数据科学与工程导论

Introduction to Data Science and Engineering

明天是星期三，我们一起去打乒乓球吧，下午3点在体育场见面，不见不散。

他总是迟到，明天会准时抵达吗？

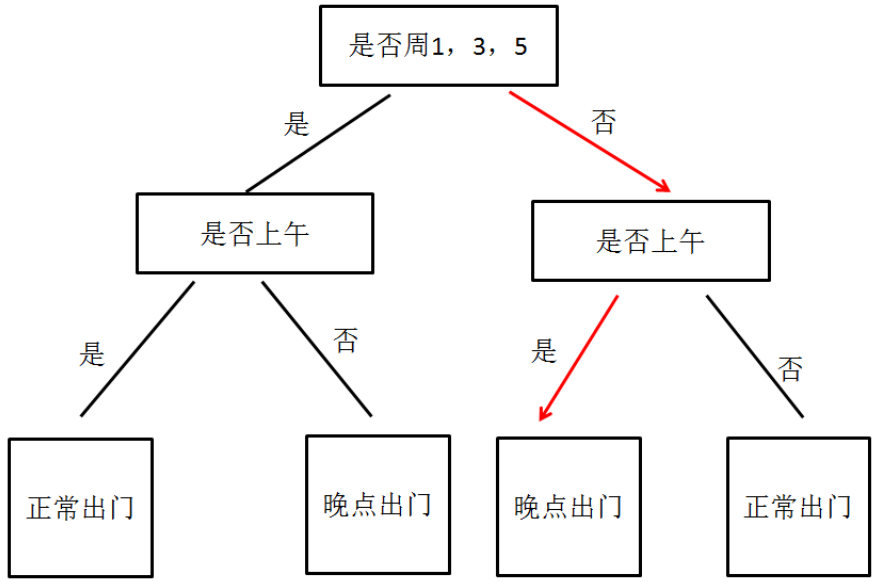
他会迟到吗?

星期三 14：30

- 依据数据同时对数据进行分析后做出判断跟机器学习的思想很相似。女孩对以往的经验（数据）处理后，得到了男孩迟到的频率，进而在当前准备出门时，根据由男孩的迟到频率，做出了当前“暂时不出门”的决策。

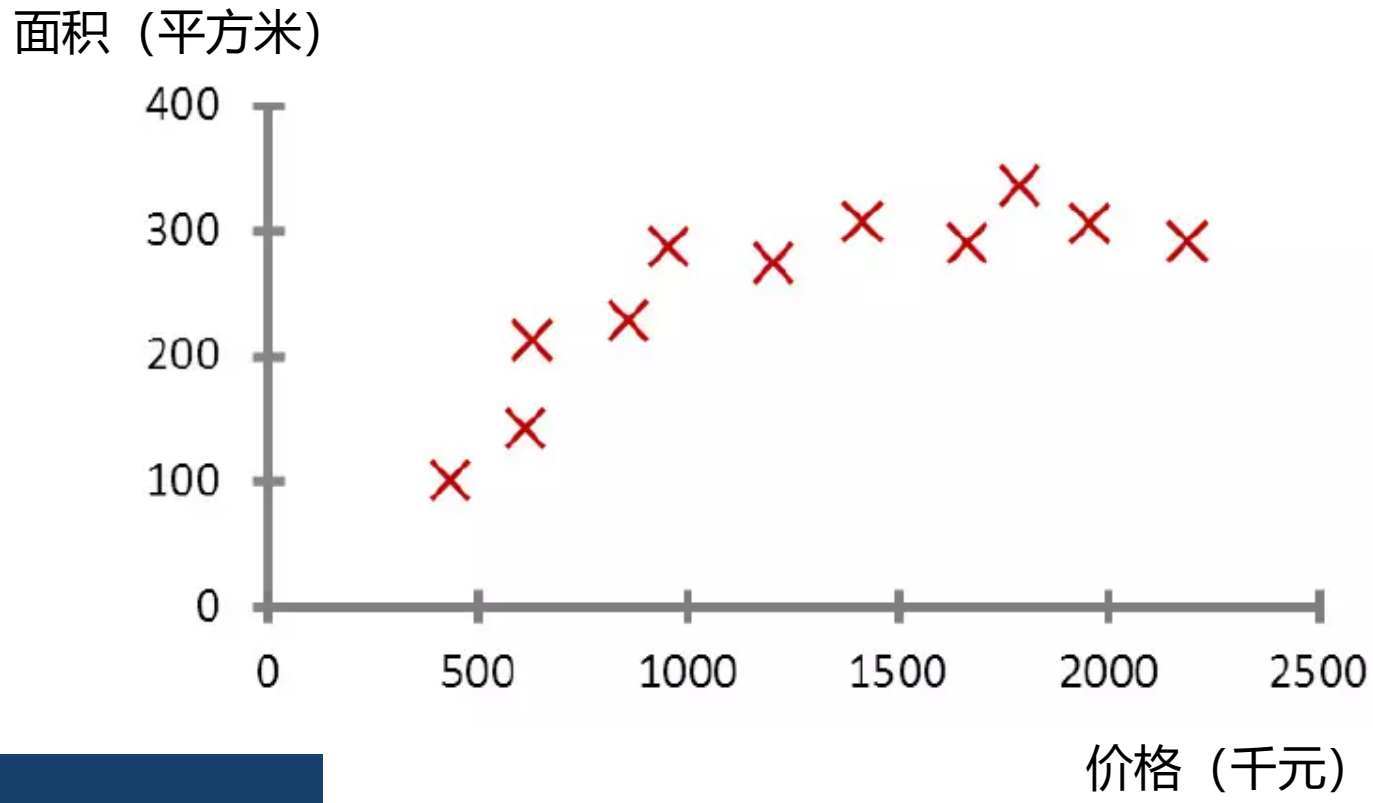
他会迟到吗?

- 刚才的思考过程只考虑“频率”这种属性。我们可以进行更细致的分析。
  - 男孩又一次约女孩周日早上9点去体育馆打羽毛球。女孩根据以往经历（数据）做了分析。
  - 女孩发现男孩很多次迟到发生在周1，3，5的下午和周2，4，6，7的上午。其他情况下，男孩基本不迟到。于是女孩可以根据以往数据建立一个决策树模型，来预测男孩周日早上9点是否迟到，如右图:



