



暨南大學  
JINAN UNIVERSITY



# 机器学习

## 第二章：机器学习基础知识

黄斐然

2021/3/8

# 目录

---

1

**机器学习的数据**

2

机器学习的主要任务

3

机器学习的学习方式

# 机器学习的数据

- 著名的鸢尾花数据集：UC Irvine的机器学习库



## Iris Data Set

Download: [Data Folder](#), [Data Set Description](#)

**Abstract:** Famous database; from Fisher, 1936



Data Set Characteristics:	Multivariate	Number of Instances:	150	Area:	Life
Attribute Characteristics:	Real	Number of Attributes:	4	Date Donated	1988-07-01
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	3858259

**Source:**

Creator:

R.A. Fisher

Donor:

Michael Marshall (MARSHALL%PLU'@'io.arc.nasa.gov)

<https://archive.ics.uci.edu/ml/datasets/iris>

**Welcome to the UC Irvine Machine Learning Repository!**

We currently maintain 585 data sets as a service to the machine learning community. You may [view all data sets](#) through our searchable interface. For a general overview of the Repository, please visit our [About page](#). For information about citing data sets in publications, please read our [citation policy](#). If you wish to donate a data set, please consult our [donation policy](#). For any other questions, feel free to [contact the Repository librarians](#).

Supported By: In Collaboration With:

**Latest News:**

- 09-24-2018: Welcome to the new Repository admins Dheeru Dua and Efi Karra Taniskidou!
- 04-04-2013: Welcome to the new Repository admins Kevin Bache and Moshe Lichman!
- 03-01-2010: Note from donor regarding Netflix data
- 10-16-2009: Two new data sets have been added.
- 09-14-2009: Several data sets have been added.
- 03-24-2008: New data sets have been added!
- 06-25-2007: Two new data sets have been added: UJI Pen Characters, MAGIC Gamma Telescope

**Featured Data Set: Zoo**

**Task:** Classification  
**Data Type:** Multivariate  
**# Attributes:** 17  
**# Instances:** 101

Artificial, 7 classes of animals

**Newest Data Sets:**

- 02-17-2021: Hungarian Chickenpox Cases
- 12-09-2020: Myocardial Infarction Complications
- 10-14-2020: Gait Classification
- 10-03-2020: Codon usage
- 09-15-2020: In-vehicle coupon recommendation
- 09-14-2020: Dry Bean Dataset
- 09-03-2020: Intelligent Media Accelerometer and Gyroscope (IM-AccGyro) Dataset
- 08-30-2020: AI4I 2020 Predictive Maintenance Dataset
- 08-25-2020: Wispsight Sentiment Corpus

**Most Popular Data Sets (hits since 2007):**

- 3858522: Iris
- 2086894: Adult
- 1613406: Wine
- 1458876: Heart Disease
- 1449685: Wine Quality
- 1448258: Breast Cancer Wisconsin (Diagnostic)
- 1411795: Bank Marketing
- 1334900: Car Evaluation
- 1095325: Human Activity Recognition Using Smartphones

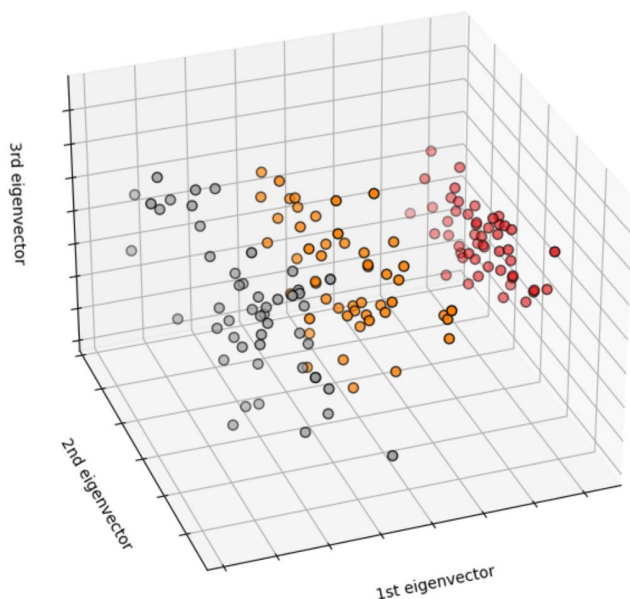
## ● 著名的鸢尾花数据集：Sklearn

### The Iris Dataset

This data sets consists of 3 different types of irises' (Setosa, Versicolour, and Virginica) petal and sepal length, stored in a 150x4 numpy.ndarray

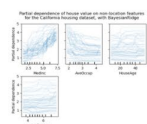
The rows being the samples and the columns being: Sepal Length, Sepal Width, Petal Length and Petal Width.

The below plot uses the first two features. See [here](https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html#sklearn.datasets.load_iris) for more information on this dataset.

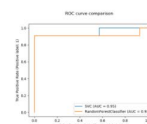


[https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load\\_iris.html#sklearn.datasets.load\\_iris](https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html#sklearn.datasets.load_iris)

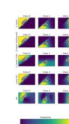
### Examples using `sklearn.datasets.load_iris`



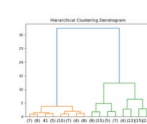
Release Highlights for  
scikit-learn 0.24



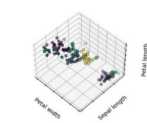
Release Highlights for  
scikit-learn 0.22



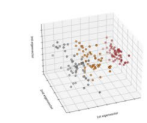
Plot classification  
probability



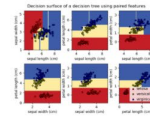
Plot Hierarchical Cluster-  
ing Dendrogram



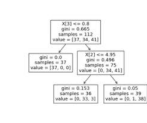
K-means Clustering



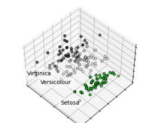
The Iris Dataset



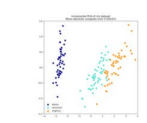
Plot the decision surface  
of a decision tree on the  
iris dataset



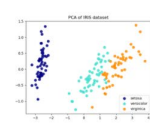
Understanding the  
decision tree structure



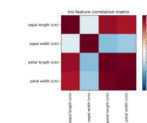
PCA example with Iris  
Data-set



Incremental PCA



Comparison of LDA and  
PCA 2D projection of Iris  
dataset



Factor Analysis (with  
rotation) to visualize  
patterns

# 1 机器学习的数据

- 著名的鸢尾花数据集：UC Irvine的机器学习库

```
from sklearn.datasets import load_iris
```

```
iris = load_iris()  
iris
```

```
{'data': array([[5.1, 3.5, 1.4, 0.2],  
               [4.9, 3. , 1.4, 0.2],  
               [4.7, 3.2, 1.3, 0.2],  
               [4.6, 3.1, 1.5, 0.2],...  
'target': array([0, 0, 0, ... 1, 1, 1, ... 2, 2, 2, ...  
'target_names': array(['setosa', 'versicolor', 'virginica'], dtype='<U10'),  
...}
```

