1 关于数据集的一些基本术语

1.1 数据集：一组记录的集合

In [6]:

```
import sklearn
from sklearn.datasets import load_iris
import pandas as pd
```

In [7]:

```
1 print(load_iris()['DESCR'])
```

```
Iris Plants Database
====================

Notes
-----
Data Set Characteristics:
    :Number of Instances: 150 (50 in each of three classes)
    :Number of Attributes: 4 numeric, predictive attributes and the class
    :Attribute Information:
        - sepal length in cm
        - sepal width in cm
        - petal length in cm
        - petal width in cm
        - class:
                - Iris-Setosa
                - Iris-Versicolour
                - Iris-Virginica
    :Summary Statistics:

    ============== ==== ==== ======= ===== ====================
                    Min  Max   Mean    SD   Class Correlation
    ============== ==== ==== ======= ===== ====================
    sepal length:   4.3  7.9   5.84   0.83    0.7826
    sepal width:    2.0  4.4   3.05   0.43   -0.4194
    petal length:   1.0  6.9   3.76   1.76    0.9490  (high!)
    petal width:    0.1  2.5   1.20  0.76     0.9565  (high!)
    ============== ==== ==== ======= ===== ====================

    :Missing Attribute Values: None
    :Class Distribution: 33.3% for each of 3 classes.
    :Creator: R.A. Fisher
    :Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
    :Date: July, 1988

This is a copy of UCI ML iris datasets.
http://archive.ics.uci.edu/ml/datasets/Iris (http://archive.ics.uci.edu/ml/datasets/
Iris)

The famous Iris database, first used by Sir R.A Fisher

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 Duda & Hart, for example.)  The
data set contains 3 classes of 50 instances each, where each class refers to a
type of iris plant.  One class is linearly separable from the other 2; the
latter are NOT linearly separable from each other.

References
----------
   - Fisher,R.A. "The use of multiple measurements in taxonomic problems"
     Annual Eugenics, 7, Part II, 179-188 (1936); also in "Contributions to
     Mathematical Statistics" (John Wiley, NY, 1950).
   - Duda,R.O., & Hart,P.E. (1973) Pattern Classification and Scene Analysis.
     (Q327.D83) John Wiley & Sons.  ISBN 0-471-22361-1.  See page 218.
   - Dasarathy, B.V. (1980) "Nosing Around the Neighborhood: A New System
     Structure and Classification Rule for Recognition in Partially Exposed
     Environments".  IEEE Transactions on Pattern Analysis and Machine
```

Intelligence, Vol. PAMI-2, No. 1, 67-71.
- Gates, G.W. (1972) "The Reduced Nearest Neighbor Rule". IEEE Transactions on Information Theory, May 1972, 431-433.
- See also: 1988 MLC Proceedings, 54-64. Cheeseman et al"s AUTOCLASS II conceptual clustering system finds 3 classes in the data.
- Many, many more ...

In [8]:

```
data = sklearn.datasets.load_iris()['data']
target = sklearn.datasets.load_iris()['target']

columns = sklearn.datasets.load_iris()['feature_names']
target_names = sklearn.datasets.load_iris()['target_names']
```

In [9]:

```
iris =    pd.concat([pd.DataFrame(data,columns=columns),
          pd.DataFrame([[i,target_names[i]] for i in target],columns=['target','target_label'])
          axis=1)
```

In [10]:

```
iris.head()
```

Out[10]:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target	target_label
0	5.1	3.5	1.4	0.2	0	setosa
1	4.9	3.0	1.4	0.2	0	setosa
2	4.7	3.2	1.3	0.2	0	setosa
3	4.6	3.1	1.5	0.2	0	setosa
4	5.0	3.6	1.4	0.2	0	setosa

In [40]:

```
def func(num):
    res = [0,0,0]
    res[num] = 1
    return res
iris['target_vertor']=iris['target'].apply(func)
```

In [41]:

```
1 iris.head()
```

Out[41]:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target	target_label	target_vertor
0	5.1	3.5	1.4	0.2	0	setosa	[1, 0, 0]
1	4.9	3.0	1.4	0.2	0	setosa	[1, 0, 0]
2	4.7	3.2	1.3	0.2	0	setosa	[1, 0, 0]
3	4.6	3.1	1.5	0.2	0	setosa	[1, 0, 0]
4	5.0	3.6	1.4	0.2	0	setosa	[1, 0, 0]

1.2 样本

数据集有许多条记录组成，关于一个时间或对象的描述，称为一个**样本**

In [27]:

```
1 #这是五条样本
2 iris[columns].head()
```

Out[27]:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2

1.3 样例

带有样本“结果”信息（这里是哪一个种类的花）的样本成为样例

In [28]:

```
1 iris.head()
```

Out[28]:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

2 机器学习的三种任务

分类，聚类，回归

分类：一种监督学习方法，如上面的鸢尾花数据集，给出一条新的鸢尾花样本，预测样本属于哪个类别（即预测的结果是离散值），是一个分类任务。只涉及两个类别的分类称为“二分类”，涉及多个类别时，称为多分类任务。

聚类：一种非监督学习方法，是指在没有样本对应的结果信息的条件下，将训练集中的数据分成若干组的方法（常见的聚类方法在sklearn中均有实现，通过API即可方便的调用，详见https://github.com/apachecn/scikit-learn-doc-zh/ (https://github.com/apachecn/scikit-learn-doc-zh/) 不过这不是我们这节课关注的重点，有兴趣的同学课后可以自行查阅）

回归：预测的是连续值，如属于某一类花的概率，则为回归。

！注：有的时候回归任务也可以等同于分类任务，比如将概率p>0.5的样例视为某一类花，则任务由回归任务变成了分类任务，具体在后面会有更加详细的讲解。

> 根据训练数据是否拥有标记信息，学习任务可大致划分为两大类：“监督学习”和“无监督学习”，分类和回归是前者的代表，而聚类则是后者的代表

——周志华《机器学习》

In [ ]:

```
1
```