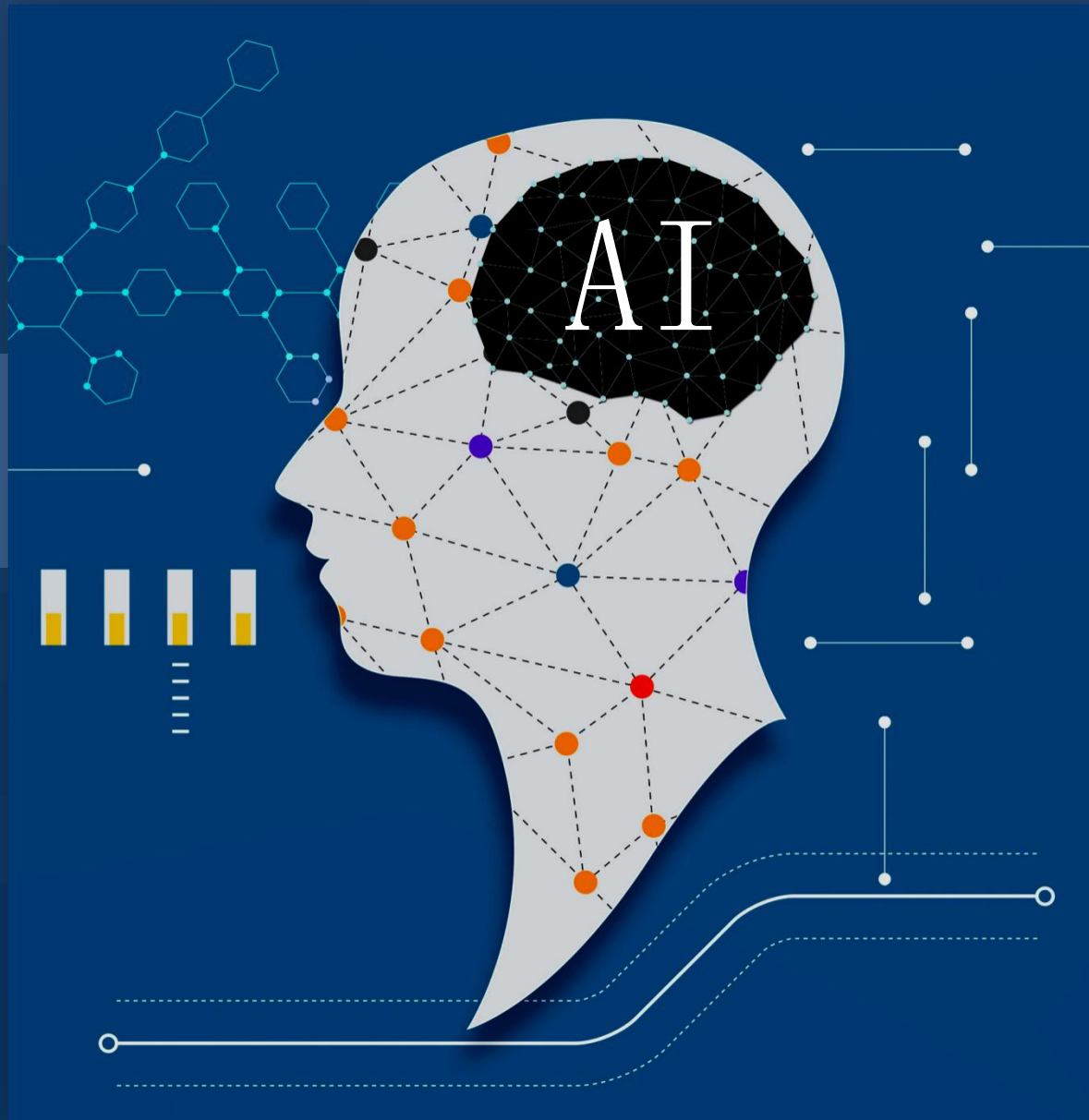


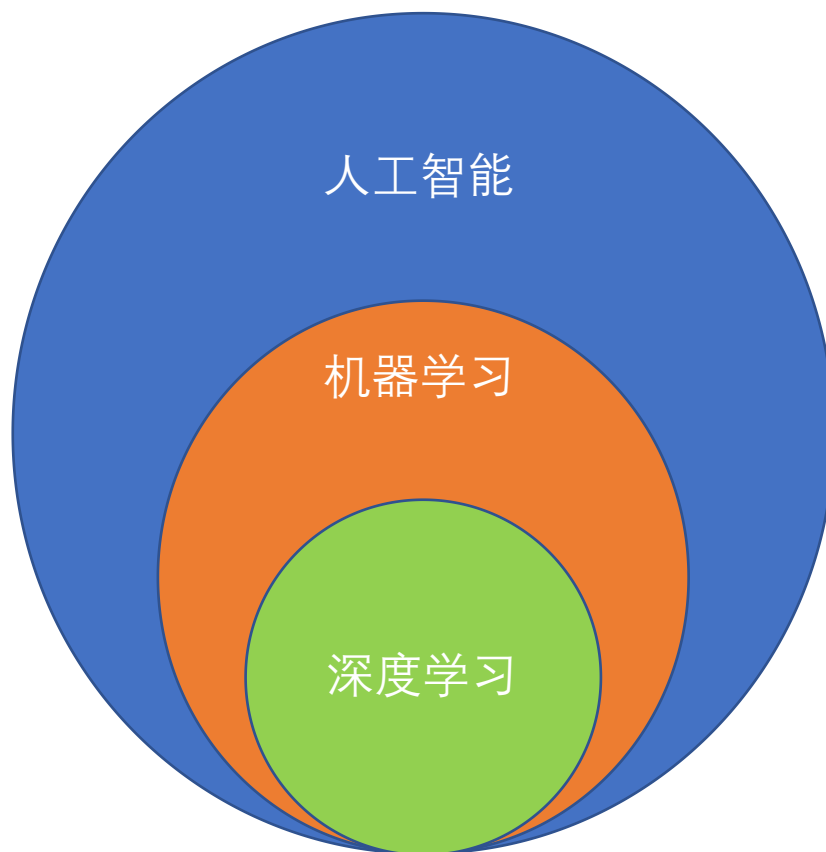


机器学习入门



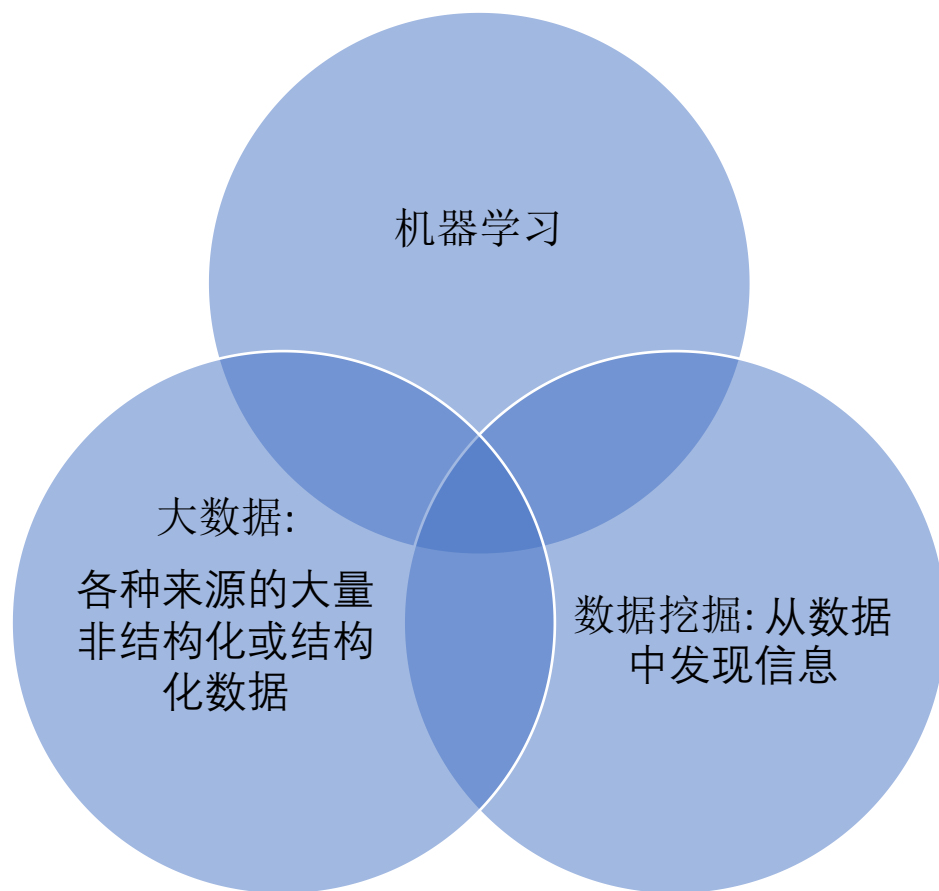
什么是机器学习

- 是人工智能(AI)的一部分, 研究如何让计算机从数据学习某种规律



