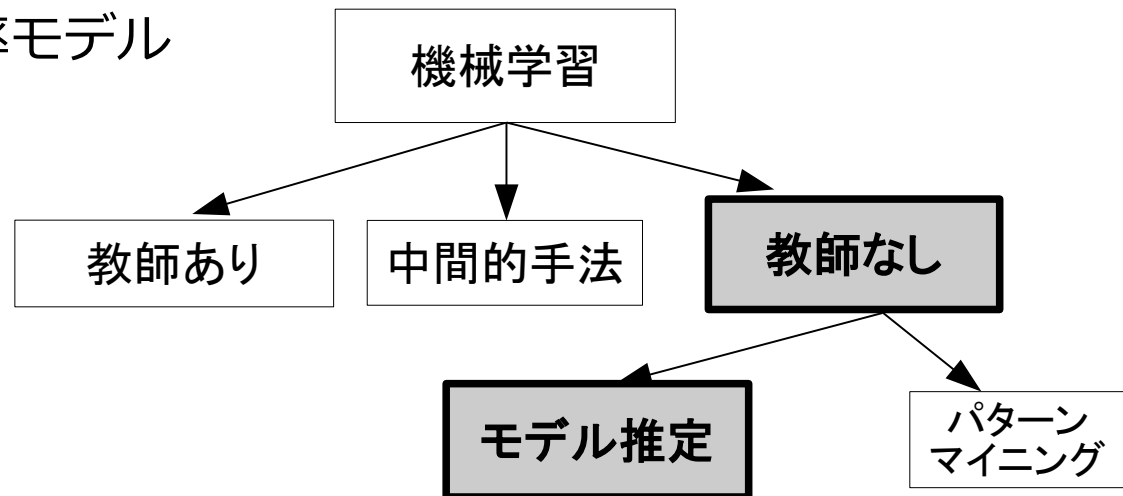


10. モデル推定

- 問題設定
 - 教師なし学習
 - 数値入力 → クラスモデル
 - クラスモデルの例
 - クラスの分割結果
 - クラスの確率モデル



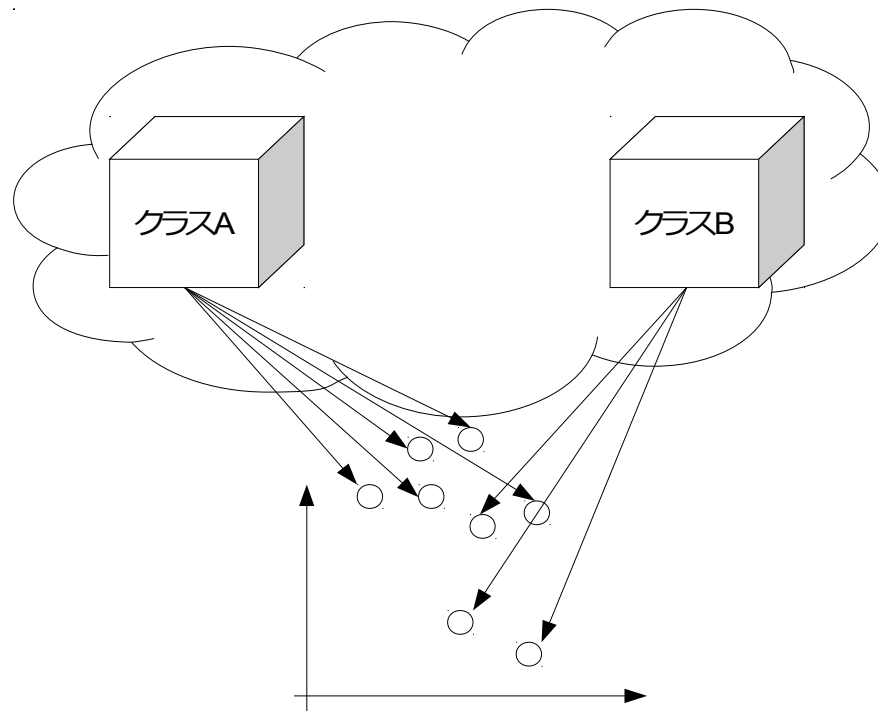
10.1 問題の定義

- 学習データ

$$\{x^{(i)}\} \quad i = 1, \dots, N$$

- 問題設定

- 特徴ベクトル x が生成された元のクラスの性質を推定する



10.2 クラスタリング

