a relation between sales of beer and diapers

Core:

So what are the facts? In 199
retail consulting group at Te
analysis of 1.2 million marke
stores. Database queries were
analysis "did discover that b
consumers bought beer and dia
the beer and diapers relation
together on the shelves. This
using query tools to find an
bland compared to the legend

Original text: http://www.dss

DSS News
D. J. Power, Editor
November 10, 2002 -- Vol. 3, No. 23
A Bi-Weekly Publication of DSSResources.COM

****
Check the article by F. Kelly "Implementing an EIS"
****

Featured:

* DSS Wisdom
* Ask Dan! - What is the "true story" about data mining, beer and diapers?
* What's New at DSSResources.COM
* DSS News Releases

****

Enhance model-driven DSS with Crystal Ball simulation software.
Download a FREE evaluation at http://www.crystalball.com/dss/

****

DSS Wisdom

Bonczek, Holsapple, and Whinston (1981) concluded "With the continued and rapid decline in computing costs, there is the potential of using computers to enhance the decision-making capabilities of individuals. A theory of the entire process of decision making should be the basis for introducing computer technology into decision processes in order to enhance decision-making capabilities. It is from such a theory of decision making that we can build generalized decision support systems (p. 380)."

What Is Frequent Pattern Analysis ?

- Frequency pattern: a pttern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set. An intrinsic and important property of datasets.

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

{ Break }

{Bread, Beer}

{Diaper, Beer, Milk}

……

What Is Frequent Pattern Analysis ?

- Frequency pattern: a pttern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set. An intrinsic and important property of datasets.
- Motivation: Finding inherent regularities in data.
  - What products were often purchased together?-Beer and dispers?!
  - What are the subsequent purchases after buying a PC?
  - What kinds of DNA are sensitive to this new drug?
  - Can we automatically classify web documents.

Basic Consepts: Frequency Patterns

- Itemset: A set of one or more items
- K-itemset: X={ x_1, …, x_k }
- (absolute) Support count of X: Frequency or occurrence of an item X
- (releative) Support, s, is the fraction of transactions that contains X (i.e., the probability that a transaction contains X)
- An itemset X is **frequent** if X's support is no less than a *minimum threshold.*

Association Analysis

- Itemset: X={ Bread, Milk, Beer, Eggs, Coke, Diaper }

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

K=1

Tid	Sub Itemsets	Support
1	Bread	4/5
2	Milk	4/5
3	Beer	3/5
4	Eggs	1/5
5	Coke	2/5
6	Diaper	4/5