机器学习 V. S. 数据挖掘 V. S. 大数据

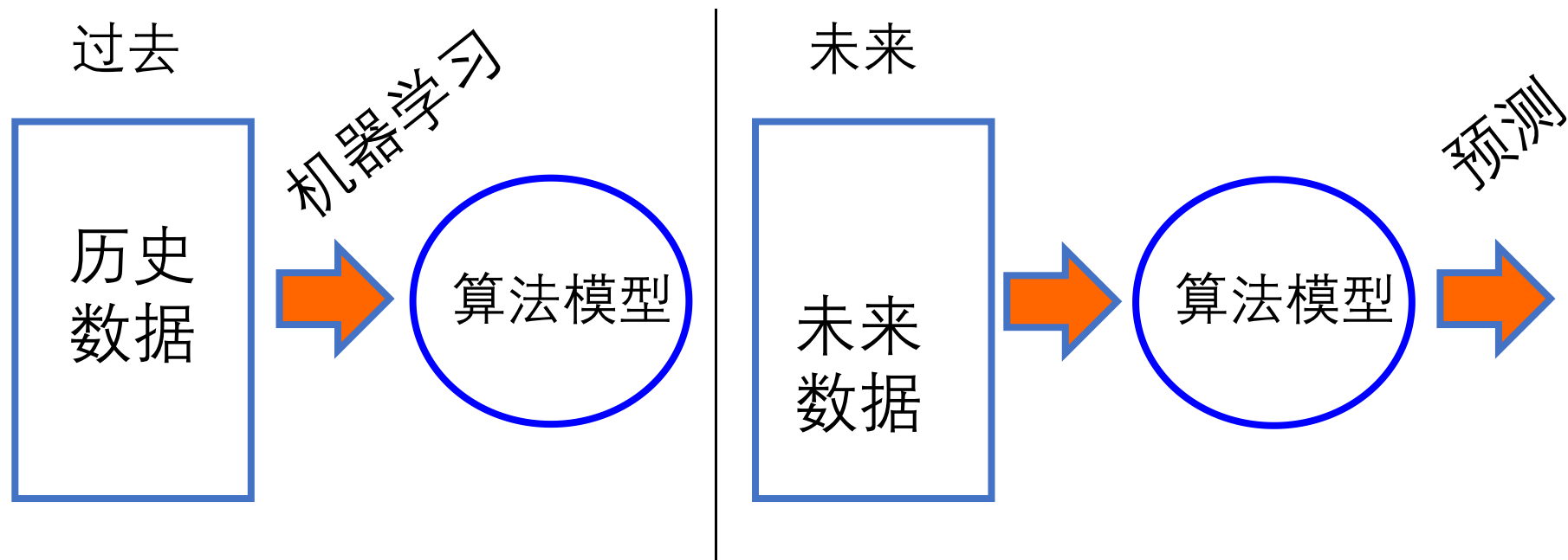
- 是人工智能(AI)的一部分, 研究如何让计算机从数据学习某种规律



什么是机器学习

- 通过计算机程序根据数据去优化某一个评价指标
- 自动的从数据发现规律, 使用这些规律做出预测
- 根据过去预测未来

什么是机器学习



机器学习的别名