- クラスタリングとは
 - 対象のデータを、
内的結合（同じ集合内のデータ間の距離は小さく）
と
外的分離（異なる集合間の距離は大きく）
が達成されるような部分集合に分割すること
- クラスタリング手法の分類
 - 階層的手法
 - ボトムアップ的にデータをまとめてゆく
 - 分割最適化手法
 - トップダウン的にデータ集合を分割してゆく

要するに
塊を見つ
けること

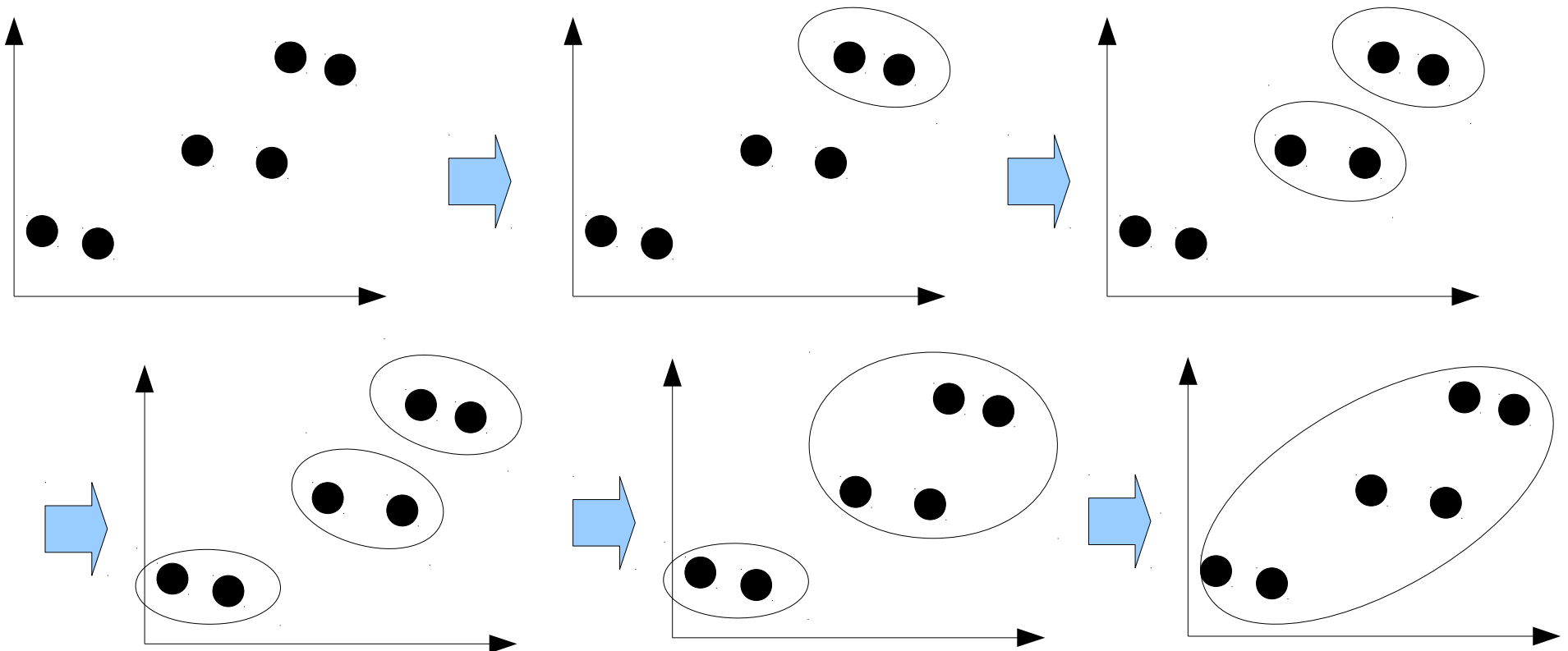
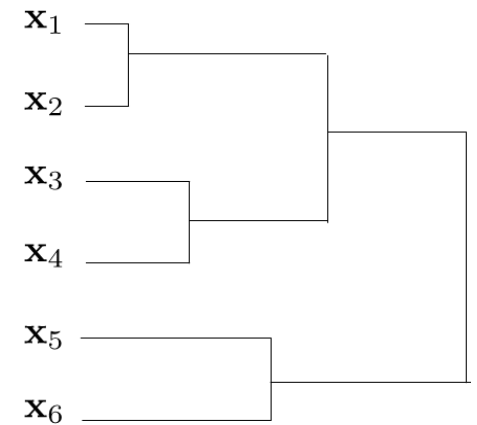
10.2.1 階層的クラスタリング