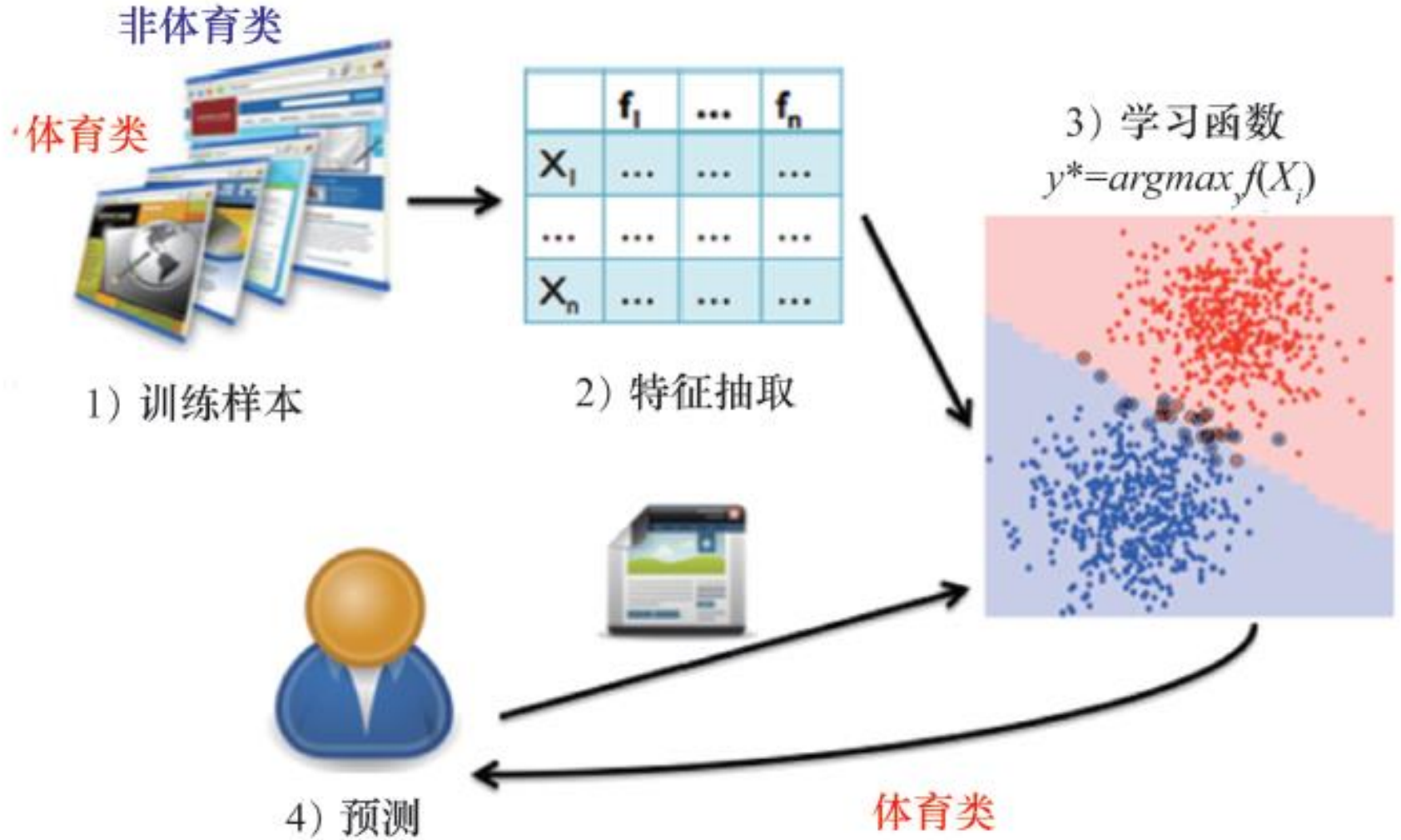
女孩根据这个决策树箭头所示，决定晚点出门

他会迟到吗?

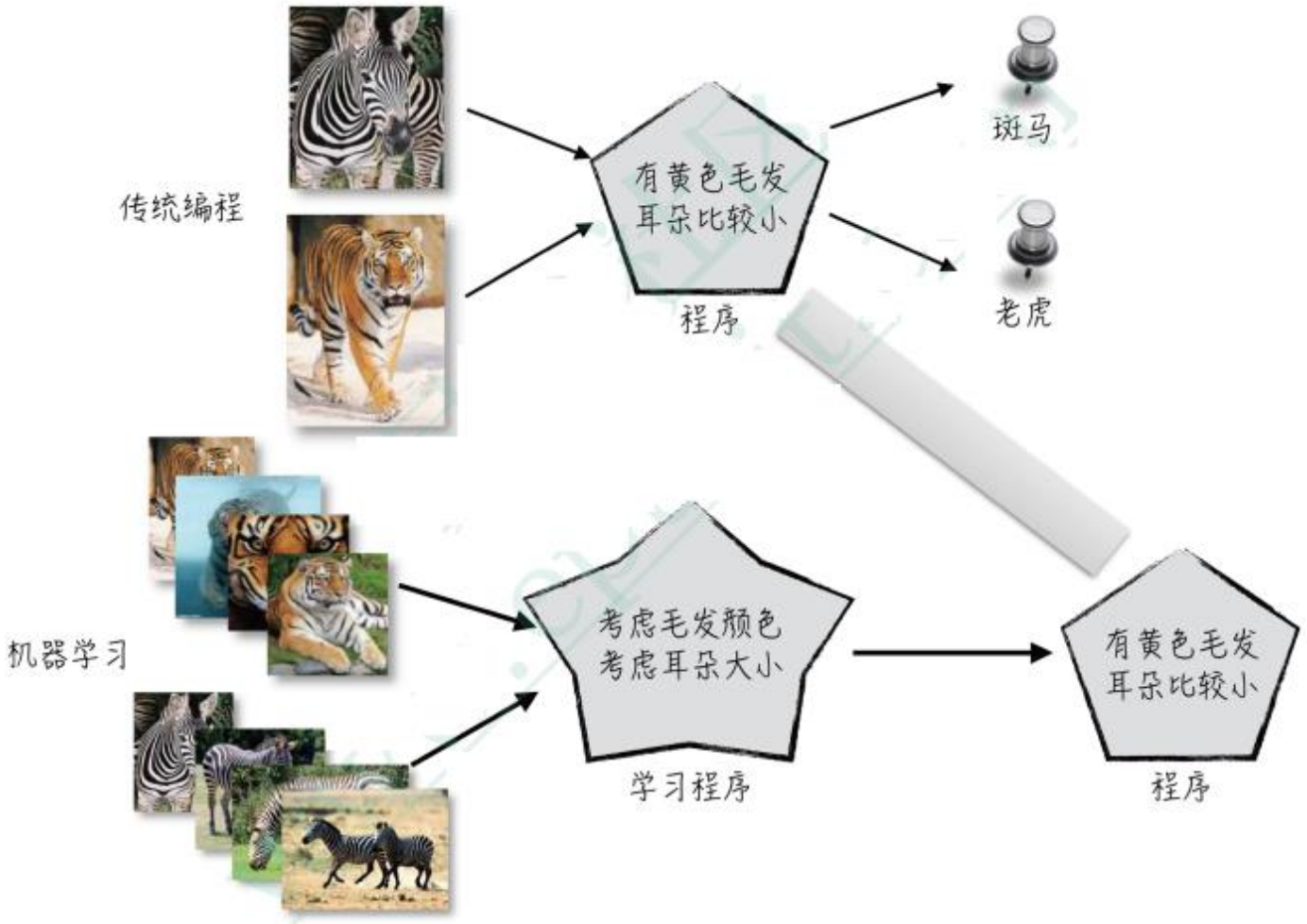


现在有一栋房子需要售卖，
应该给它标上多大的价格？
房子的面积是150平方米
标注价格是70万，80万，还是90万？

房子应该标价多少?