Classes	3
Samples per class	50
Samples total	150
Dimensionality	4
Features	real, positive

sepal length:	4.3	7.9	5.84
sepal width:	2.0	4.4	3.05
petal length:	1.0	6.9	3.76
petal width:	0.1	2.5	1.20

<https://archive.ics.uci.edu/ml/datasets/iris>



*Iris versicolor*



*Iris setosa*



*Iris virginica*

# 1 机器学习的数据

## ■ 著名的鸢尾花数据

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	label
126	6.2	2.8	4.8	1.8	2
60	5.0	2.0	3.5	1.0	1
31	5.4	3.4	1.5	0.4	0
87	6.3	2.3	4.4	1.3	1
125	7.2	3.2	6.0	1.8	2
42	4.4	3.2	1.3	0.2	0
131	7.9	3.8	6.4	2.0	2
22	4.6	3.6	1.0	0.2	0

# 1 机器学习的数据

## ■ 著名的鸢尾花数据集

- **数据集**：数据整体叫数据集（dataset）。
- **样本**：每一行数据可以称为一个样本（sample），或一条数据。
- **特征**：除最后一列，每一列表示数据的一个特征（feature），数据的特征的向量的集合可用 $X$ 表示。
- **标签**：最后一列称为标签（label），数据标签集可以用 $Y$ 表示。

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	label
126	6.2	2.8	4.8	1.8	2
60	5.0	2.0	3.5	1.0	1
31	5.4	3.4	1.5	0.4	0
87	6.3	2.3	4.4	1.3	1
125	7.2	3.2	6.0	1.8	2
42	4.4	3.2	1.3	0.2	0
131	7.9	3.8	6.4	2.0	2
22	4.6	3.6	1.0	0.2	0

$X$   $y$

# 1 机器学习的数据

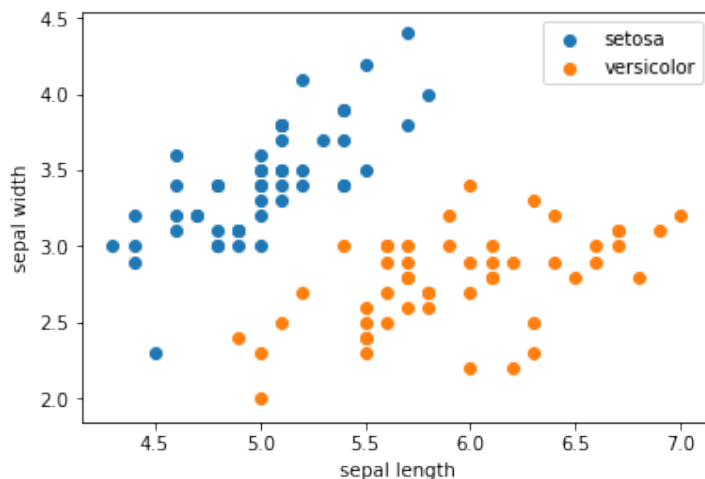
## 著名的鸢尾花数据集

- **特征**：除最后一列，每一列表示数据的一个特征。
- **特征向量**：每一条样本的特征组成的向量称为特征向量。
- **特征空间**：所有特征向量张成的空间称为特征空间。

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	label
126	6.2	2.8	4.8	1.8	2
60	5.0	2.0	3.5	1.0	1
31	5.4	3.4	1.5	0.4	0
87	6.3	2.3	4.4	1.3	1
125	7.2	3.2	6.0	1.8	2
42	4.4	3.2	1.3	0.2	0
131	7.9	3.8	6.4	2.0	2
22	4.6	3.6	1.0	0.2	0