- 階層的クラスタリングとは

- 1.1 データ 1 クラスタからスタート

- 2.最も近接するクラスタをまとめる

- 3.全データが 1 クラスタになれば終了



10.2.2 分割最適化クラスタリング — k-means アルゴリズム —

- 分割最適化クラスタリングとは
 - データ分割の良さを評価する関数を定め、その評価関数の値を最適化することを目的とする
 - ただし、全ての可能な分割に対して評価値を求めることは、データ数 N が大きくなると、不可能
 - 2 分割で 2^N 通り
 - 探索によって、準最適解を求める

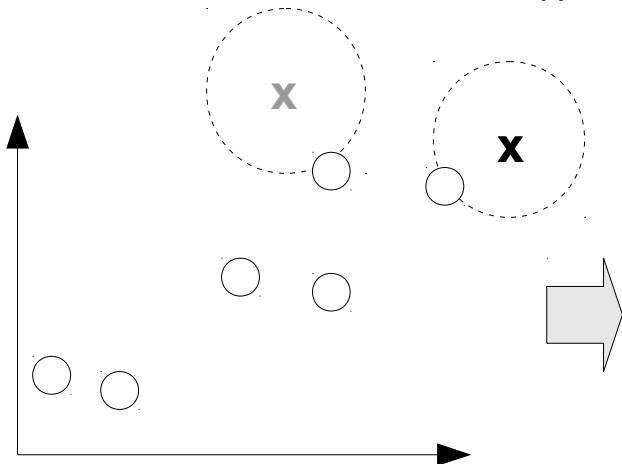
10.2.2 分割最適化クラスタリング — k-means アルゴリズム —

- k-Means アルゴリズム

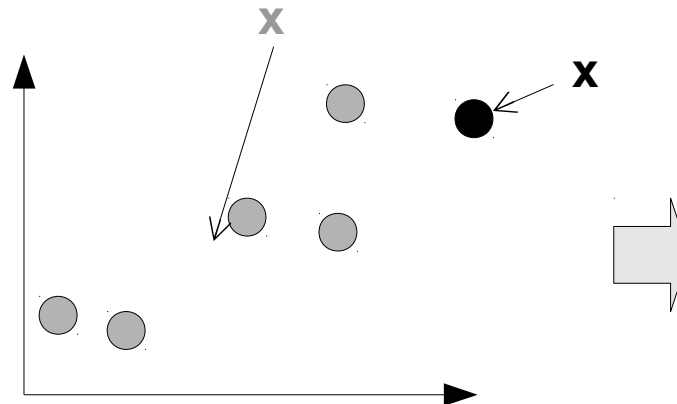
1. 分割数 k を予め与える

2. 乱数で k 個のクラスタ中心を設定し、逐次更新

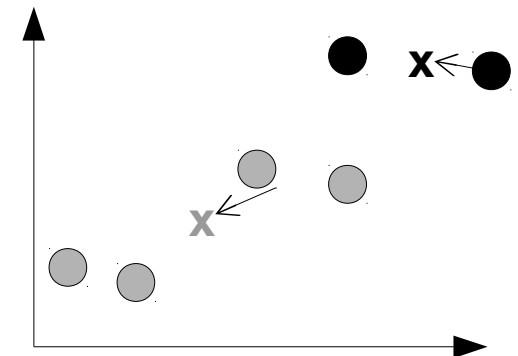
$k=2$ とし、初期値として
乱数でクラスタ中心を配置



全データを近い方のクラスタ
中心に所属させる。そして、
クラスタ中心を所属している
データの平均へ移動。



左の処理を繰り返す。



10.2.3 自動分割最適化クラスタリング — X-means アルゴリズム —

- k-means 法の問題点
 - 分割数 k を予め決めなければならない
- 解決法 \Rightarrow X-means アルゴリズム
 - 2 分割から始めて、分割数を適応的に決定する
 - 分割の妥当性の判断： BIC(Bayesian information criterion) が小さくなれば、分割を継続

$$BIC = -2 \log L + q \log N$$

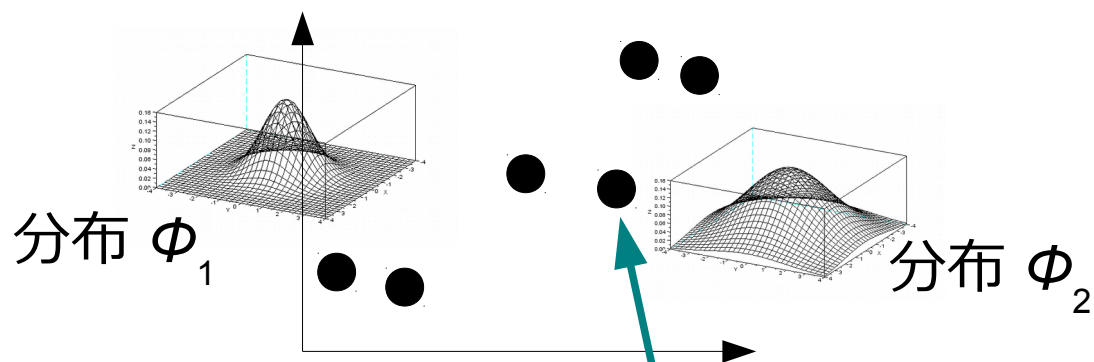
- L : モデルの尤度
- q : モデルのパラメータ数
- N : データ数

パラメータで表される