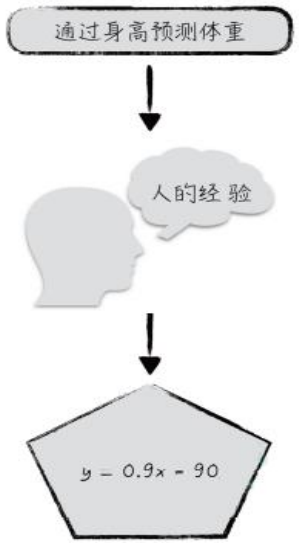


网页分类



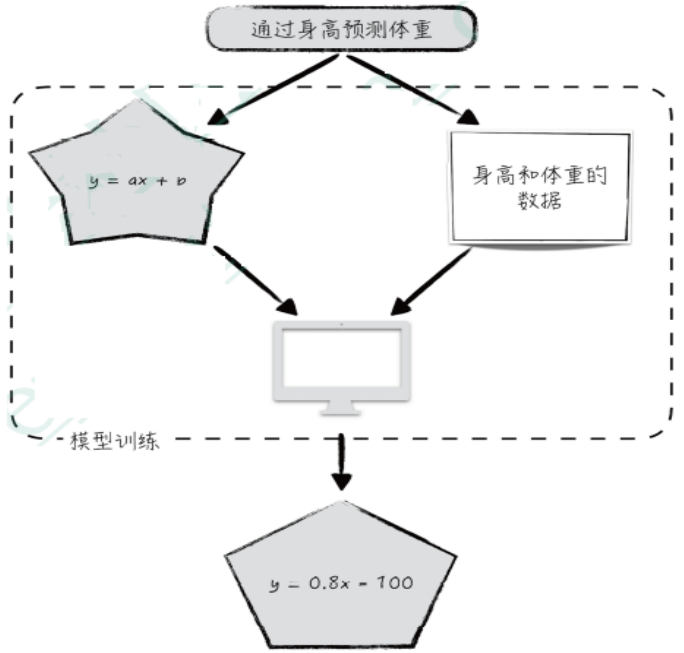
从编程的角度来看，机器学习是一种能自动生成程序的特殊程序。

机器学习方法

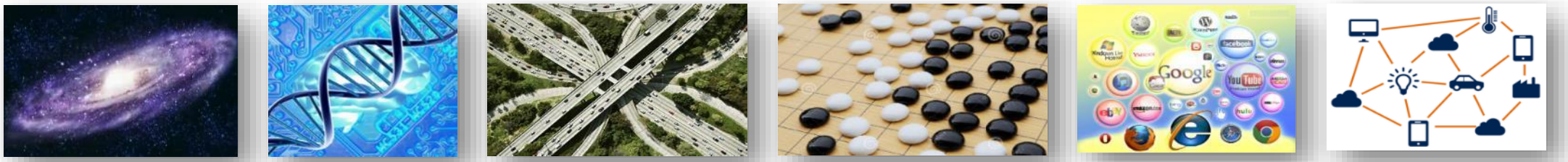


经验驱动的
算法模型

数据驱动的
算法模型



经验 → 算法模型 ← 数据



机器学习：数据驱动的问题求解

机器学习方法

第 11 章
机器学习方法

1 机器学习的发展历史

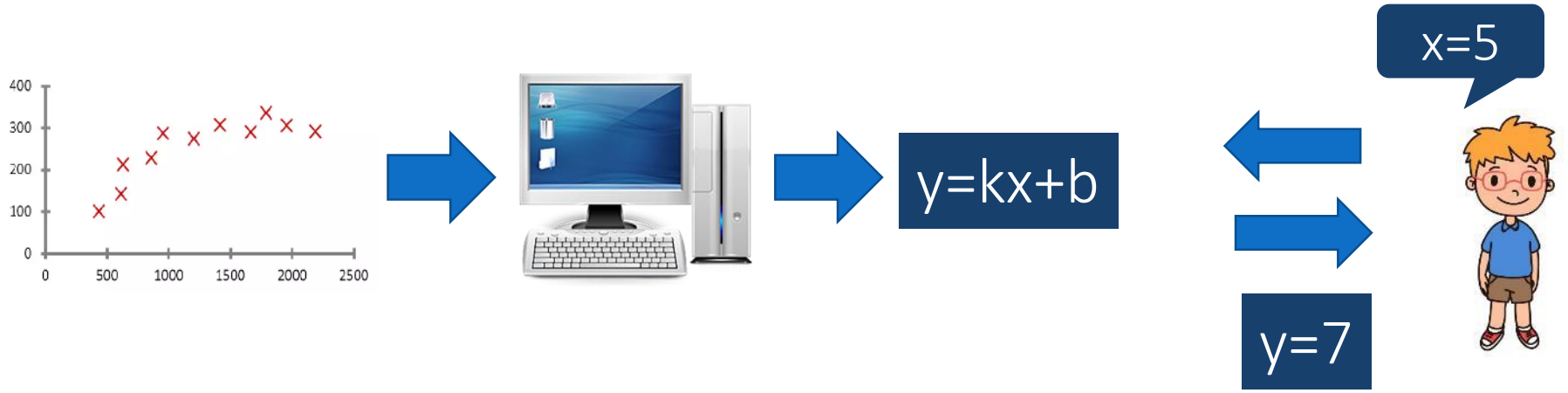
2 机器学习的方法

3 机器学习最新发展

- 学习是人类具有的一种重要智能行为，但究竟什么是学习，长期以来却众说纷纭。
  - 社会学家、逻辑学家和心理学家都各有其不同的看法。
  - 至今，还没有统一的“机器学习”定义，而且也很难给出一个公认的和准确的定义。
- 机器学习是研究如何使用机器来模拟人类学习活动的一门学科。
- 稍为严格的提法是：机器学习是一门研究机器获取新知识和新技能，并识别现有知识的学问。

什么是机器学习

- 从广义上来说，机器学习是一种能够赋予机器学习的能力以此让它完成直接编程无法完成的功能的方法。
- 但从实践的意义上来说，机器学习是一种通过利用数据，训练出模型，然后使用模型预测的一种方法。

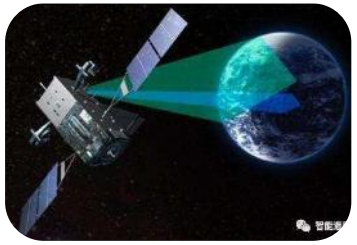


什么是机器学习

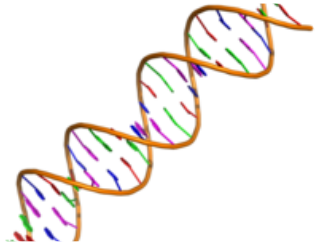
工业过程控制

信息安全

机器人

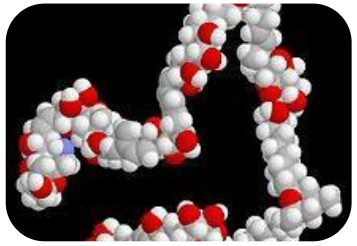


遥感信息处理



生物信息学

计算金融学



分子生物学

行星地质学

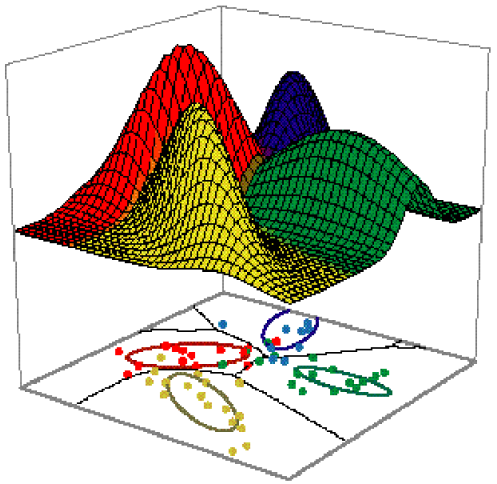
美国JPL实验室的科学家在《Science》（2001年9月）上撰文指出：

- 机器学习对科学研究的整个过程正起到越来越大的支持作用，……，该领域在今后的若干年内将取得稳定而快速的发展

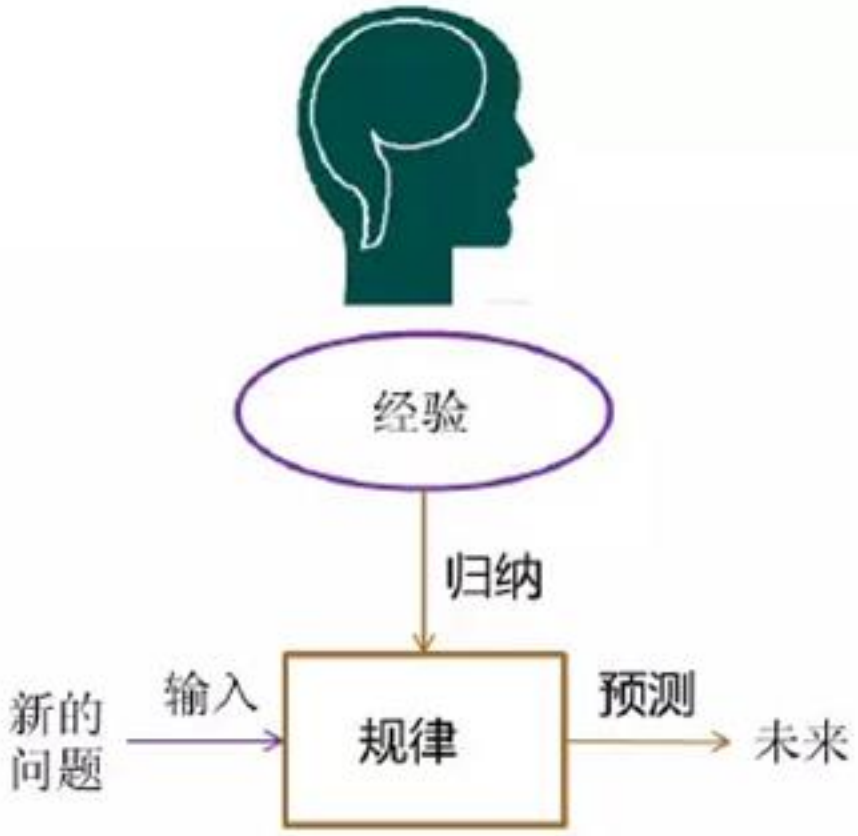
机器学习的重要性

- We believe machine learning will lead to appropriate, partial automation of every element of scientific method, from hypothesis generation to model construction to decisive experimentation. Thus, machine learning has the potential to amplify every aspect of a working scientist's progress to understanding. It will also, for better or worse, endow intelligent computer systems with some of the general analytic power of scientific thinking.