$x_i$

$y_i$



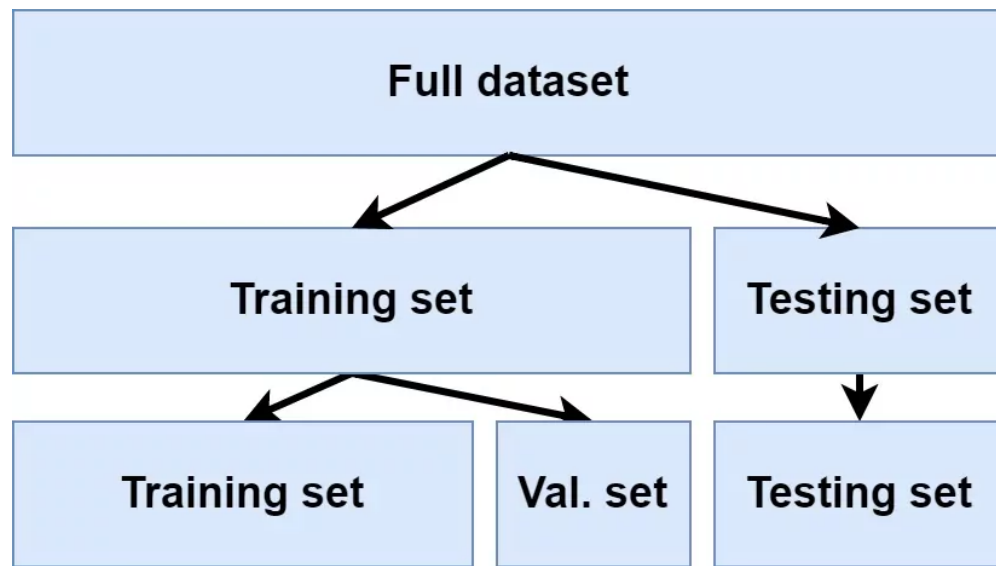
只考虑花萼特征和前  
两类鸢尾花所形成的  
征空间



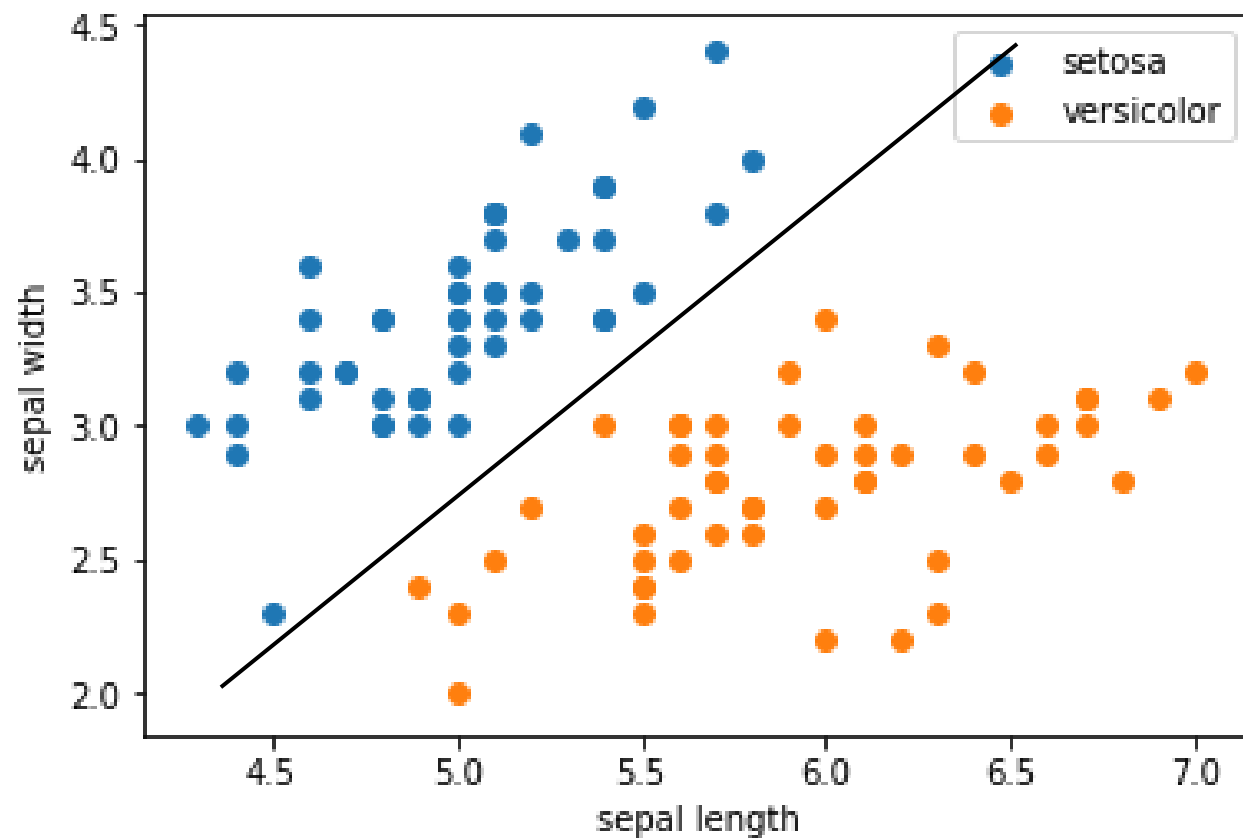
# 1 机器学习的数据

## ■ 数据集分类

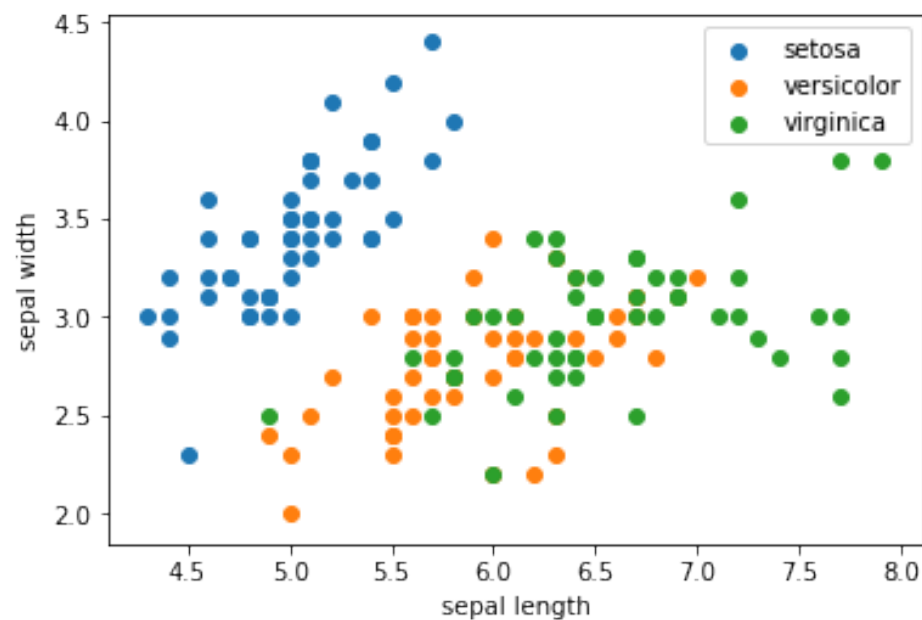
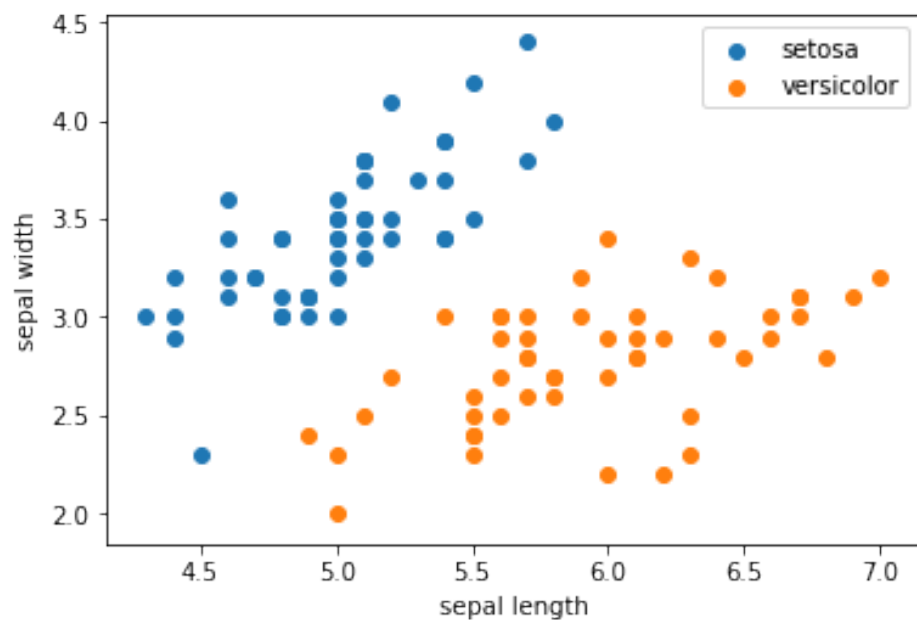
- **训练数据** (training data) : 在训练过程中使用的数据称为训练数据, 每一个样例称为**训练**样本, 全体训练样本集合称为训练集。
- **测试数据** (testing data) : 用于测试学习得到模型的数据称为测试数据, 每一个样例称为**测试**样本, 全体测试样本集合称为测试集。
- **验证数据** (Validation data) 是训练模型时所保留的数据样本, 我们在**调整模型超参数**时, 需要根据它来对模型的能力进行评估。



- 分类任务就是在标签指导下对特征空间进行切分



- 分类任务就是在标签指导下对特征空间进行切分



## 1

[illegible]