統計モデルの選択基準
(小さいほどよいモデル)

10.4 確率密度推定

- 教師なし学習で識別器を作る問題
 - クラスタリング結果からは、1 クラス 1 プロトタイプ
の単純な識別器しかできない
 - 各クラスの事前確率や確率密度関数も推定したい

➡ EM アルゴリズム



分布 ϕ_1 の再計算の際、
重み 0.2 だけ寄与する

$$0.2\phi_1 + 0.8\phi_2$$

10.4 確率密度推定

- k-means 法の一般化
 - k 個の平均ベクトルを乱数で決める
⇒ k 個の正規分布を乱数で決める
 - 平均ベクトルとの距離を基準に、各データをいずれかのクラスタに所属させる
⇒ 各分布が各データを生成する確率を計算し、
各クラスタにゆるやかに帰属させる
 - 所属させたデータをもとに平均ベクトルを再計算
⇒ 各データのクラスタへの帰属度に基づき各分布のパラメータ（平均値、共分散行列）を再計算

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... Open UR... Open DB... Generate... Undo Edit... Save...

Filter
Choose None Apply

Current relation
Relation: iris Instances: 150 Attributes: 5 Sum of weights: 150

Attributes
All None Invert Pattern

No.	Name
1	sepal.length
2	sepal.width
3	petal.length
4	petal.width
5	class

Selected attribute
Name: class Missing: 0 (0%) Type: Nominal Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	Iris-setosa	50	50.0
2	Iris-versicolor	50	50.0
3	Iris-virginica	50	50.0