K=2

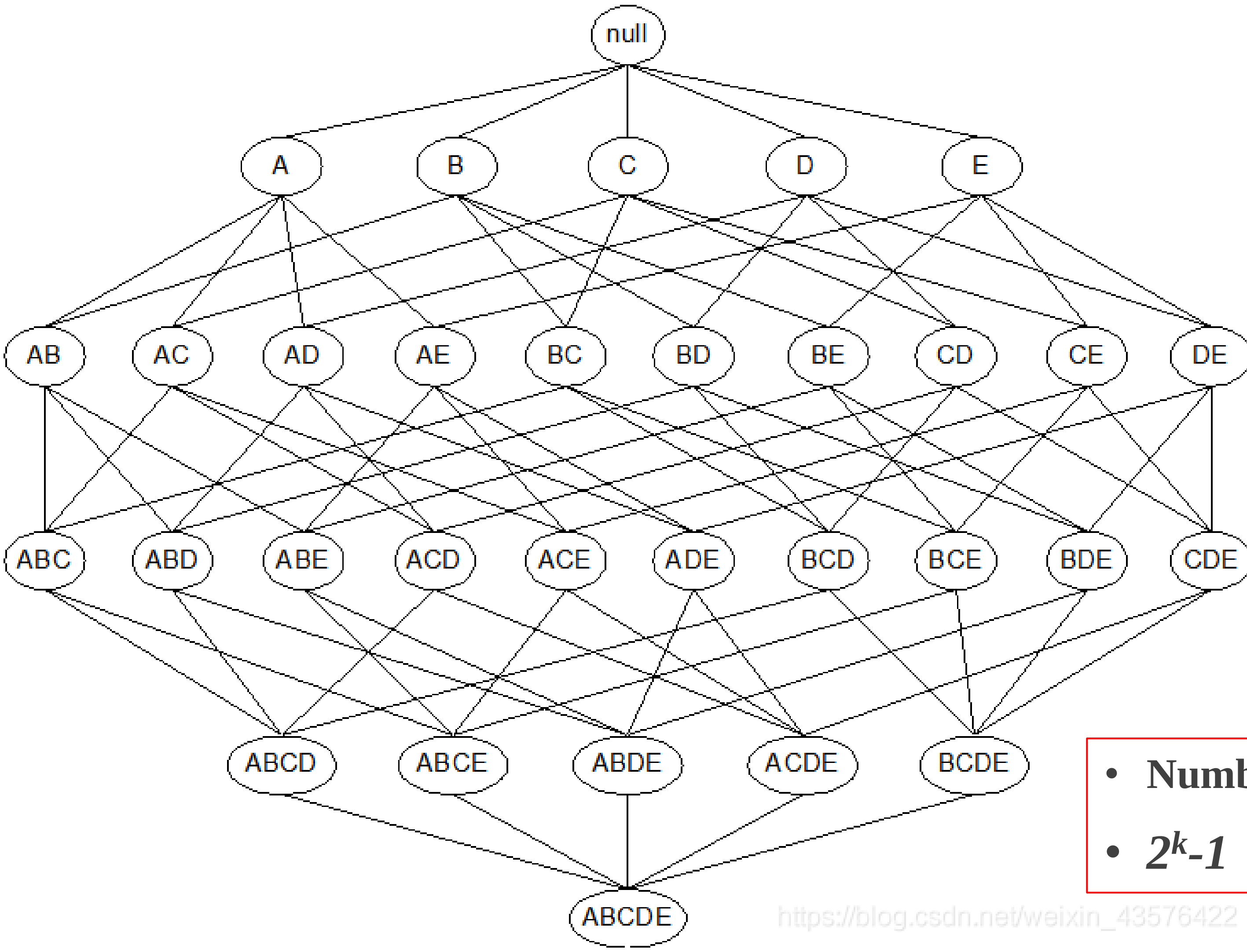
Tid	Sub Itemsets	Support
1	Break, Milk	3/5
2	Bread, Beer	2/5
3	Bread, Eggs	1/5
4	Bread, Coke	1/5
5	Bread, Diaper	3/5
6	Milk, Beer	2/5
7	Milk, Eggs	0/5
8	Milk, Coke	2/5
9	Milk, Diaper	3/5
10	Beer, Eggs	1/5
11	Beer, Coke	1/5
12	Beer, Diaper	3/5
13	Eggs, Coke	0/5
14	Eggs, Diaper	1/5
15	Coke, Diaper	2/5

K=3

Tid	Sub Itemsets	Support
1	Break, Milk, Beer	1/5
2	Break, Milk, Eggs	0/5
3	Break, Milk, Coke	1/5
4	Break, Milk, Diaper	2/5
5	…….	…
6		
7		
8		
9		
10		
11		
12		
13		
14		
…	…	……