- 数据挖掘: 机器学习应用于“数据库”
- 推理/估计: 统计学
- 模式识别

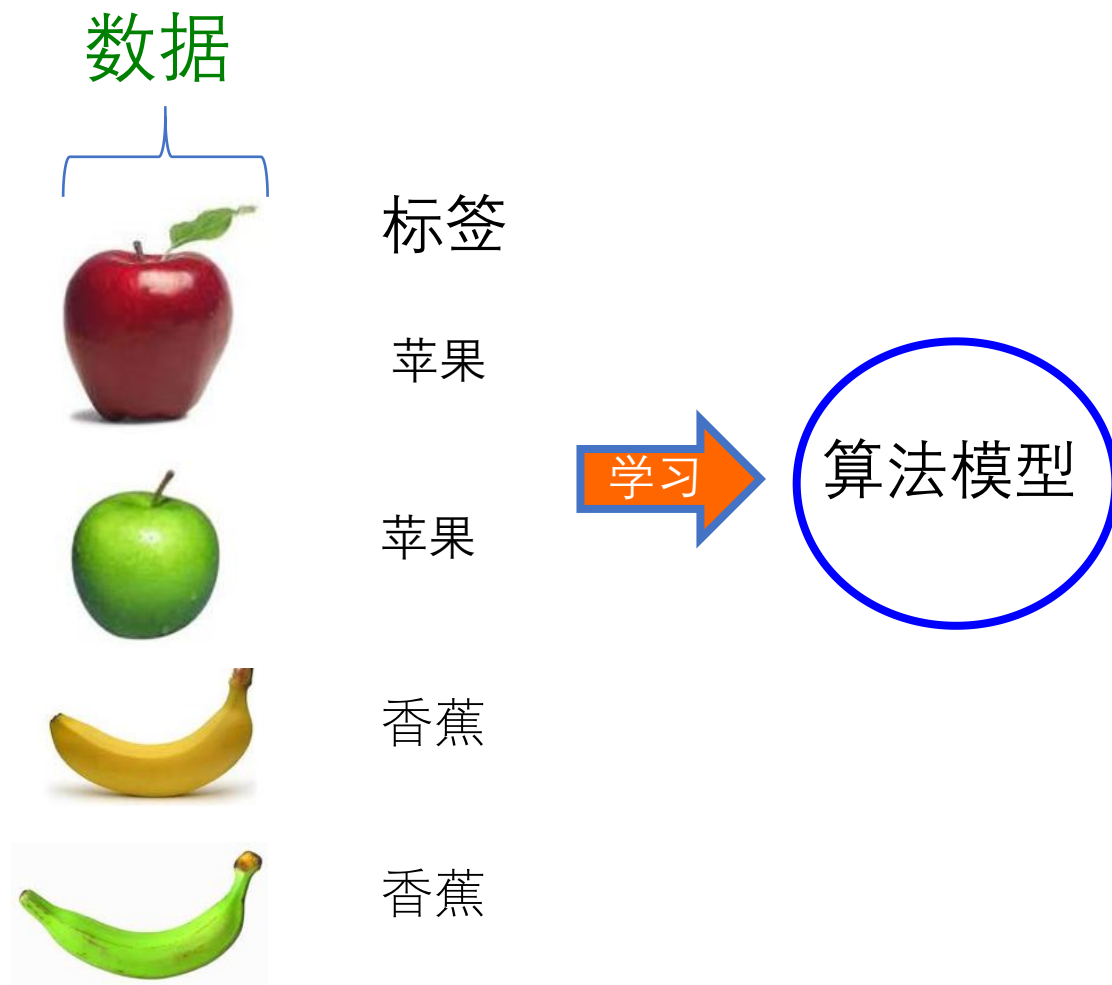
机器学习家族

- 监督式学习
 - 分类
 - 回归
- 非监督式学习
 - 聚类

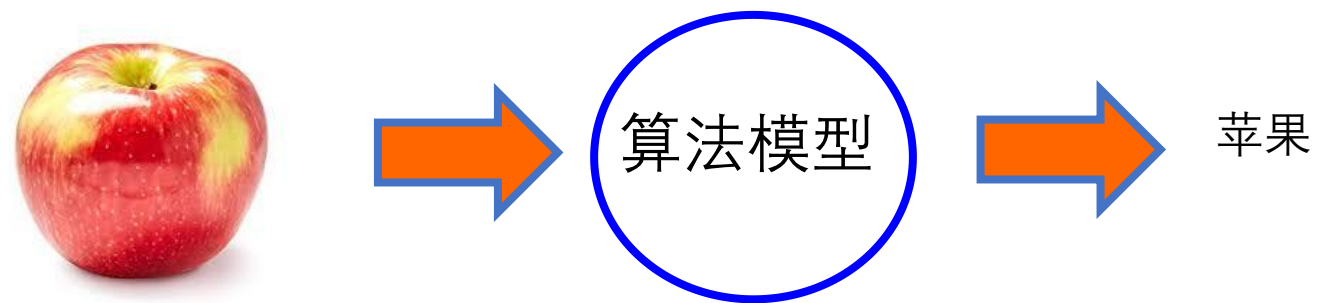
监督式机器学习（分类）



监督式机器学习（分类）



监督式机器学习（分类）