Class: class (Nom) Visualize All

50 50 50

Remove

Status OK Remove selected attributes. Log x 0

class ラベルを消す

Weka の HierarchicalCluster

The screenshot shows the Weka Explorer application window. The 'Cluster' tab is selected. The 'Clusterer' dropdown is set to 'HierarchicalClusterer -N 2 -L WARD -P -A "weka.core.EuclideanDistance" -R first-last'. The 'Cluster mode' section has 'Use training set' selected. The 'Percentage split' is set to 66%. The 'Classes to clusters evaluation' dropdown is set to '(Num) petalwidth'. The 'Store clusters for visualization' checkbox is checked. The 'Ignore attributes' button is visible. The 'Start' button is highlighted. The 'Result list' shows two entries: '17:46:06 - HierarchicalClusterer' and '17:46:48 - HierarchicalClusterer'. The 'Clusterer output' pane displays the following text:

```
Scheme: weka.clusterers.HierarchicalClusterer -N 2 -L WARD -P -A "weka.core.EuclideanDistance" -R first-last
Relation: iris-weka, filters, unsupervised, attribute, Remove-R5
Instances: 150
Attributes: 4
    sepallength
    sepalwidth
    petallength
    petalwidth
Test mode: evaluate on training data

=== Clustering model (full training set) ===

Cluster 0
((((((0.2:0.03254,0.2:0.03254):0.01891,((0.2:0.02778,0.2:0.02778):0.00843,

Cluster 1
((((((1.4:0.07344,1.5:0.07344):0.08446,((1.5:0.09914,1.4:0.09914):0.0122, (

Time taken to build model (full training data) : 0.13 seconds

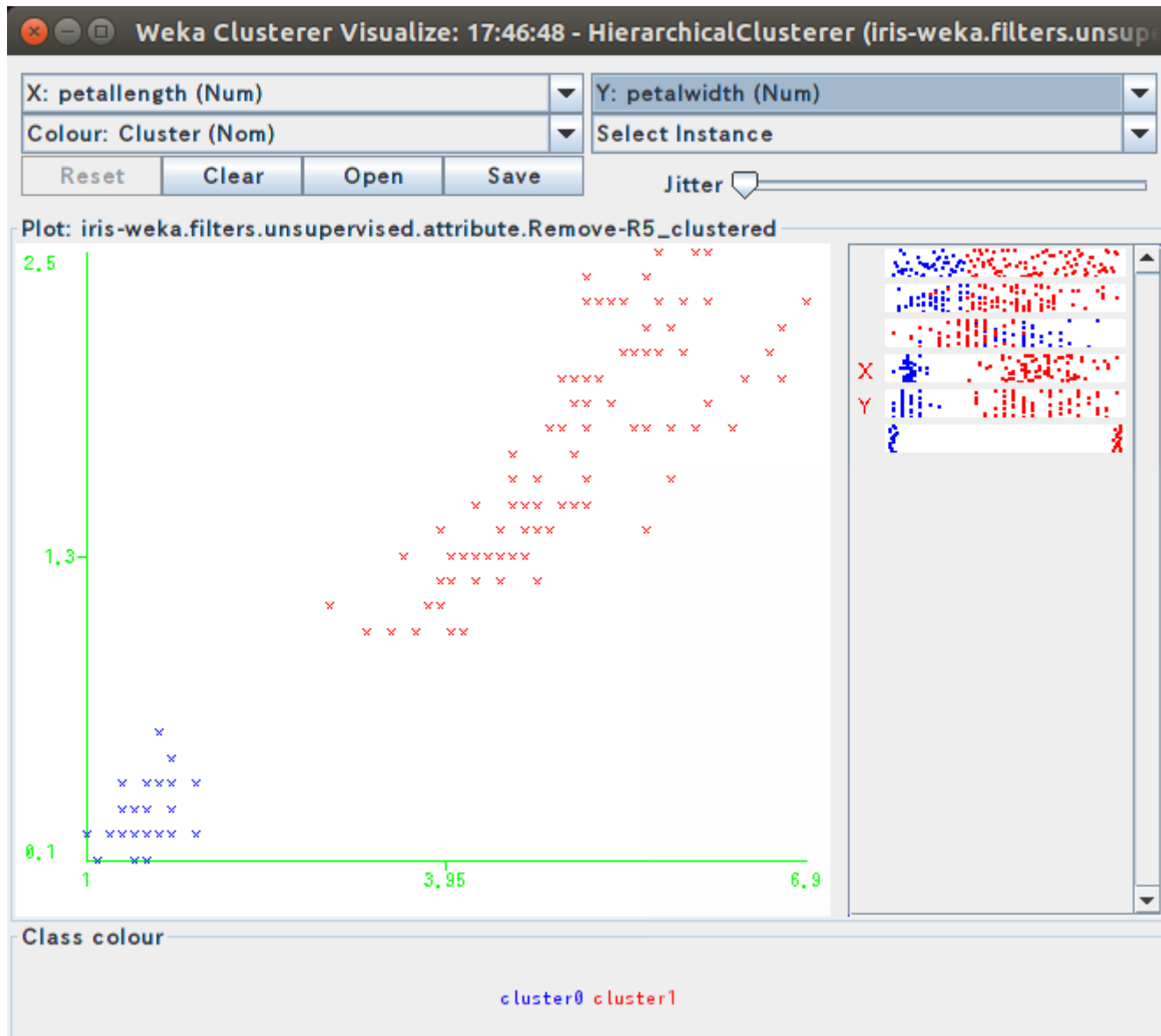
=== Model and evaluation on training set ===

Clustered Instances

0      50 ( 33%)
1      100 ( 67%)
```