Association Analysis

- Frequent Itemset Generation $I=\{A,B,C,D,E\}$

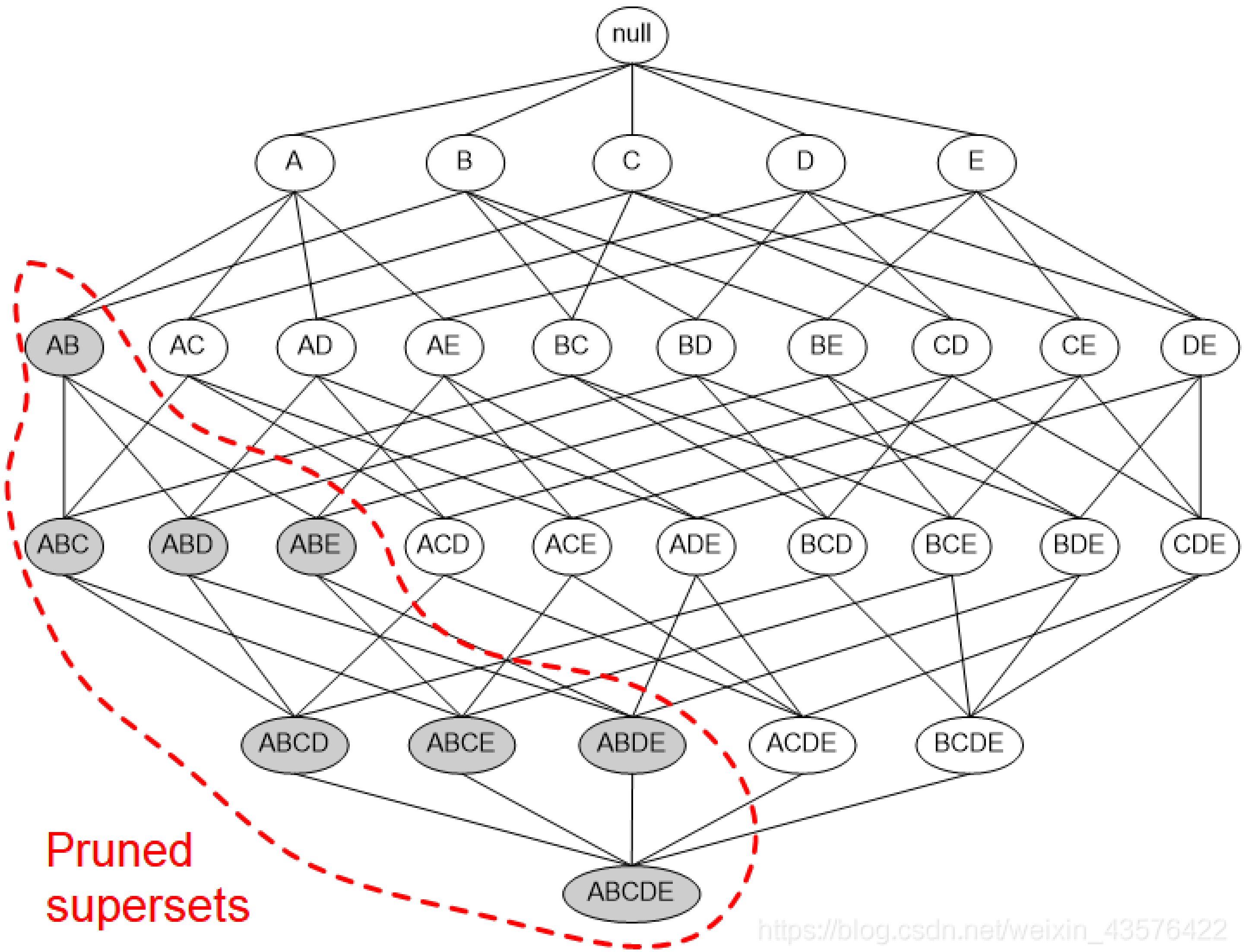


- **Number of sub itemsets:**
- 2^k-1

Apriori: A candidate Generation & Test Approach

- Apriori pruning principle: If there is any itemset which is infrequent, its superset should not be generated/tasted!
- Method:
  - Initially, scan DB once to get frequent 1-itemset
  - Generate length (k+1) candidate itemsets from length k frequent itemsets
  - Test the candidates against DB
  - Terminate when no frequent of candidate set can generated

Association Analysis



The Apriori Algorithm – Example

$Sup_{min} = 2$

Database TDB

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

1^{st} scan

C_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

L_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

C_2

Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

2^{nd} scan

C_2

Itemset	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

L_2

Itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

C_3

Itemset
{B, C, E}

3^{rd} scan

L_3

Itemset	sup
{B, C, E}	2

Association Analysis- *The Apriori Algorithm*

- minimum threshold of support for frequently pattern = 3

TID	Items
1	Bread, Milk
2	Beer, Bread, Diaper, Eggs
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Bread, Coke, Diaper, Milk

Items (1-itemsets)

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Items (2-itemsets)

Itemset
{Bread,Milk}
{Bread, Beer }
{Bread,Diaper}
{Beer, Milk}
{Diaper, Milk}
{Beer,Diaper}

Itemset	Count
{Bread,Milk}	3
{Beer, Bread}	2
{Bread,Diaper}	3
{Beer,Milk}	2
{Diaper,Milk}	3
{Beer,Diaper}	3

Items (3-itemsets)

Itemset	Count
{ Beer, Diaper, Milk}	2
{ Beer,Bread, Diaper}	2
{Bread, Diaper, Milk}	2
{Beer, Bread, Milk}	1

Thanks!