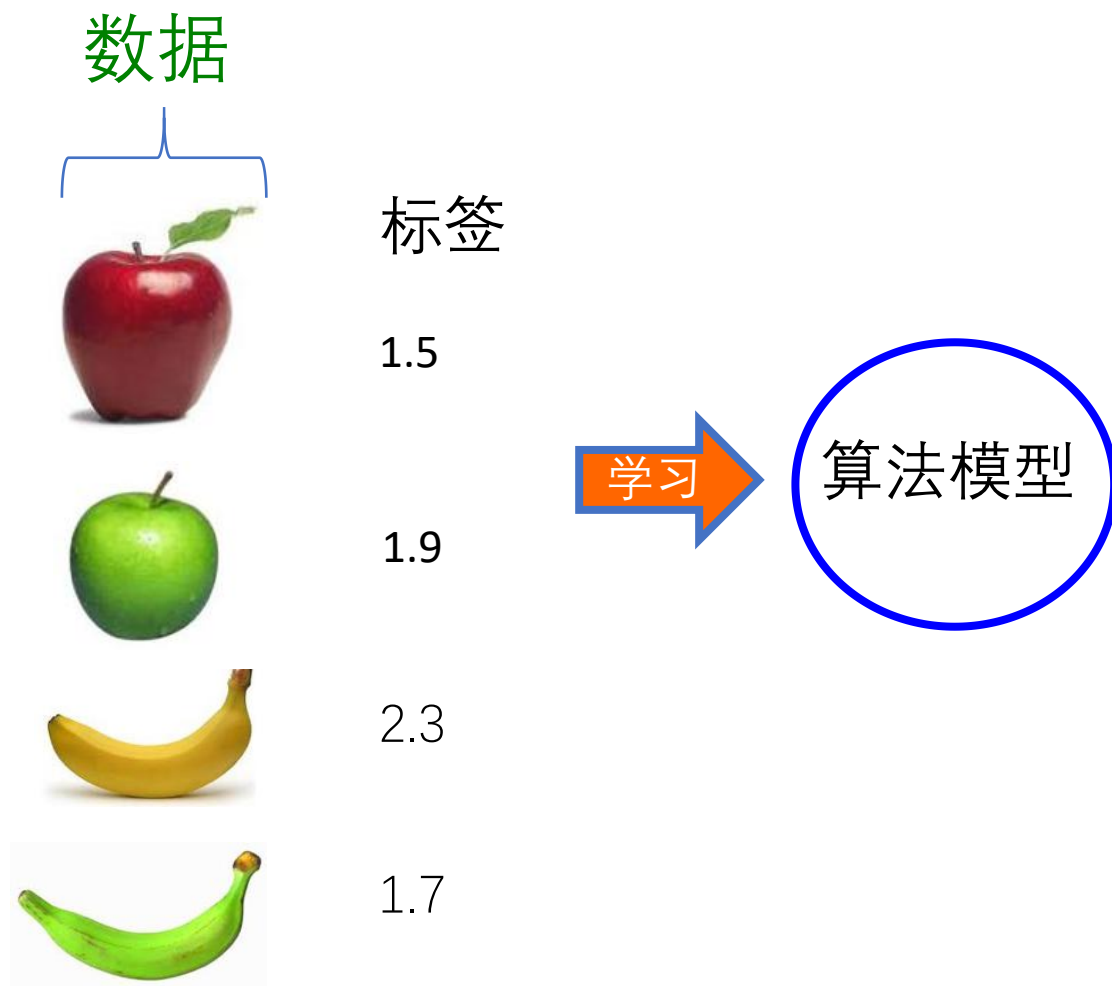
监督式学习 (分类) 实例

- 垃圾邮件/短信检测
- 自动车牌号识别
- 人脸识别
- 手写字符识别
- 语音识别
- 医疗图片的病症诊断
-

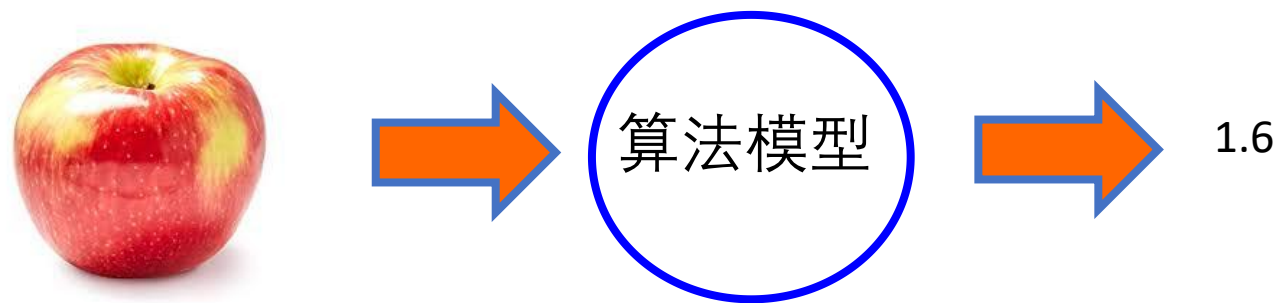
监督式机器学习（回归）



监督式机器学习（回归）



