

情報工学実験4: データマイニング班

(補足) クラスタリング

1. 問題設定例(クラスタリング)
2. 同一データセットでも手法により結果が異なる例(scikit-learn)
3. クラスタリングにおける「似ているものの同士を同一クラスタにする」問題

実験ページ: <http://ie.u-ryukyu.ac.jp/~tnal/2014/info4/dm/>

Example: *Iris* flower data set **WITHOUT** classes

http://en.wikipedia.org/wiki/Iris_flower_data_set

(1) What is experience E?

(2) What is task T?

(3) How to measure the performance P?

• Clustering

- is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense or another) to each other than to those in other groups (clusters).
- training data consists of a set of input vectors **x** **without any corresponding target values**.
- Dataset = **samples** vs. **features**

4 features

Fisher's *Iris* Data

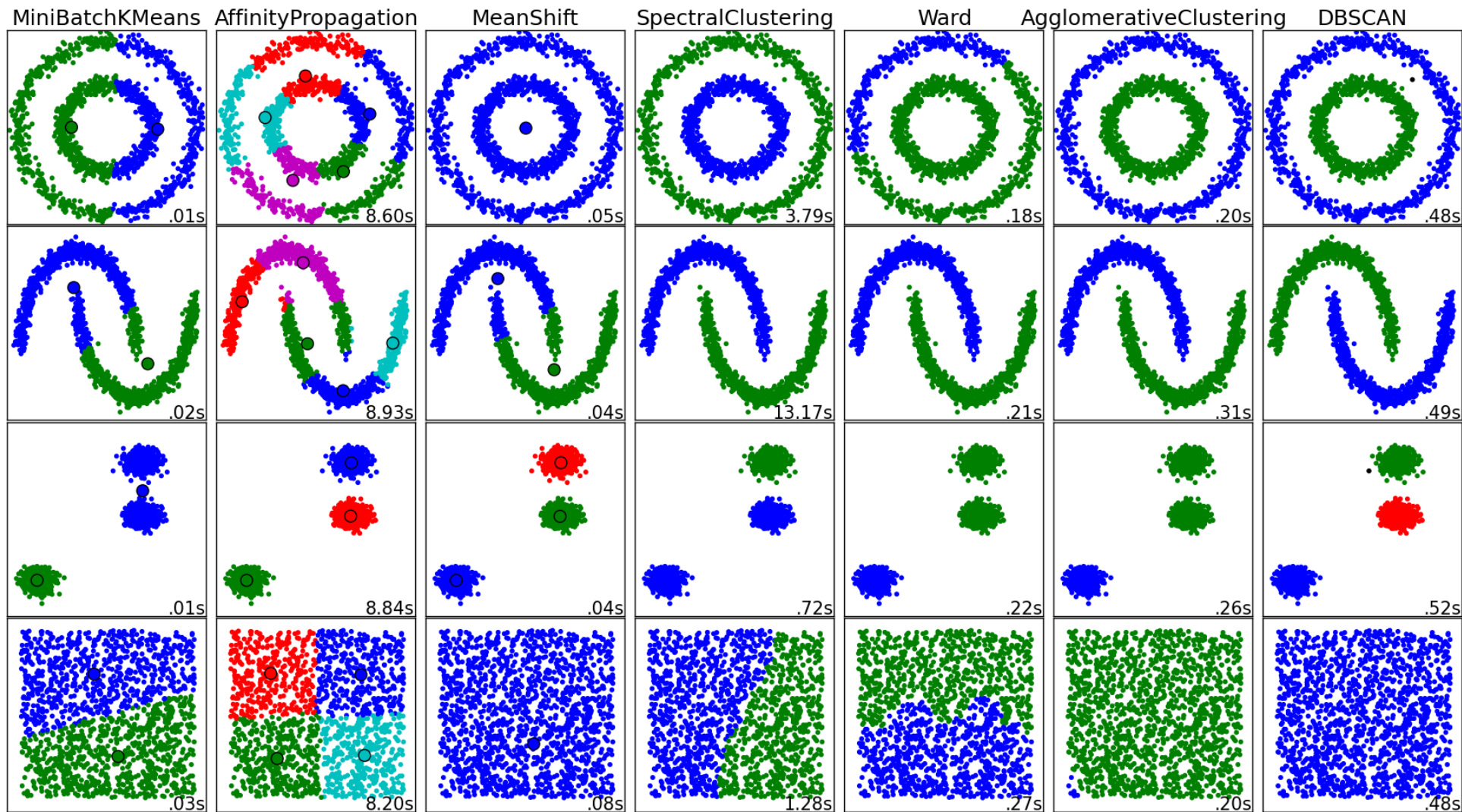
Don't use at learning

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	<i>I. setosa</i>
4.9	3.0	1.4	0.2	<i>I. setosa</i>
4.7	3.2	1.3	0.2	<i>I. setosa</i>
4.6	3.1	1.5	0.2	<i>I. setosa</i>
5.0	3.6	1.4	0.2	<i>I. setosa</i>

1 sample

(scikit-learn) Overview of clustering methods

<http://scikit-learn.org/stable/modules/clustering.html#clustering>



more similar? same group?