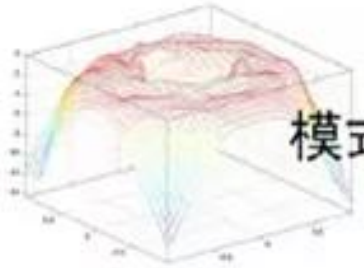
——*Science, 14 September, 2001*



机器学习的重要性



机器学习与人类思考



模式识别

计算机视觉

数据挖掘

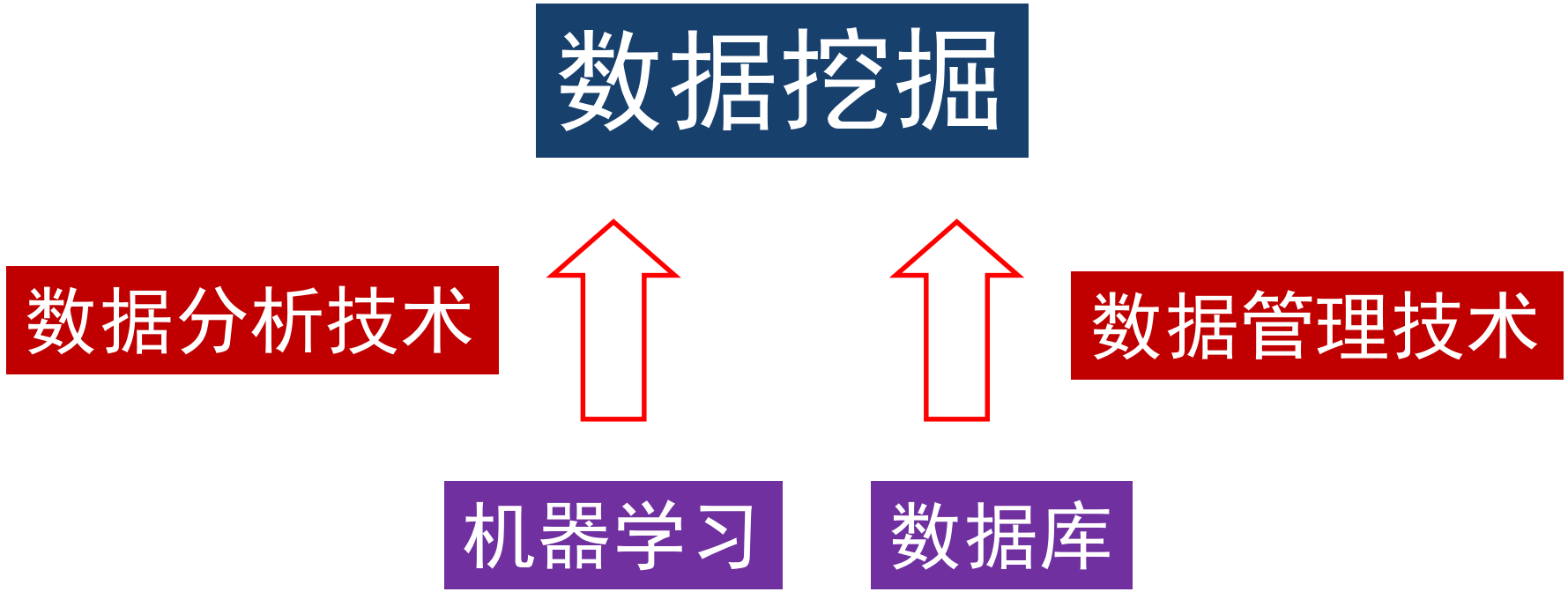
机器学习

语音识别

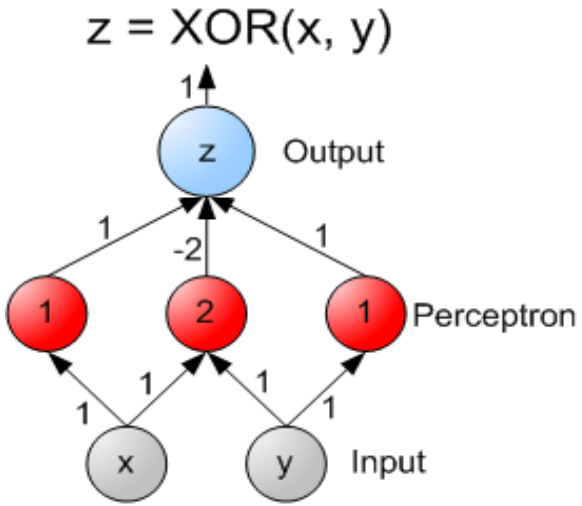
统计学习

自然语言处理

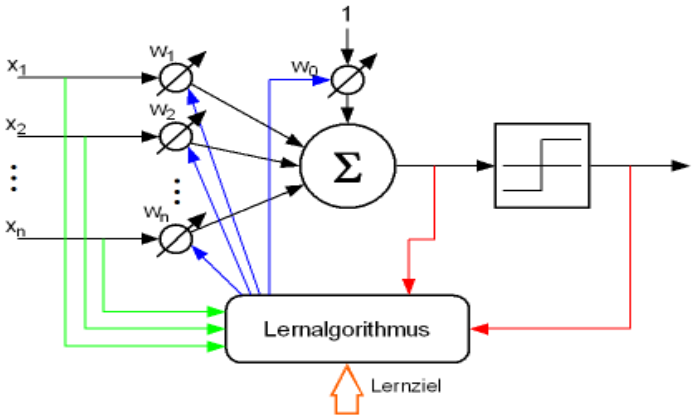
机器学习与相关学科



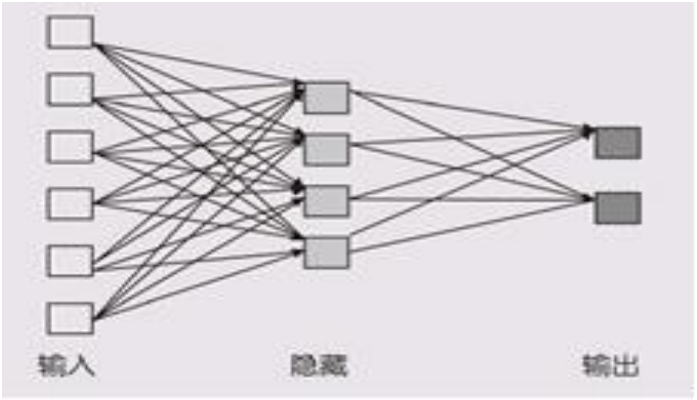
机器学习与相关学科



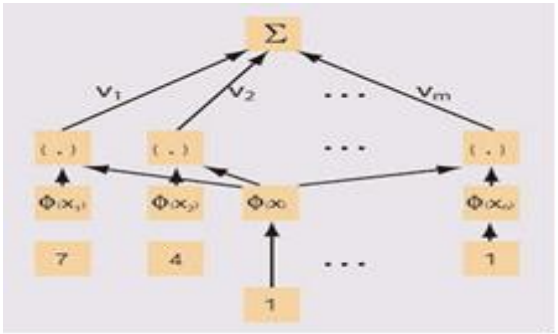
感知机



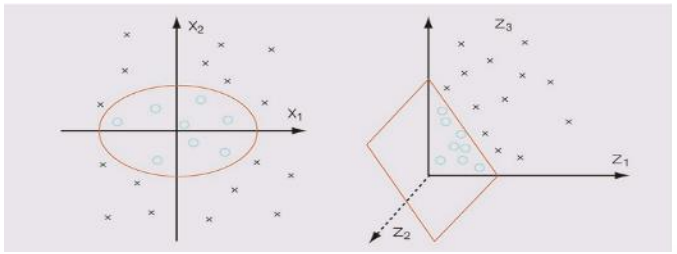
线性适应元



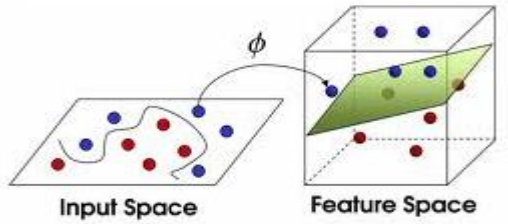
连接主义学习模型



统计学习模型



“核方法”