监督式机器学习（回归）

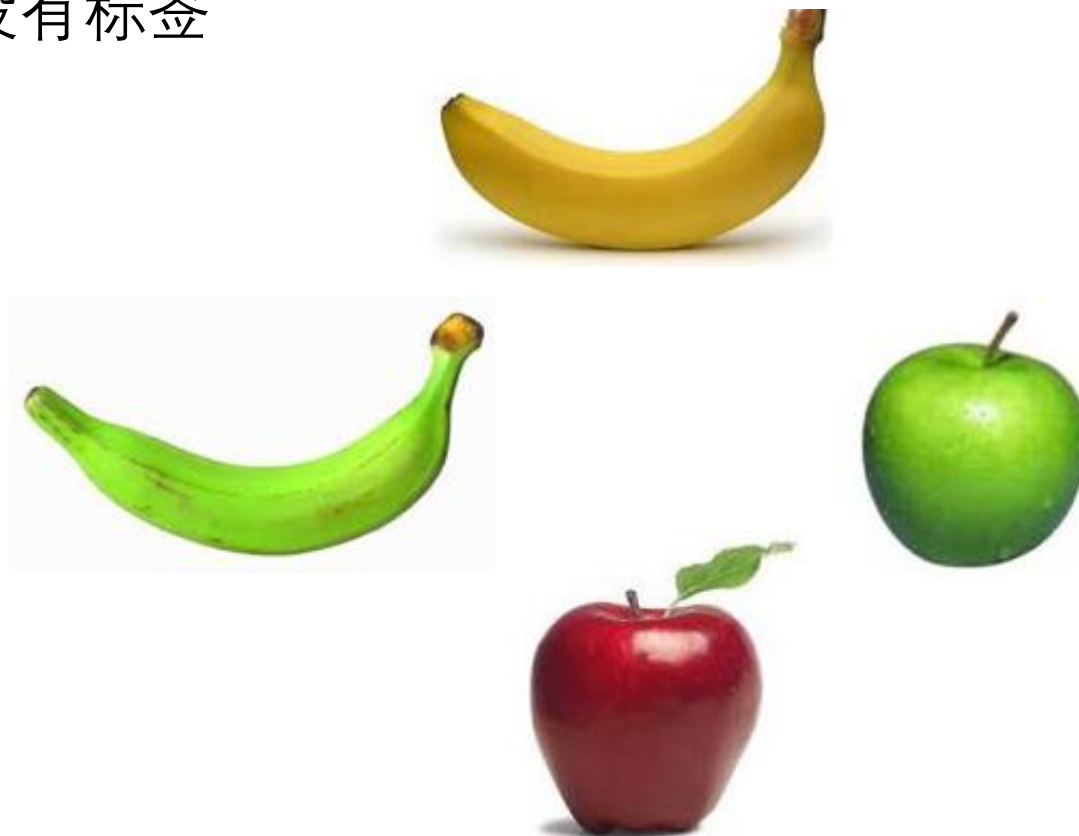


监督式机器学习（回归）实例

- 自动为二手车估价
- 预测股票价格
- 预测未来气温
- 自动驾驶
-

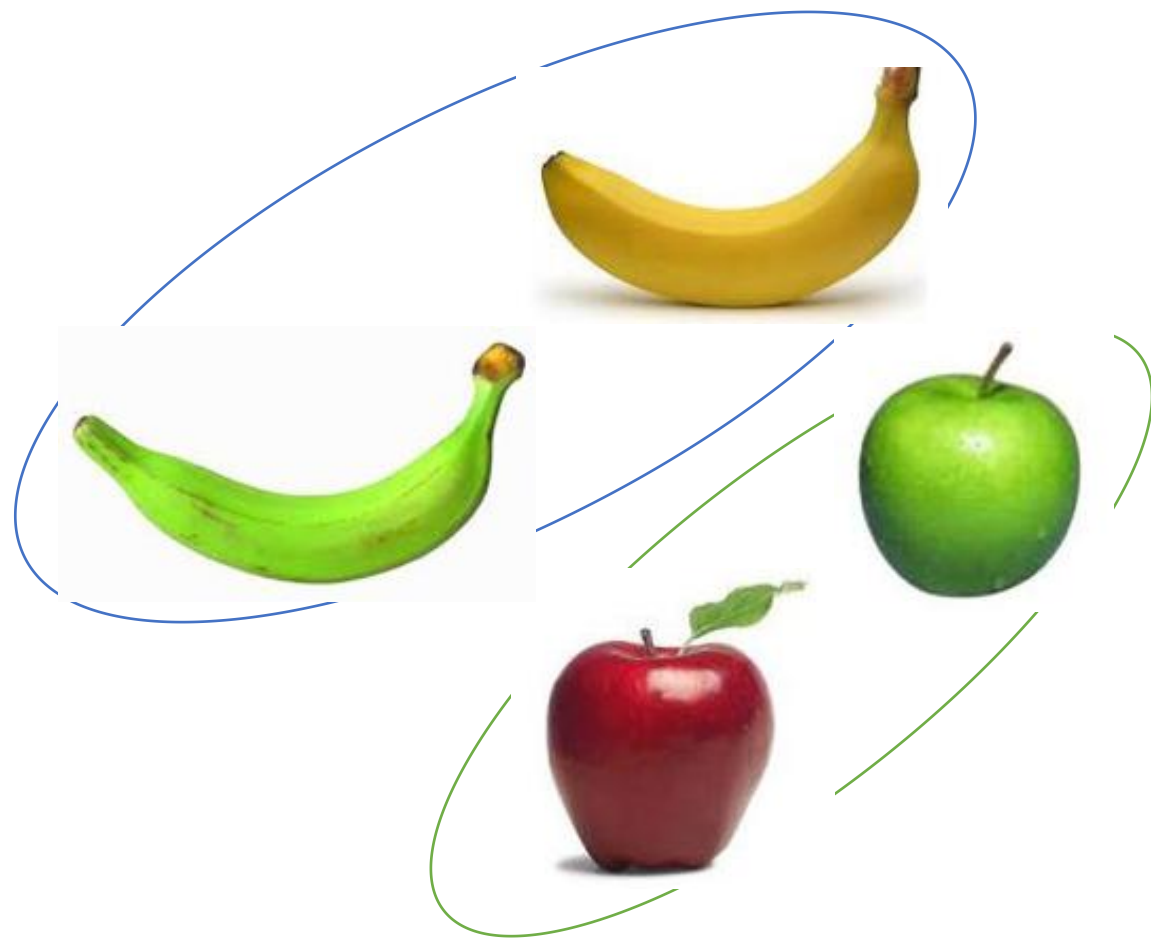
非监督式机器学习 (聚类)

- 只有数据, 没有标签



非监督式机器学习(聚类)