- 图像，每一个像素点都是特征
- 28\*28总共784、个特征
- 如果是彩色图·像，特征更多。

# 目录

---

1

机器学习的数据

2

**机器学习的主要任务**

3

机器学习的学习方式

## 2 机器学习的主要任务

### ■ 基本任务

分类

回归

## 2 机器学习的主要任务

### ■ 分类任务



## 2 机器学习的主要任务

### ■ 分类任务

MNIST数字识别数据集





## 2 机器学习的主要任务

### ■ 二分类

- 判断垃圾邮件
- 判断信用卡是否盗刷
- 判断股票涨还是跌
- 检测是否是网络入侵



## 2 机器学习的主要任务

### ■ 多分类

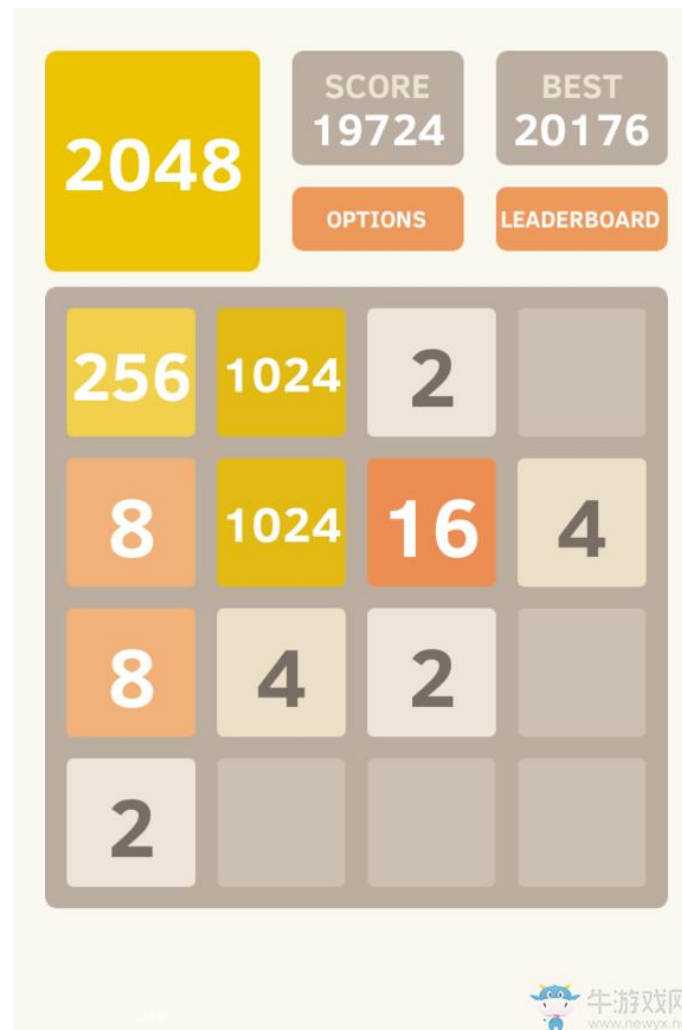
- 数字识别
- 图像识别
- 信用卡风险评级



## 2 机器学习的主要任务

### ■ 多分类

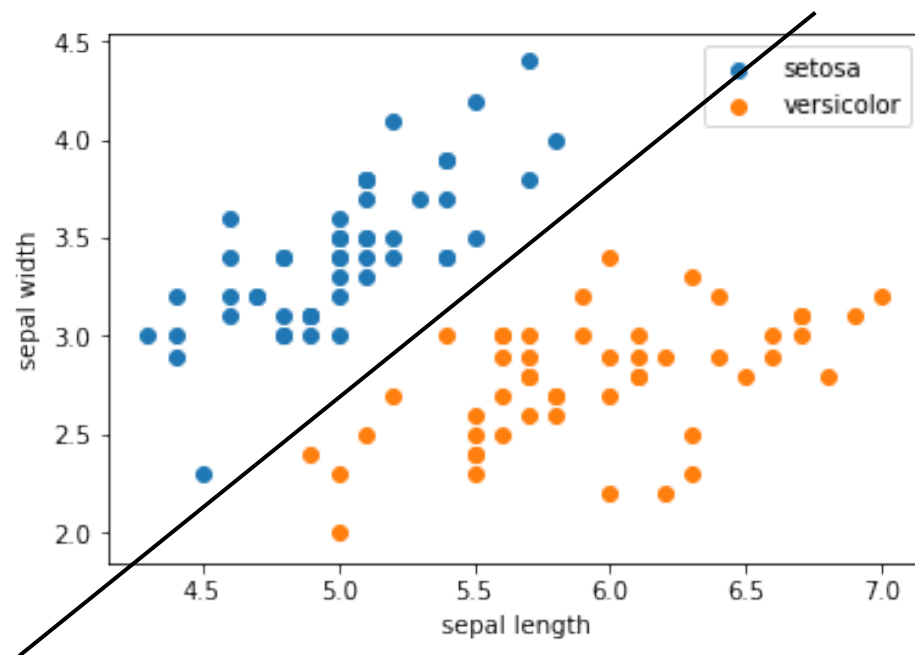
- 数字识别
- 图像识别
- 信用卡风险评级
- 很多复杂的任务也能转为多分类任务



## 2 机器学习的主要任务

### ■ 多分类

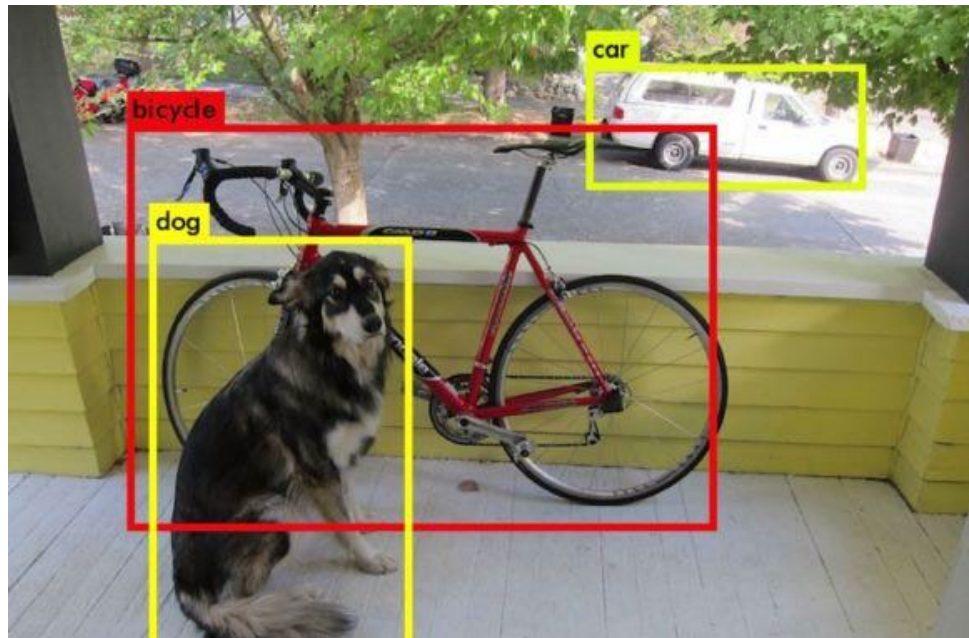
- 有一些算法只支持完成二分类任务
  - SVM、感知机、逻辑回归
- 但是多分类任务可以转化为多个二分类任务
- 有一些算法可以直接支持多分类任务
  - 朴素贝叶斯、决策树