机器学习发展历史

开启人工智能的新时代

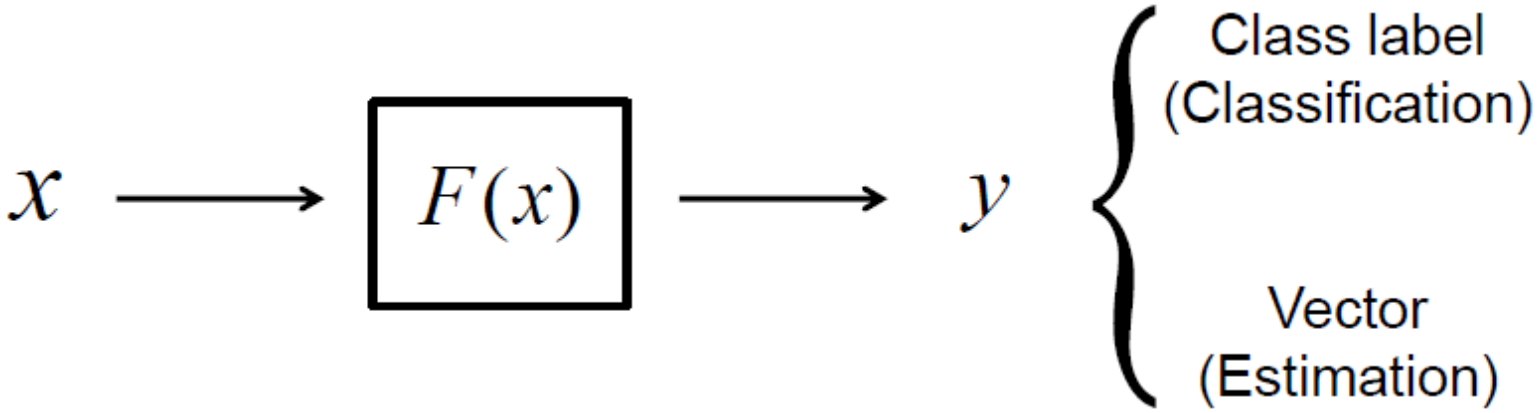
机器学习发展历史

第 11 章
机器学习方法

1 机器学习的发展历史

2 **机器学习的方法**

3 机器学习最新发展



Object recognition

{dog, cat, horse,, …}

机器学习的基本任务

特征

求职者	笔试成绩	语言	工作经验	是否录取
1	A	Java	2年	录取
2	B	Python	2年	录取
3	C	C++	1年	录取
4	A	C	0年	不录取

样本

- 数据集:记录的集合，例如上表中4个求职者的所有信息。
- 样本（示例）:描述一个对象或者事件的记录。例如表中的描述每个求职者的记录就是一个样本。
- 特征:反应对象在某方面的表现或者性质的事项，例如笔试成绩，语言，工作经验。
- 属性（样本）空间:属性张成的空间。例如我们把笔试成绩，语言，工作经验作为3个坐标轴，每个求职者都能在这3维空间中找到自己的位置。
- 由于空间中每个点对应一个向量，因此我们可以把每一个样本称为一个**特征向量**。

基本术语

模型是机器学习中的一个重要概念。简单的讲，模型是样本空间到输出空间的映射，一般由模型的假设函数与参数w组成，例如房价示例中简单的线性模型：

$$房子价格 = beta * 房屋面积 + alpha$$

模型参数为beta和alpha。当训练好模型参数beta与alpha后，就可以输入一个房屋面积，进而预测房屋价格。

- 从数据中学得模型的过程称为“学习”(learning)或“训练”(training)，这个过程通过执行某个学习算法来完成。比如给定上述求职者的信息，面试后再给每个求职者一个标记“录取”或者“不录取”。我们可以通过给定标记（录取或者不录取）的数据中训练一棵决策树，从而不用面试新的求职者，直接用决策树模型就可以判断是否录取。

基本术语

- 训练过程中使用的数据称为“训练数据”(training data)，其中每一个样本称为一个训练样本(training sample)，训练样本组成的集合称为训练集(training set)。
- 学得的模型对应了关于数据的某种潜在的规律，因此称模型为“假设”(hypothesis)。这种“潜在规律”称为“真相”。学习的过程就是为了找出或者逼近真相。
- 若我们欲预测的是离散值，例如“生存”或者“死亡”，此类学习任务称为分类；若预测的值是连续值，例如预测房价，则此类学习任务称为“回归”。

基本术语

- 按照训练的数据有无标签，可以将机器学习方法分为监督学习算法和无监督学习算法，但推荐算法较为特殊，既不属于监督学习，也不属于非监督学习，是单独的一类。
  - 监督学习算法：线性回归，逻辑回归，神经网络，SVM
  - 无监督学习算法：聚类算法，降维算法
  - 特殊算法：推荐算法

机器学习的方法分类

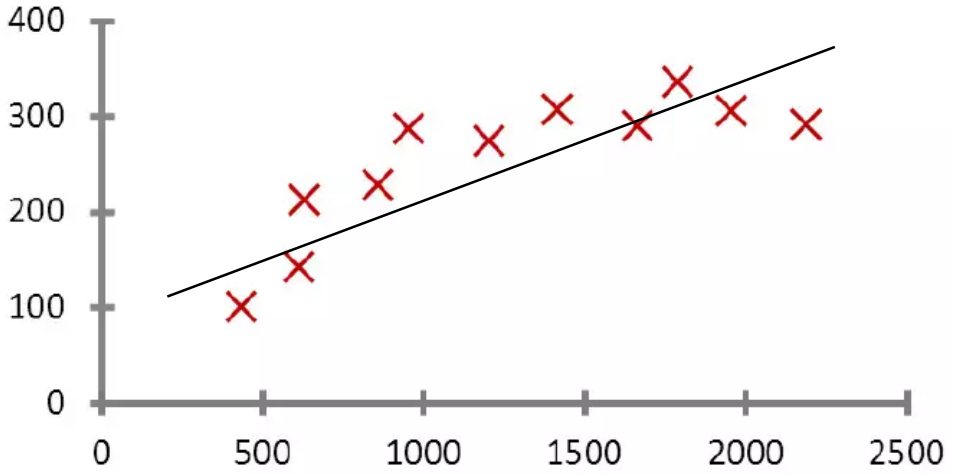
求职者	笔试成绩	语言	工作经验	是否录取
1	A	Java	2年	录取

标识

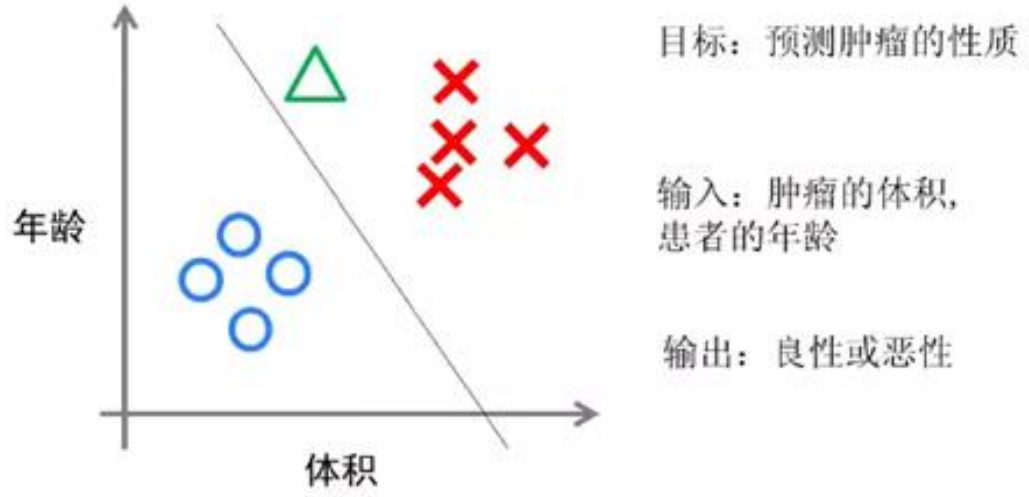
- 在监督式学习下，训练数据集中的每一个训练数据都有一个明确的标识或标记。如上面求职者的数据中(初试成绩=A；语言=java；工作经验=2年)，录取)，则“录取”为该训练数据的标识。在建立预测模型的时候，监督式学习建立一个这样的一个学习过程，将预测结果与训练数据的的实际结果进行比较，不断地调整预测模型，直到模型的预测结果达到一个预期的准确率。监督式学习常见的应用场景为回归问题和分类问题。常见的算法有线性回归，逻辑回归，决策树等。

有监督学习

- 线性回归就是我们前面说过的房价求解问题。如何拟合出一条直线最佳匹配我所有的数据？例如：“最小二乘法”来求解。
- 回归算法有两个重要的子类：即线性回归和逻辑回归。