- 把对象分成不同的子集
(subset), 使得属于同一个子集中的成员对象都有相似的一些属性



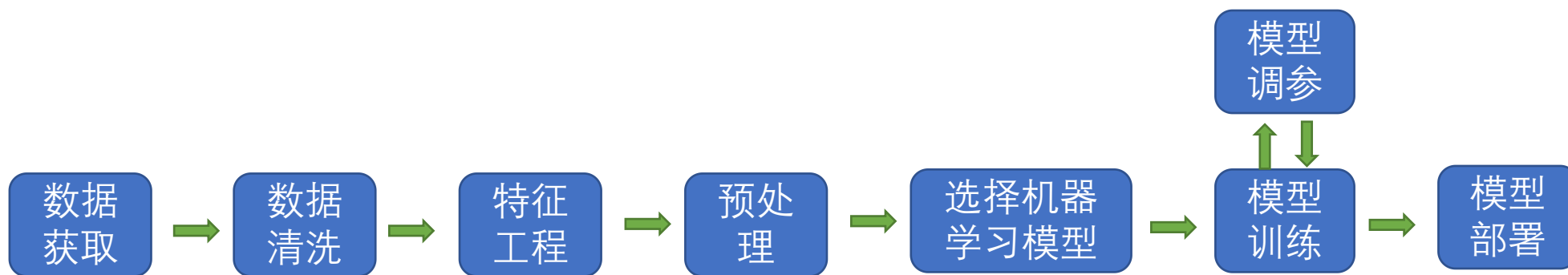
非监督式机器学习(聚类)

- 应用领域
 - 客户分类(市场研究)
 - 用户分组(社交网络)
 - 图像分割
 - 推荐系统
 - 消除歧义(自然语言处理)
 -

总结

- 监督式学习
 - 训练数据包含输入和预期的输出
- 非监督式学习
 - 训练数据只有输入, 没有预期的输出

机器学习流程



单项选择题

- 如下场景中, 哪一个不是监督学习的应用:
 - A. 手机使用指纹识别代替密码登录
 - B. 机场使用CT做安检, 检测是否乘客携带非法物品
 - C. 智能音响提供语音下单购物
 - D. 战场使用无人机跟踪敌方目标

判断题

- 机器学习的内容囊括了人工智能和大数据
- 是
- 不是

数据预处理

- 特征提取
- 处理缺失数据
- 数据定标
- 数据转换: One-Hot encoding, One/Two/MultiGram, Bag of words, 取对数

数据预处理之特征提取

- 以基于图像进行行人检测为例, 需要提取图像的梯度直方图

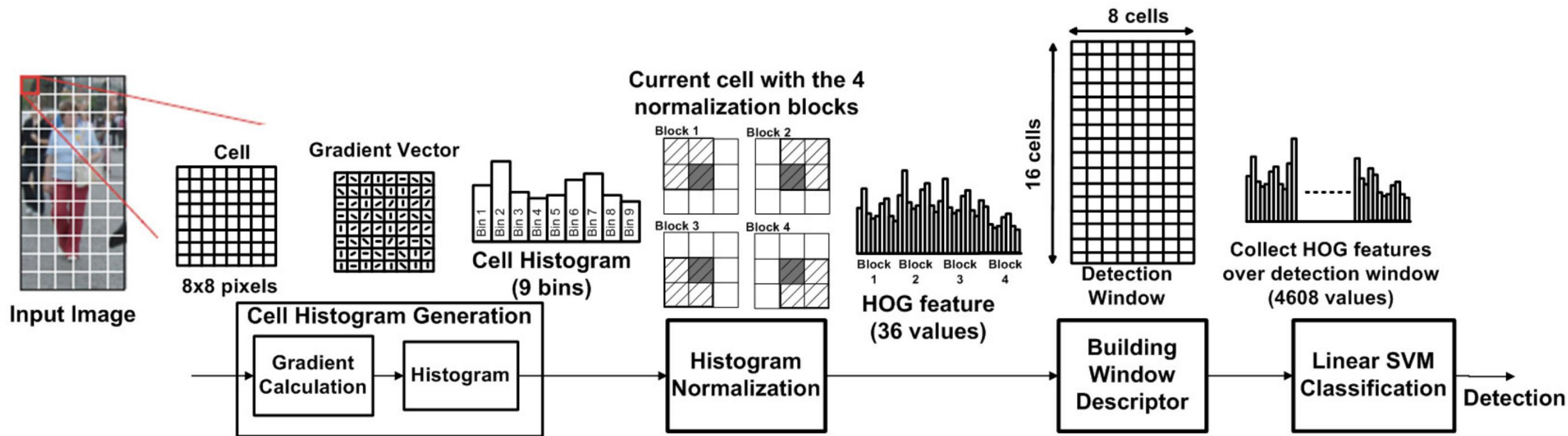


Fig. 2: Object detection algorithm using HOG features.

Reference: https://www.researchgate.net/publication/267868361_Energy-Efficient_HOG-based_Object_Detection_at_1080HD_60_fps_with_Multi-Scale_Support

数据预处理之特征提取


- 以自然语言处理为例, 需要提取文字的n-gram

Natural Language Processing
by National Research University Higher School of Economics

We can count token pairs, triplets, etc.

- Also known as n-grams
 - 1-grams for tokens
 - 2-grams for token pairs
 - ...

good movie	1	1	0	0	...
not a good movie	1	1	0	1	...
did not like	0	0	1	0	...




The diagram illustrates the process of converting natural language text into a numerical representation using n-grams. On the left, a list of phrases is shown: 'good movie', 'not a good movie', and 'did not like'. An arrow points from this list to a matrix on the right. The matrix has five columns: 'good movie', 'movie', 'did not', 'a', and '...'. Each row corresponds to one of the phrases from the list. The values in the matrix are binary (0 or 1), indicating the presence or absence of each token in the phrase. For example, the phrase 'good movie' has a value of 1 for 'good movie' and 'movie', and 0 for 'did not', 'a', and '...'. The phrase 'not a good movie' has a value of 1 for 'good movie', 'movie', and 'a', and 0 for 'did not' and '...'. The phrase 'did not like' has a value of 1 for 'did not', and 0 for 'good movie', 'movie', 'a', and '...'. The National Research University Higher School of Economics logo is visible in the top right corner of the slide.