## 2 机器学习的主要任务

### ■ 多标签分类

- 一个样本可以归类成多种类别，不仅限于一个。



## 2 机器学习的主要任务

### ■ 回归

装修	面积	销售价格（万元）
0	123	250
20	150	320
17	87	160
8	102	220

## 2 机器学习的主要任务

### ■ 回归

- 需要预测的是一个连续的值，而非离散的类别。
  - 房屋价格
  - 股票股价
  - 营业额预测
  - 学生成绩
- 回归和分类在某种条件下可以相互转化

装修	面积	销售价格（万元）
0	123	250
20	150	320
17	87	160
8	102	220

## 2 机器学习的主要任务

### ■ 什么是机器学习:

#### 数据集

西瓜数据集

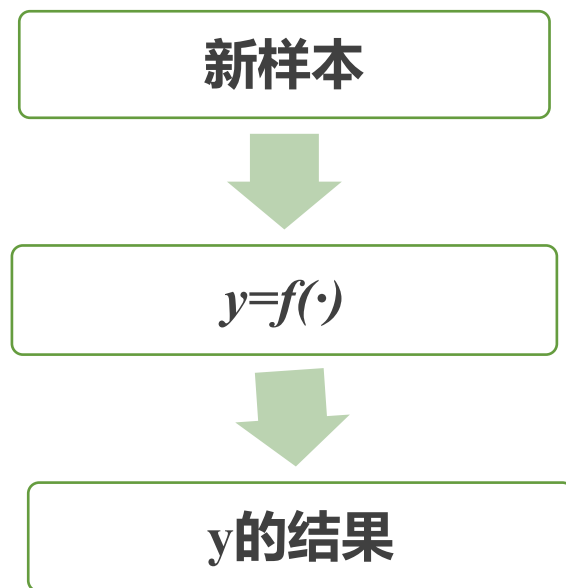
编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	浊响	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否

装修	面积	销售价格 (万元)
0	123	250
20	150	320
17	87	160
8	102	220

#### 归纳学习



#### 应用模型





# 目录

---

1

机器学习的数据

2

机器学习的主要任务

3

**机器学习的学习方式**

### 3 机器学习的学习方式

- 按照学习方式的不同，机器学习可分为很多的类型：
  - 监督学习
  - 无监督学习
  - 强化学习
  - 半监督学习

### 3 机器学习的学习方式

#### ■ 监督学习：

- 利用一组**已知标注的样本**调整模型的参数，使其达到所要求性能的过程，也称为监督训练。
- 包括分类和回归。



### 3 机器学习的学习方式

#### ■ 监督学习：

- 图像已经拥有了标定信息
- 银行已经积累了一定的客户信息和他们信用卡的信用情况
- 医院已经积累了一定的病人信息和他们最终确诊是否患病的情况
- 市场积累了房屋的基本信息和最终成交的金额

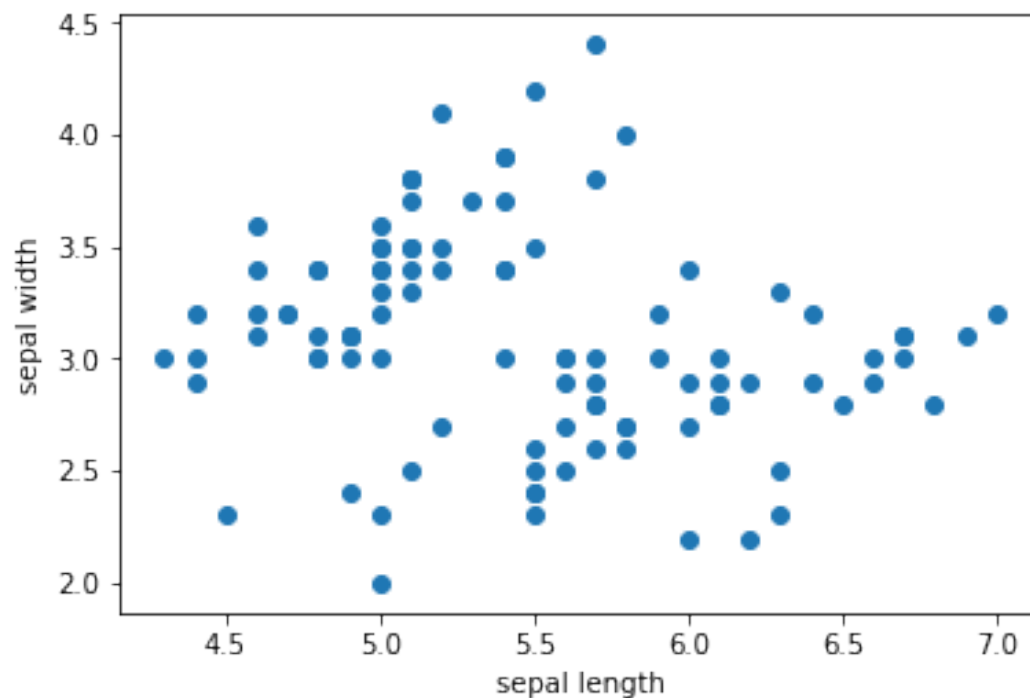
### 3 机器学习的学习方式

- 这门课程的大部分算法属于监督学习：
  - K近邻算法
  - 线性回归
  - 逻辑回归
  - 朴素贝叶斯
  - 支持向量机
  - 决策树
  - 随机森林