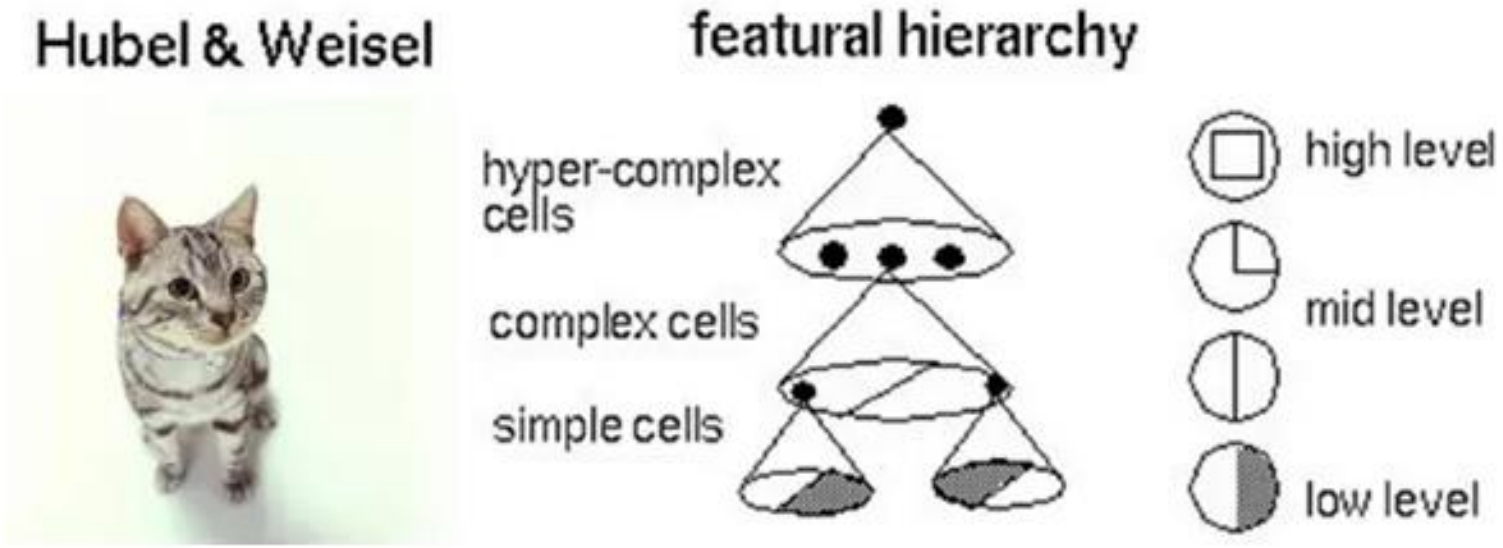


回归方法



有监督学习

- 神经网络(也称之为人工神经网络，ANN)，是8o年代机器学习界非常流行的方法，其诞生起源于对大脑工作机理的研究。简单来说，就是分解与整合。

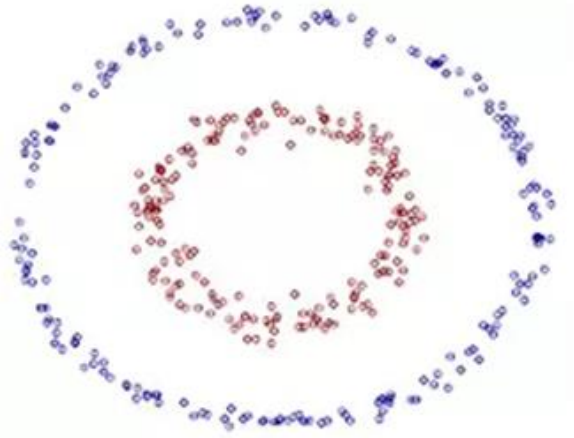


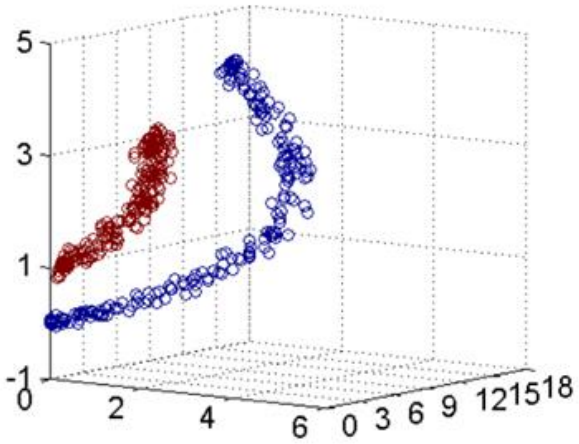
Hubel-Wiesel试验与大脑视觉机理

神经网络

有监督学习

- 支持向量机诞生于统计学习界，从某种意义上来说是逻辑回归算法的强化：通过给予逻辑回归算法更严格的优化条件，支持向量机算法可以获得比逻辑回归更好的分类界线。通过跟高斯“核”的结合，支持向量机可以表达出非常复杂的分类界线，从而达成很好的的分类效果。





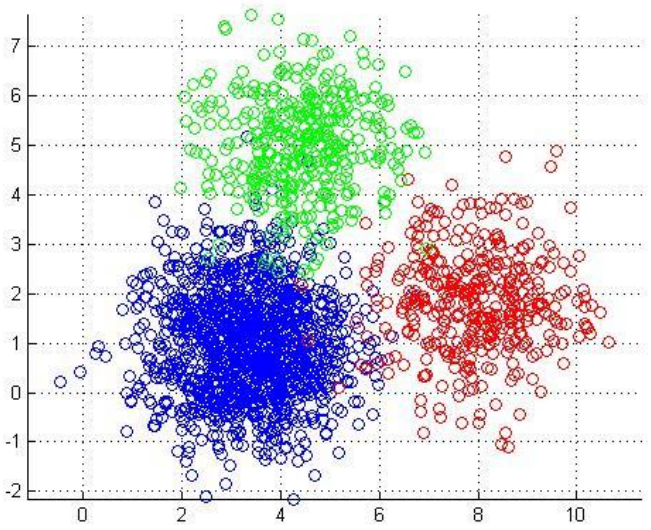
SVM（支持向量机）

有监督学习

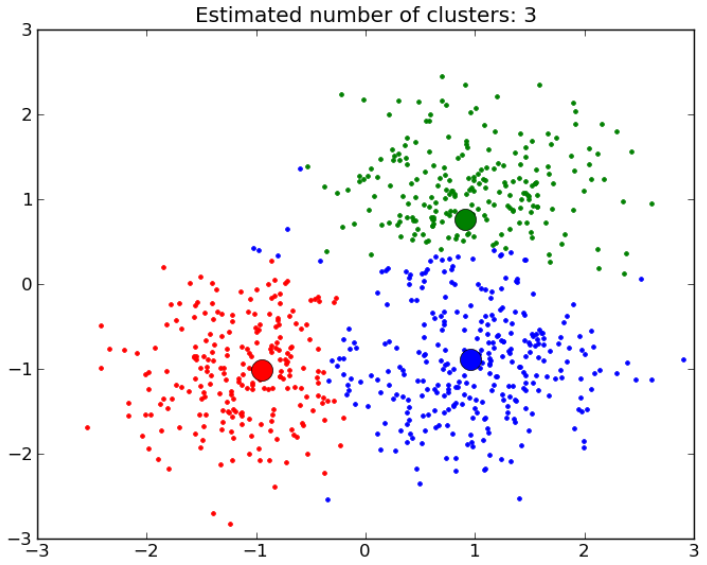
- 在无监督学习下，训练数据不被特别标识，学习模型是为了推断出数据的一些内在结构。常见的应用场景包括关联规则的学习和聚类。常见的算法有K-Means算法和Apriori算法。

无监督学习

- 这类方法有一个统称，即无监督算法，其中最典型的代表就是聚类。聚类就是计算种群中的距离，根据距离的远近将数据划分为多个族群。
- 聚类算法中最典型的代表就是K-Means算法。



聚类方法

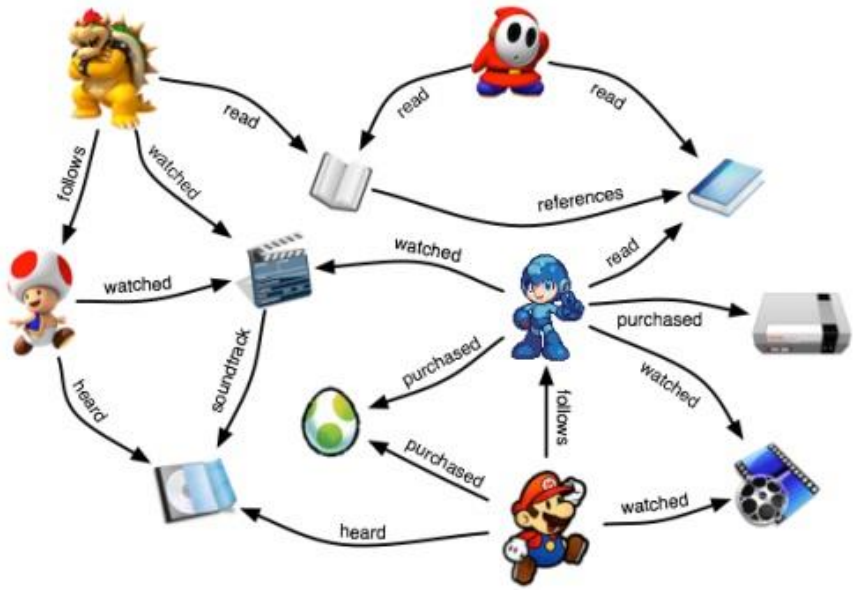


无监督学习

- 在半监督学习下，输入数据部分被标识，部分没有被标识。这种学习模型可以用来进行预测，但模型首先需要学习数据的内在结构，以便合理地组织数据进行预测。其应用场景包括分类与回归。常见的算法包括一些对常用的监督式学习算法的延伸。这些算法首先试图对未标识的数据进行建模，然后在此基础上对标识的数据进行预测，如期望最大化算法(EM)。