数据预处理之处理缺失数据

- 以Titanic数据集为例, 部分乘客的年龄, 80%乘客的仓位有缺失

Titanic: Machine Learning from Disaster
Start here! Predict survival on the Titanic and get familiar with ML basics

 Kaggle · 11,408 teams · Ongoing

[Overview](#) [Data](#) [Kernels](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#)



Dealing with Missing values for Age & Cabin

posted in [Titanic: Machine Learning from Disaster](#) 6 years ago

数据预处理之处理缺失数据

- 以Titanic数据集为例, 部分乘客的年龄, 80%乘客的仓位有缺失
 - 处理方式:
 - 1. 使用均值或者中间值(median)代替数值类型(年龄)的缺失数据
 - 2. 使用众数(mode)代替分类数据(性别)的缺失数据
 - 3. 使用聚类的方式, 找到相似的数据点, 使用这些相似数据点的均值等替代缺失数据
 - 4. 如果某一个特征的数据丢失率太高, 直接丢弃这个特征的数据也许更好

数据预处理之数据定标

- Normalization/Min-Max-Scaler (归一化)

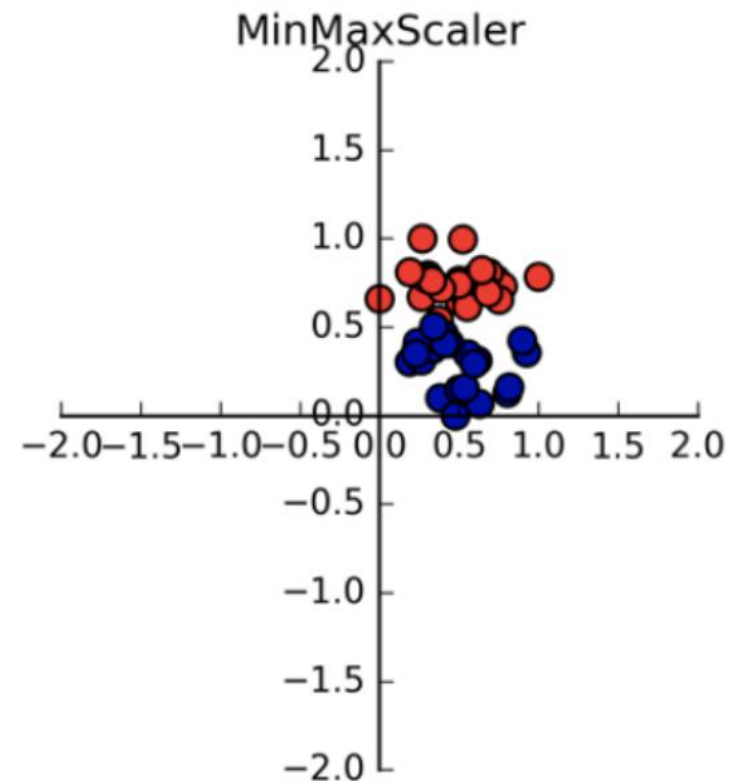
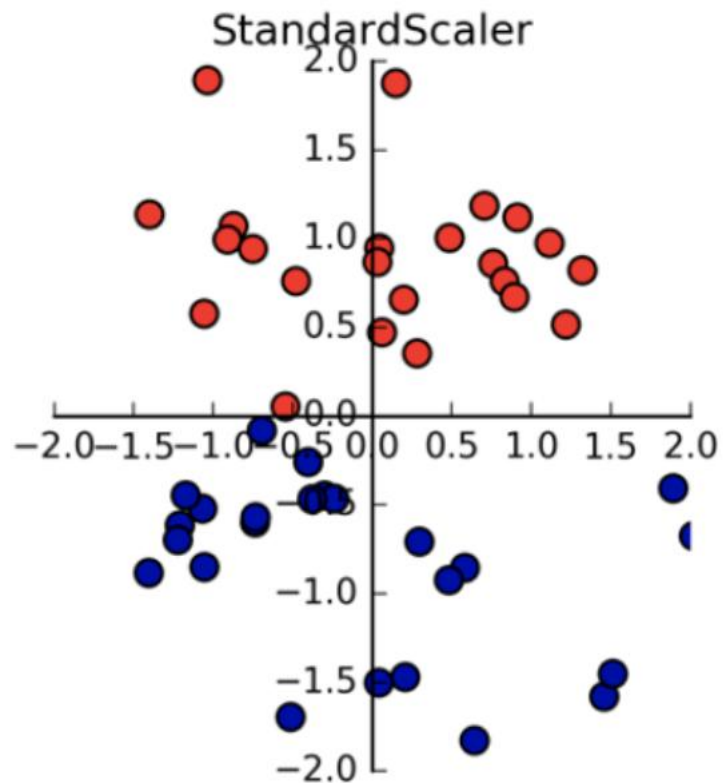
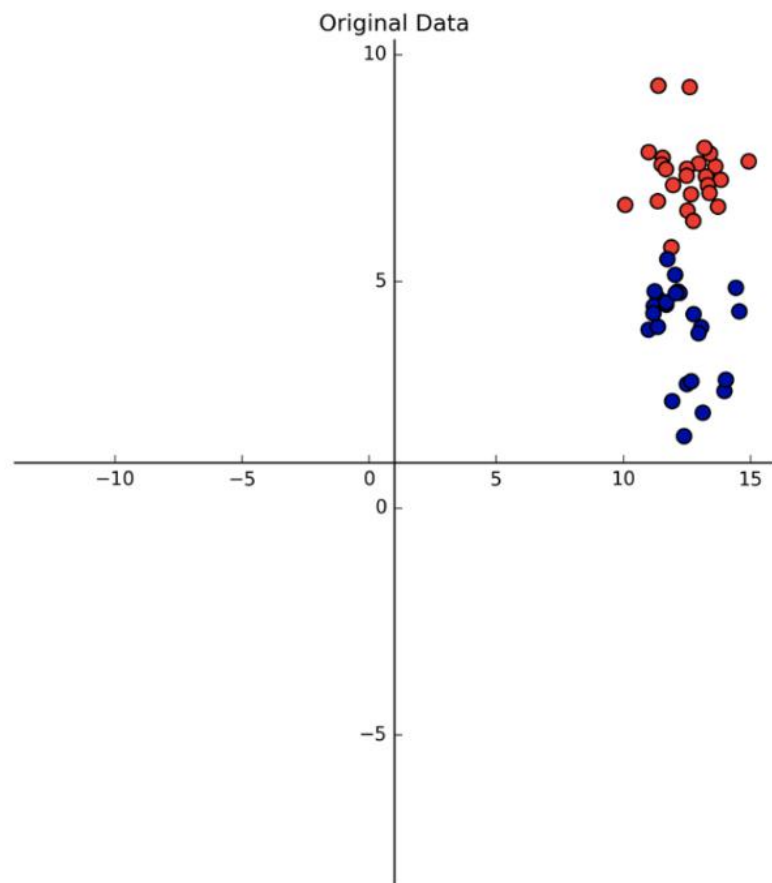
$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- Standardization (标准化)