### 3 机器学习的学习方式

#### ■ 无监督学习：

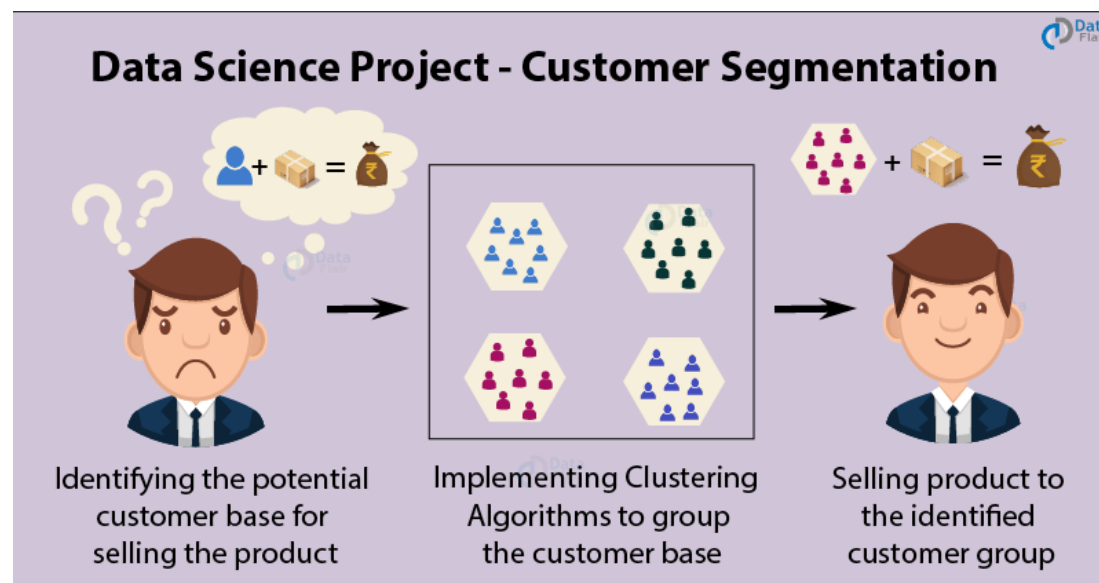
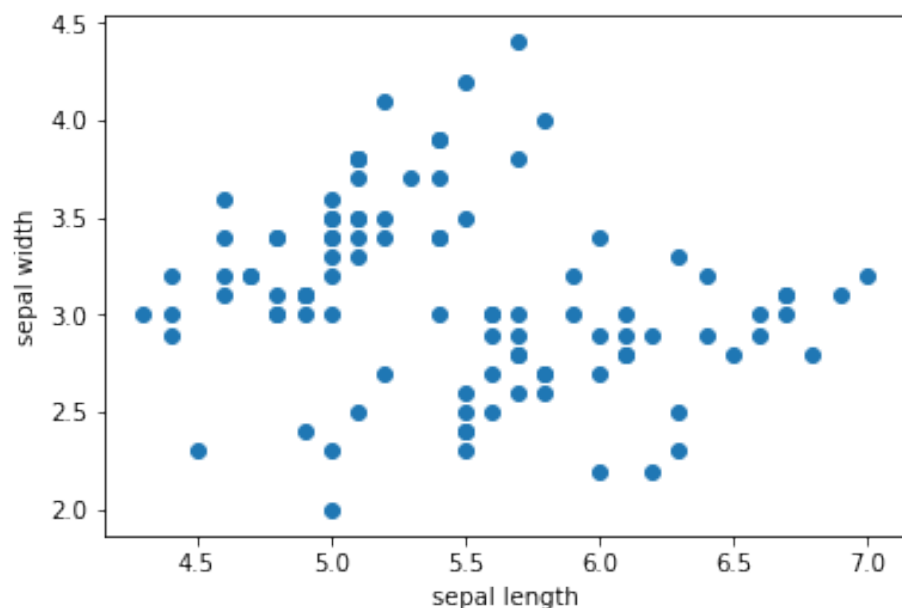
- 给机器的训练数据没有任何“标记”或者“答案”



### 3 机器学习的学习方式

#### ■ 无监督学习:

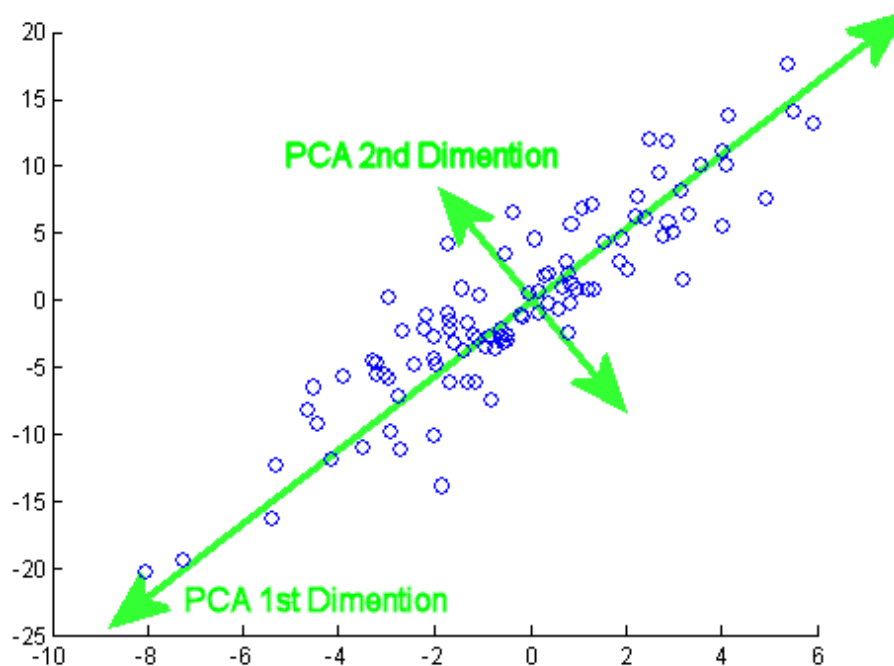
- 聚类分析: 让机器通过数据的特征自动去判断, 哪些数据比较像, 归到一类



### 3 机器学习的学习方式

#### ■ 无监督学习：

- 数据降维：机器学习领域中的降维就是指采用某种映射方法，将原高维空间中的数据点映射到低维度的空间中。
- 主成分分析算法（PCA）
- PCA是一种常见的数据分析方式，常用于高维数据的降维，可用于提取数据的主要特征分量。

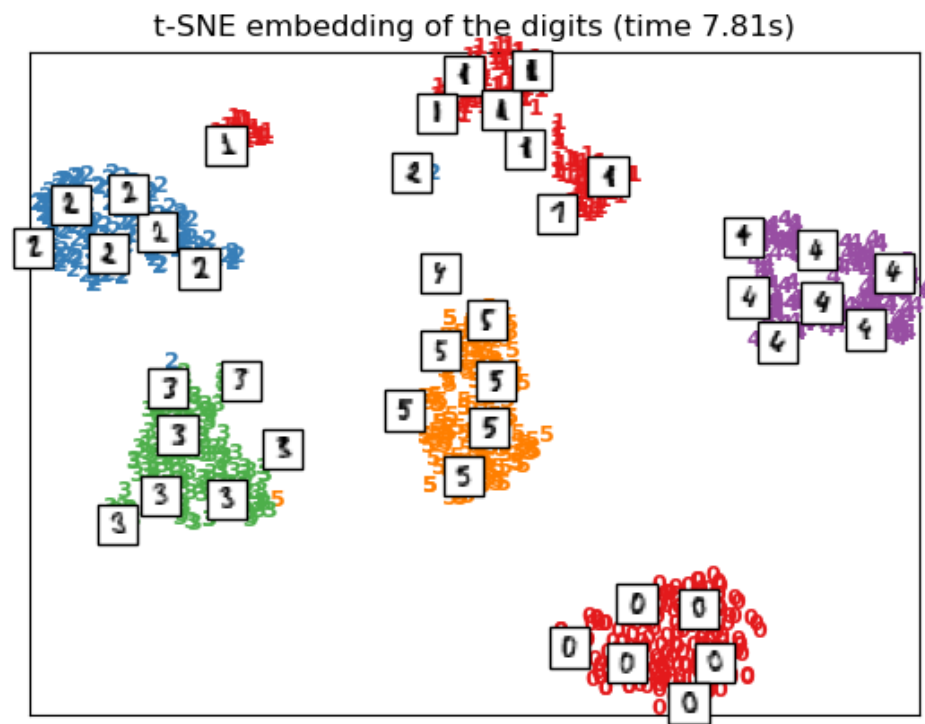




### 3 机器学习的学习方式

#### ■ 无监督学习：

- 数据降维：机器学习领域中的降维就是指采用某种映射方法，将原高维空间中的数据点映射到低维度的空间中。
- t-分布邻域嵌入算法（T-SNE）
- T-SNE是一种降维技术，用于在二维或三维的低维空间中表示高维数据集，从而使其可视化。