半监督学习

- 推荐算法是目前业界非常火的一种算法，在电商界，如亚马逊，天猫，京东等得到了广泛的运用。推荐算法的主要特征就是可以自动向用户推荐他们最感兴趣的东西，从而增加购买率，提升效益。



推荐算法

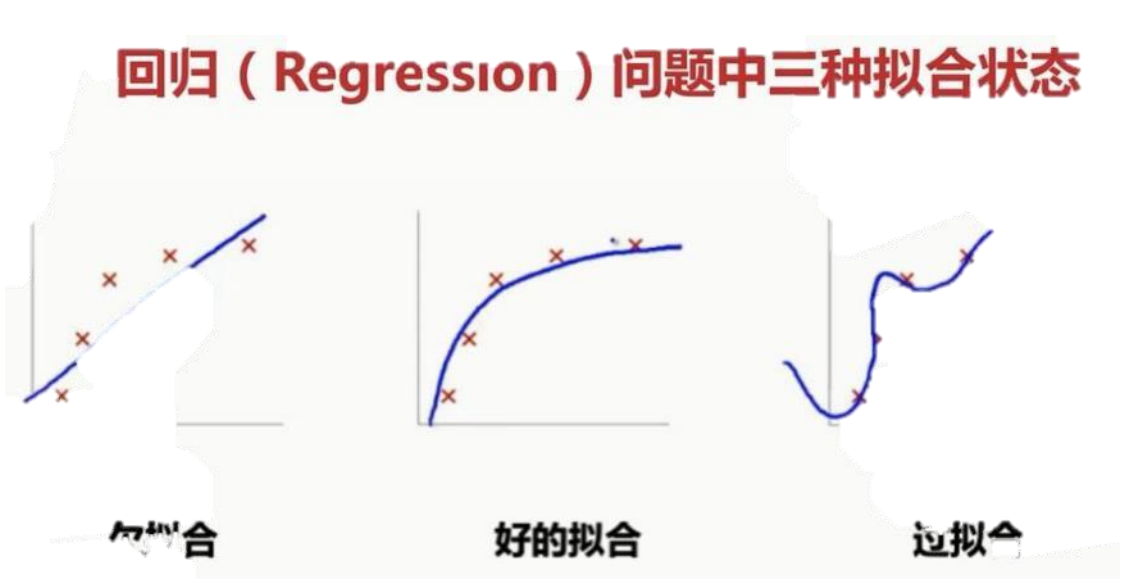
半监督学习

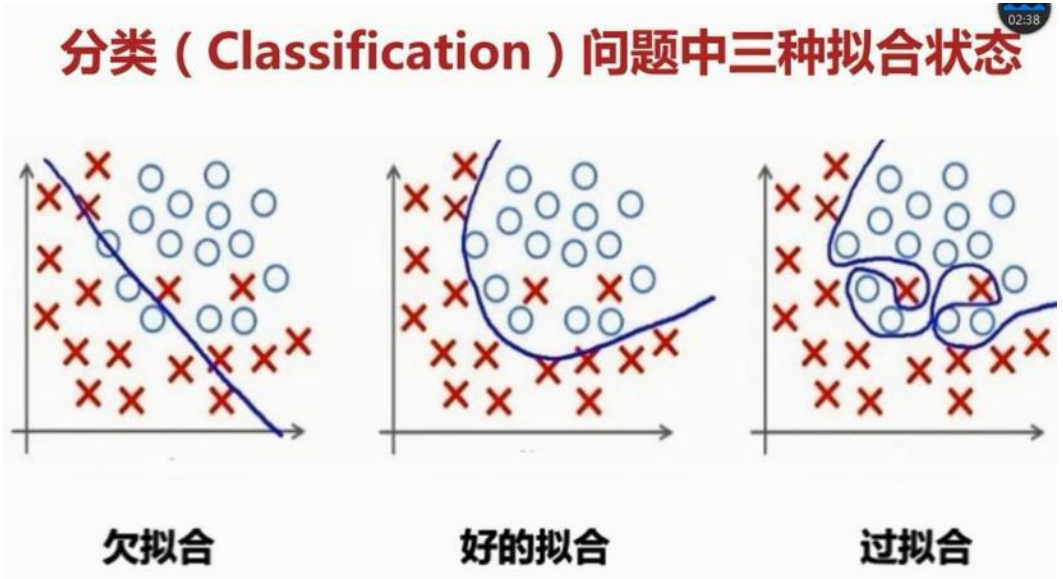
- 训练误差：根据训练集学得模型之后，模型在训练集上的误差称为“训练误差”或“经验误差”。若是回归任务，则训练误差为模型的预测值与真实值的差的平方；若是分类任务，则是误分类的训练样本与总样本的比值，称这个比值为错误率。
- 泛化误差：模型在新样本（没有出现在训练集中）上的误差称为“泛化误差”。
- 泛化误差是我们追求的目标。但实际中我们并不知道新样本是什么，我们能做的是学得一个训练误差很小、在训练集上表现良好的学习器。

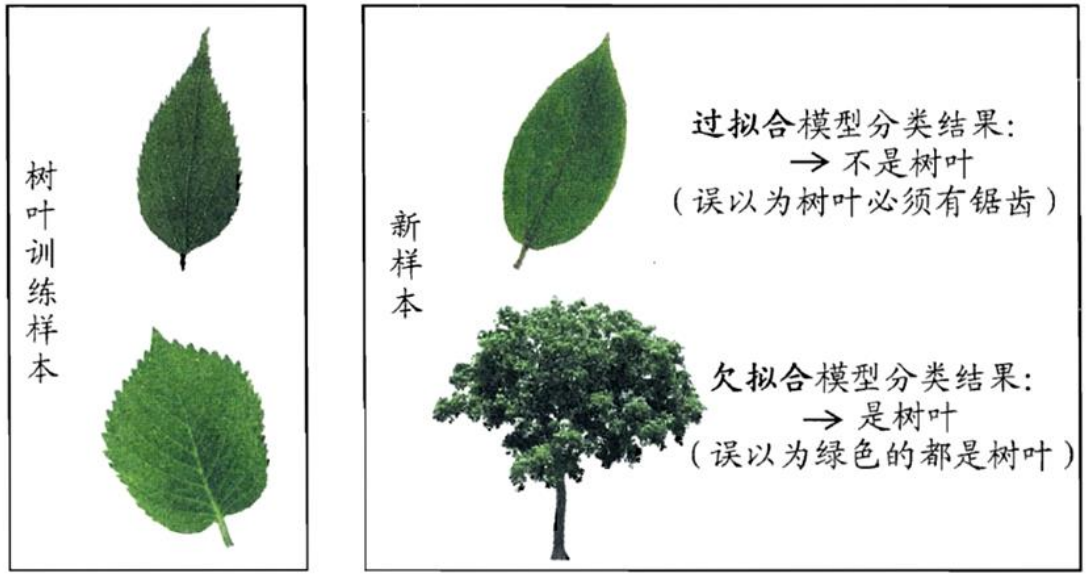
过拟合与欠拟合

- 由于实际中，我们能做的是降低训练误差。但当模型很复杂时，很多时候能导致过拟合。
- 过拟合：当模型把训练集学得“太好”了，即模型在训练集上表现的非常好。这很可能把训练样本自身的一些特点当作了所有潜在样本都会具有的一般性质，这样就会导致泛化能力下降，这种现象称为过拟合。
- 欠拟合：当模型过于简单，对训练集的一般性质尚未学好，模型在训练集上都表现的不好，此现象称为欠拟合。欠拟合通常意味着模型不够好，需要继续改进模型。