$$x_{std} = \frac{x - \mu}{\sigma}$$

input	standardized	normalized
0.0	-1.336306	0.0
1.0	-0.801784	0.2
2.0	-0.267261	0.4
3.0	0.267261	0.6
4.0	0.801784	0.8
5.0	1.336306	1.0

数据预处理之数据定标(续, 二维数据)



数据转换：One-Hot encoding

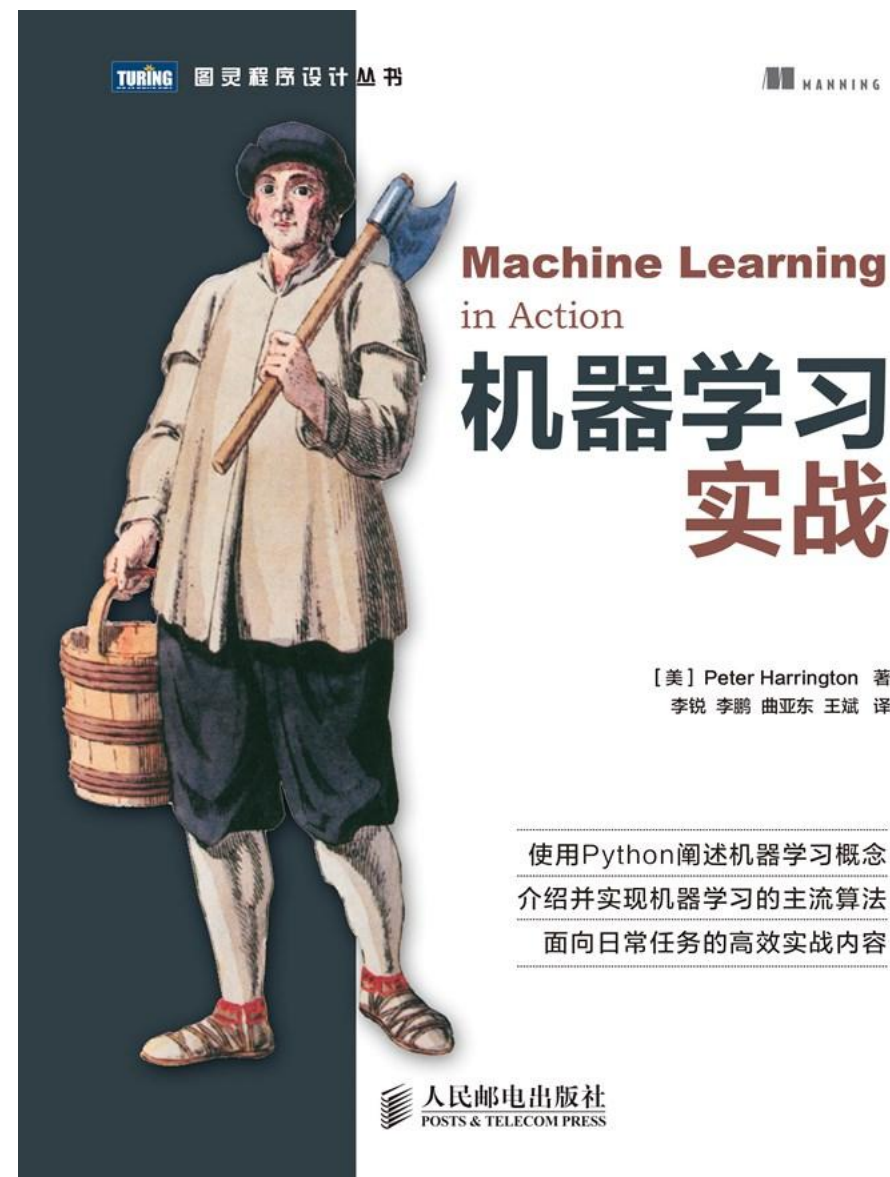
Color		Red	Yellow	Green
Red		1	0	0
Red		1	0	0
Yellow		0	1	0
Green		0	0	1
Yellow		0	0	1

课后练习

- 气温会随着海拔高度的升高而降低, 我们可以通过测量不同海拔高度的气温来预测海拔高度和气温的关系.
- 我们假设海拔高度和气温的关系可以使用如下公式表达:
- $y(\text{气温}) = a * x(\text{海拔高度}) + b$
- 理论上讲, 确定以上公式 a 和 b 的值只需在两个不同高度测试, 就可以算出来 a 和 b 的值了. 但是由于所有的设备都是有误差的, 而使用更多的高度测试的值可以使得预测的值更加准确.
- 我们提供了在9个不同高度测量的气温值, 请你根据今天学习的线性回归方法预测 a 和 b 的值. 根据这个公式, 我们预测一下在8000米的海拔, 气温会是多少?
- 数据文件请见exercise/height.vs.temperature.csv

推荐教材

- <http://www.ituring.com.cn/book/1021>





贪心科技 让每个人享受个性化教育服务

THANKS

贪心学院讲师：袁源