### 3 机器学习的学习方式

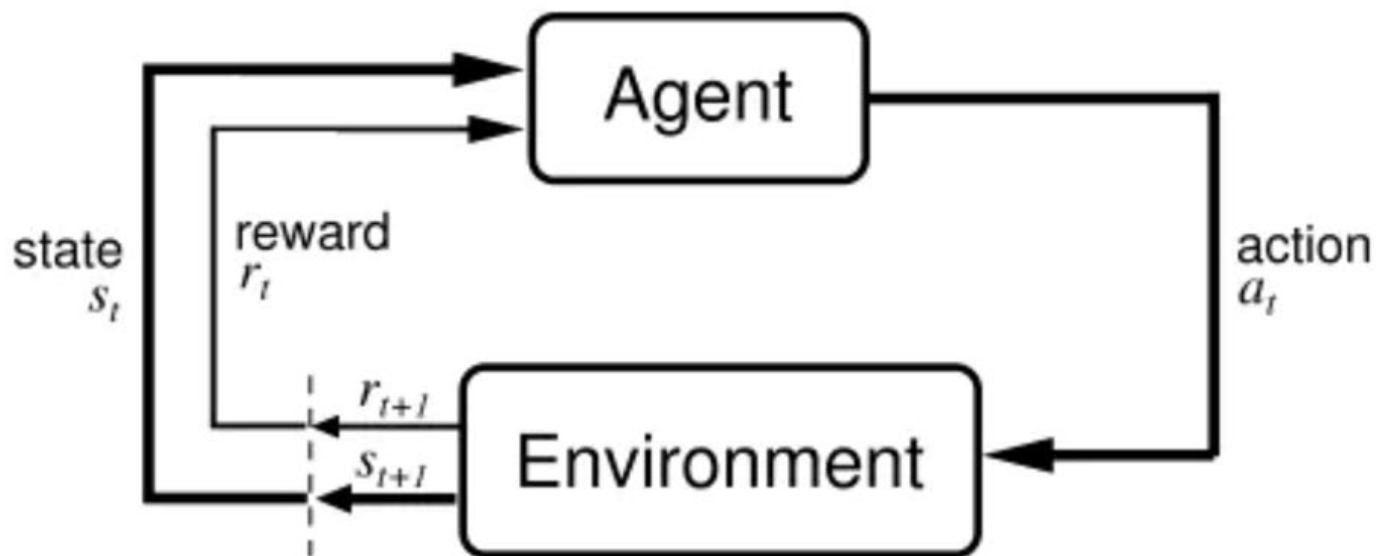
#### ■ 半监督学习：

- 半监督学习是监督学习与无监督学习相结合的一种学习方法。半监督学习中一部分数据有“标记”或者“答案”，另一部分数据没有标记。
- 在很多实际问题中，只有少量的带有标记的数据，因为对数据进行标记的代价有时很高，比如在生物学中，对某种蛋白质的结构分析或者功能鉴定，可能会花上生物学家很多年的工作，而大量的未标记的数据却很容易得到。
- Self-training
- Active-learning

### 3 机器学习的学习方式

#### ■ 强化学习：

- 根据周围环境的情况，采取行动，根据采取行动的结果，学习行动方式。



### 3 机器学习的学习方式

#### ■ 强化学习：

- 根据周围环境的情况，采取行动，根据采取行动的结果，学习行动方式。



### 3 机器学习的学习方式

- 按照学习方式的不同，机器学习可分为很多的类型：
  - 监督学习
  - 无监督学习
  - 强化学习
  - 半监督学习

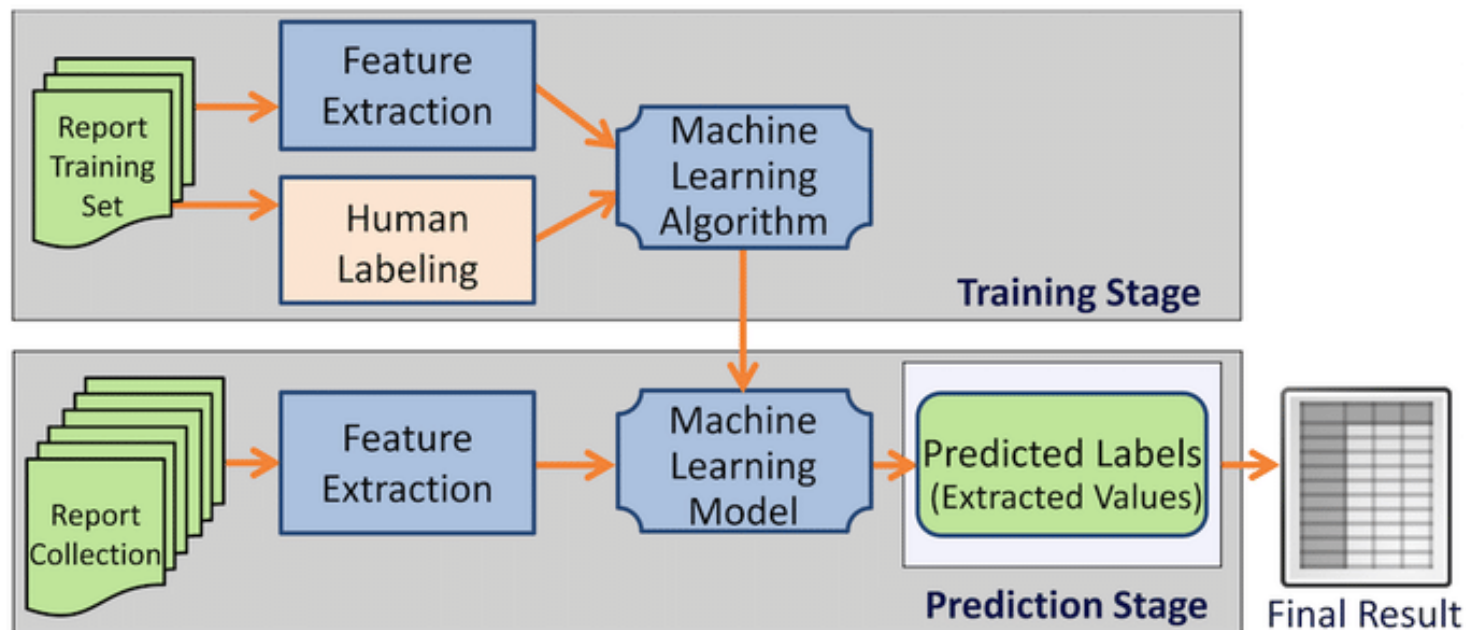
### 3 机器学习的学习方式

- 机器学习还可以分为以下的类型：
  - 在线学习 (online-learning)
  - 批量学习 (batch-learning) , 也可成为离线学习。