The 'Status' bar at the bottom shows 'OK'. The 'Log' button and a small icon are also visible.

Weka の HierarchicalCluster



Weka の SimpleKMeans

The screenshot shows the Weka Explorer interface with the SimpleKMeans clustering algorithm applied to a dataset. The 'Clusterer' tab is active, and the 'Cluster mode' is set to 'Use training set'. The 'Clusterer output' pane displays the results of the clustering process, including the within-cluster sum of squared errors, initial starting points, missing values handling, final cluster centroids, and a table of clustered instances.

Clusterer
Choose SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A 'weka.core

Cluster mode
☒ Use training set
☐ Supplied test set Set...
☐ Percentage split % 66
☐ Classes to clusters evaluation (Num) petalwidth
☒ Store clusters for visualization

Ignore attributes
Start Stop

Result list (right-click for options)
17:49:29 - HierarchicalClusterer
17:50:11 - SimpleKMeans

Clusterer output
Within cluster sum of squared errors: 6.998114004826762
Initial starting points (random):
Cluster 0: 6.1,2.9,4.7,1.4
Cluster 1: 6.2,2.9,4.3,1.3
Cluster 2: 6.9,3.1,5.1,2.3
Missing values globally replaced with mean/mode
Final cluster centroids:
Attribute Full Data Cluster#
(150.0) (61.0) (50.0) (39.0)

sepalength 5.8433 5.8885 5.006 6.8462
sepalwidth 3.054 2.7377 3.418 3.0821
petallength 3.7587 4.3967 1.464 5.7026
petalwidth 1.1987 1.418 0.244 2.0795
Time taken to build model (full training data) : 0.02 seconds
=== Model and evaluation on training set ===
Clustered Instances
0 61 (41%)
1 50 (33%)
2 39 (26%)