### 3 机器学习的学习方式

#### ■ 批量学习：

- 一次性批量输入给学习算法，可以被形象的称为填鸭式学习。



### 3 机器学习的学习方式

#### ■ 批量学习：

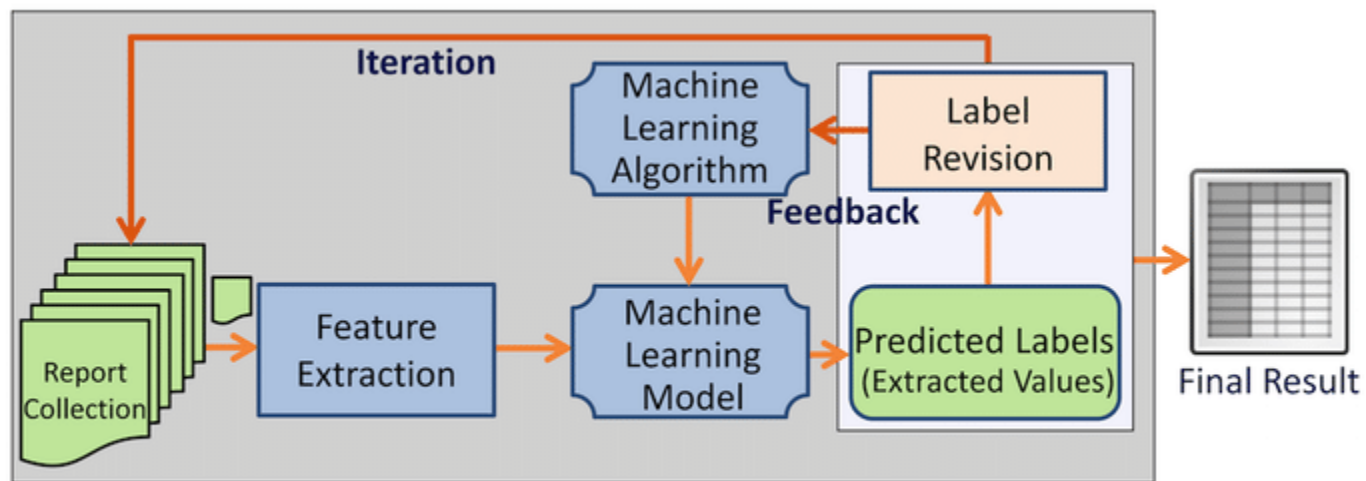
- 优点：简单
- 问题：如何适应环境变化？
- 解决方案：定时重新批量学习
- 缺点：每次重新批量学习，运算量巨大，在某些环境变化非常快的情况下，甚至不可能的。



### 3 机器学习的学习方式

#### ■ 在线学习:

- 按照顺序，循序的学习，不断的去修正模型，进行优化。



(b) Online Machine Learning

# 3 机器学习的学习方式

## ■ 在线学习：

### ● 优点：

- (1) 容易执行
- (2) 对于大规模和困难模式分类问题它提供有效解。
- (3) 随机性使得不容易陷入局部极值点
- (4) 存储量少得多

### ● 缺点：

- 学习速度慢

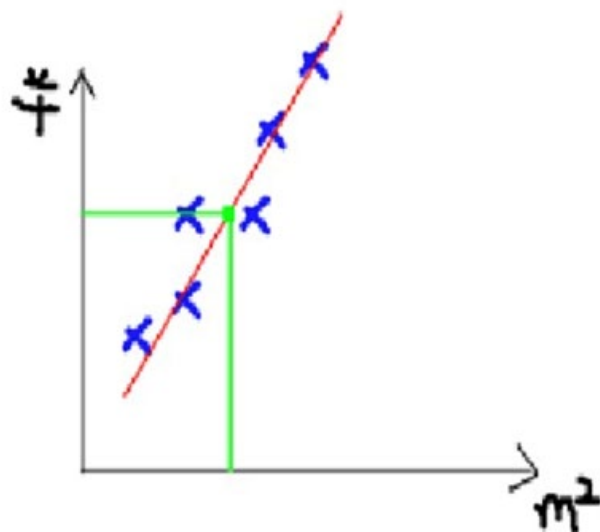
### 3 机器学习的学习方式

- 机器学习还可以分为以下的类型：
  - 参数学习
  - 非参数学习

### 3 机器学习的学习方式

#### ■ 参数学习：

- 参数学习算法是一类有固定数目参数，以用来进行数据拟合的算法。线性回归即使参数学习算法的一个例子。
- 一旦学习到了参数，就不再需要原有数据。

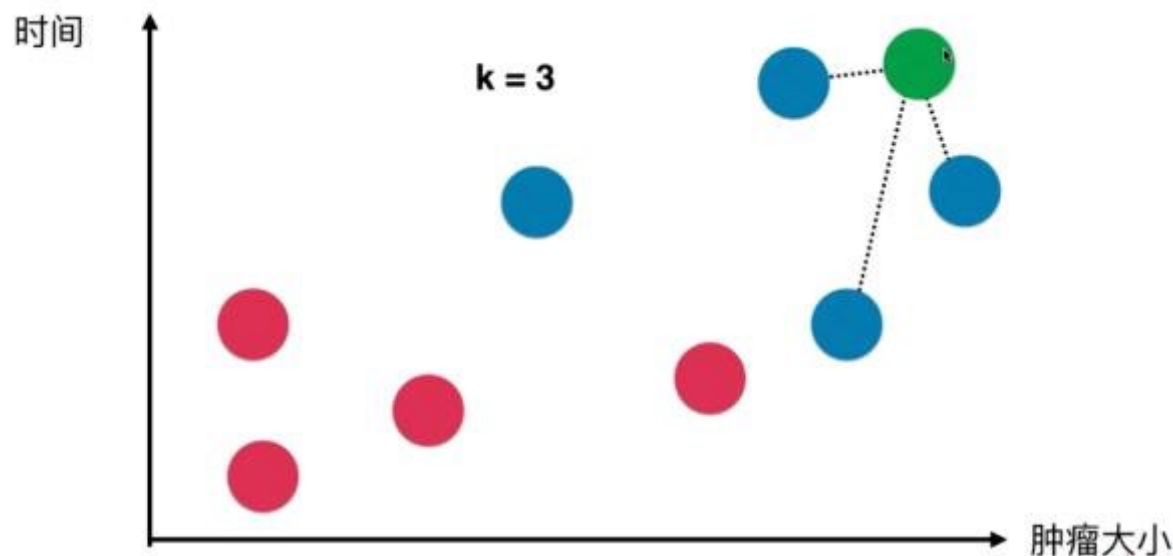


●  $y=5x+1$

### 3 机器学习的学习方式

#### ■ 非参数学习：

- 在预测新样本值时候每次都会依赖数据集得到新的参数值，也就是说每次预测新样本都会依赖训练数据集合，所以每次得到的参数值是不确定的。
- 代表性算法为K近邻算法。



问题？



暨南大學  
JINAN UNIVERSITY