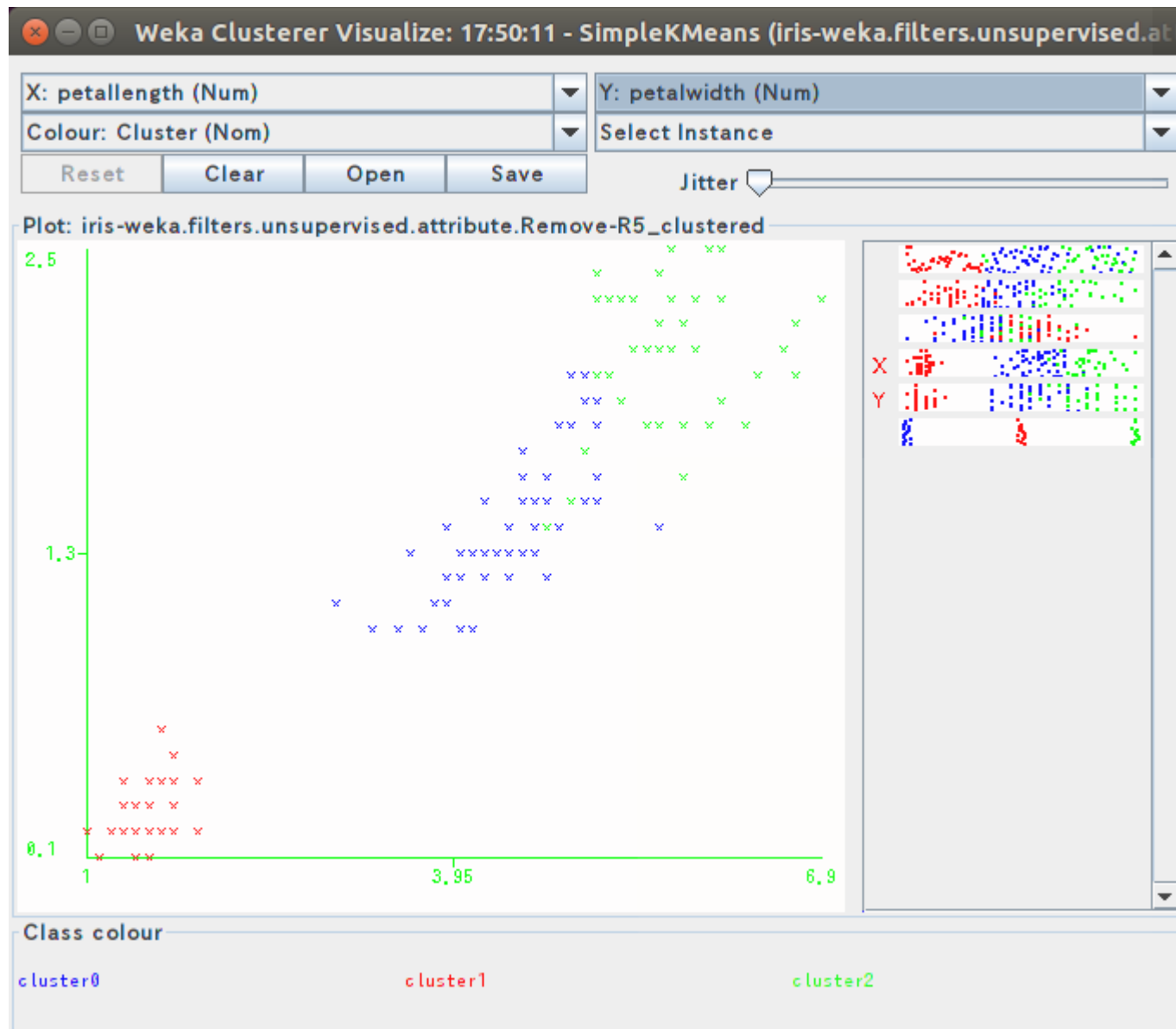
Status
OK

Log x 0

クラスタ中心
が得られている

クラスタ数 = 3

Weka の SimpleKMeans



Weka の EM

Weka Explorer

Preprocess | Classify | **Cluster** | Associate | Select attributes | Visualize

Clusterer

Choose EM -I 100 -N 3 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100

Cluster mode

☒ Use training set

☐ Supplied test set Set...

☐ Percentage split % 66

☐ Classes to clusters evaluation (Num) petalwidth

☒ Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

- 17:49:29 - HierarchicalClusterer
- 17:50:11 - SimpleKMeans
- 17:51:30 - EM
- 17:51:50 - EM**

Clusterer output

Attribute	Cluster 0	Cluster 1	Cluster 2
	(0.41)	(0.33)	(0.25)
=====			
sepal.length			
mean	5.9275	5.006	6.8085
std. dev.	0.4817	0.3489	0.5339
sepal.width			
mean	2.7503	3.418	3.0709
std. dev.	0.2956	0.3772	0.2867
petal.length			
mean	4.4057	1.464	5.7233
std. dev.	0.5254	0.1718	0.4991
petal.width			
mean	1.4131	0.244	2.1055
std. dev.	0.2627	0.1061	0.2456

Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0	64 (43%)
1	50 (33%)
2	36 (24%)

確率分布が得られている

Status OK

Log x 0

Weka の EM