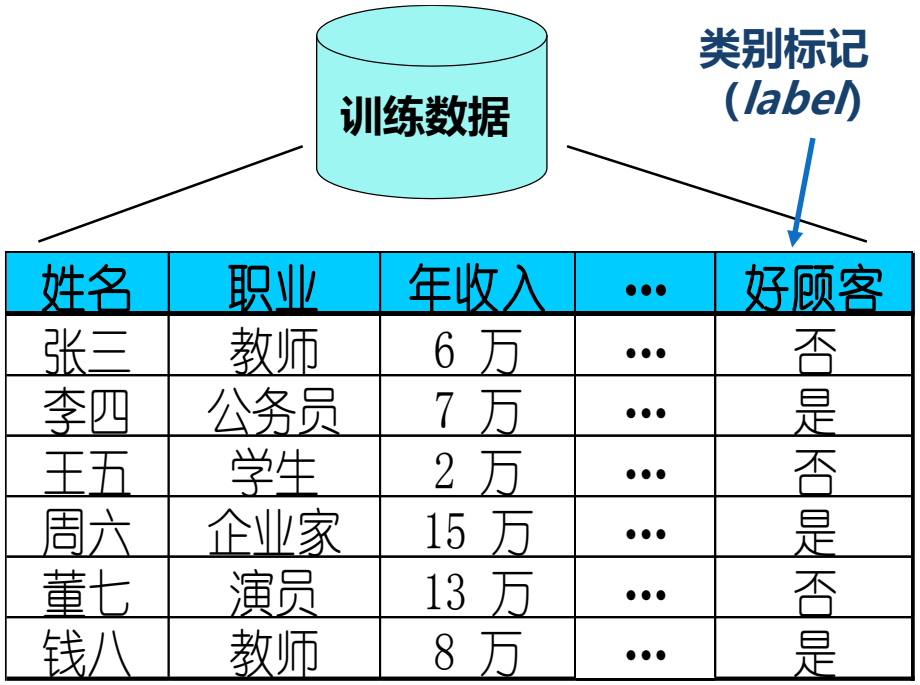
过拟合与欠拟合


回归（Regression）问题中三种拟合状态
好的拟合


分类（Classification）问题中三种拟合状态
欠拟合
好的拟合
过拟合

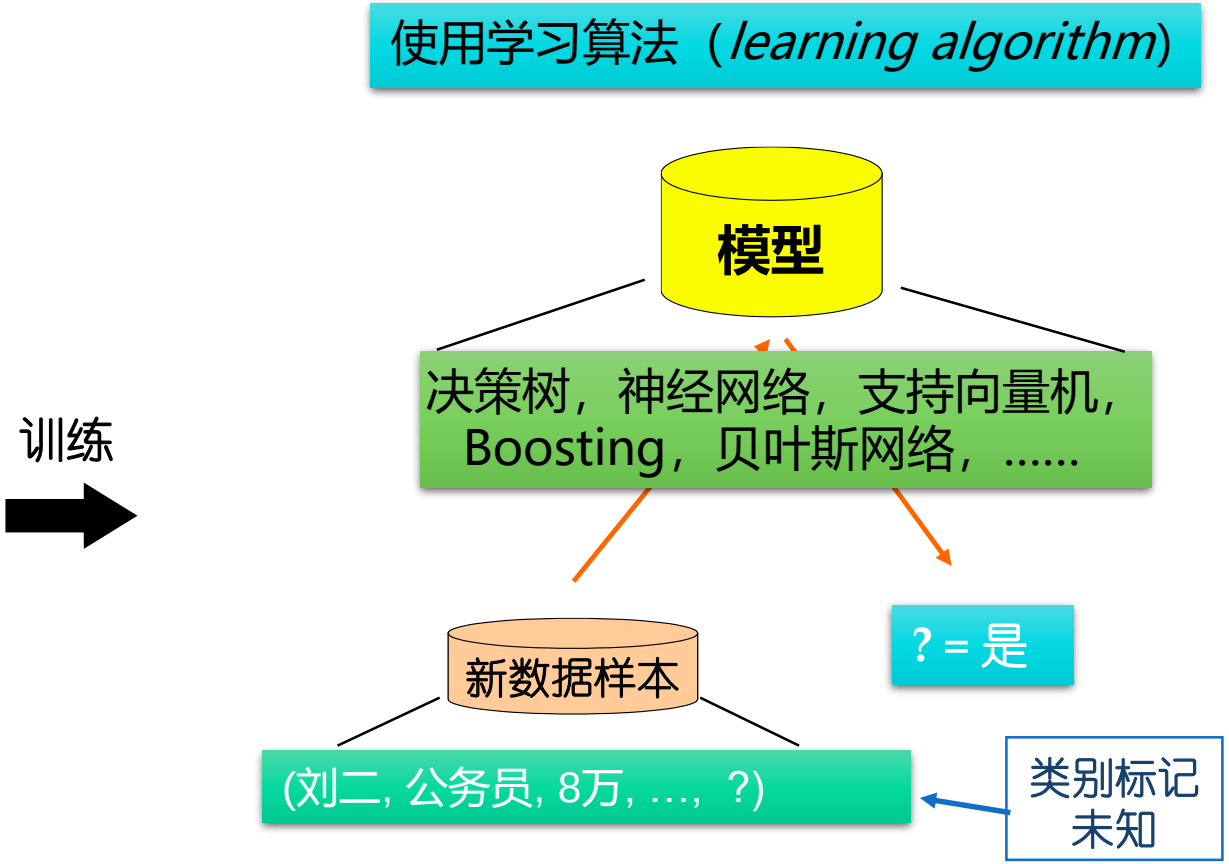

树叶训练样本
新样本
过拟合模型分类结果：
→ 不是树叶
（误以为树叶必须有锯齿）
欠拟合模型分类结果：
→ 是树叶
（误以为绿色的都是树叶）

过拟合与欠拟合



姓名	职业	年收入	…	好顾客
张三	教师	6 万	…	否
李四	公务员	7 万	…	是
王五	学生	2 万	…	否
周六	企业家	15 万	…	是
董七	演员	13 万	…	否
钱八	教师	8 万	…	是

训练



典型的机器学习过程

- 训练出模型之后，我们需要一个测试集对模型进行评估。
- 问题：我们只有一个数据集，怎么才能做到既要训练，又要测试？
- 答案是：通过对数据集进行划分，从中产生训练集S和测试集T。
- 常见的划分数据集方法：
  - 留出法
  - 交叉验证法
  - 自助法
- 训练集S和测试集T比例通常为2:1~4:1

模型评估

第 11 章
机器学习方法

1 机器学习的发展历史

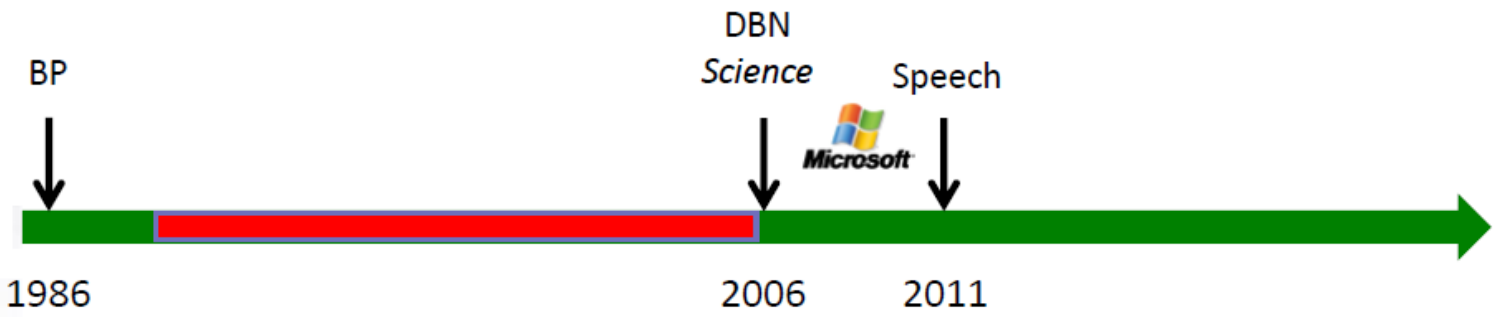
2 机器学习的方法

3 **机器学习最新发展**

- 大数据
  - 语音图像视频
- 计算能力
  - 并行计算平台
  - GPU 大量部署
- 开放的社区
  - 开源，开放数据

机器学习也要借助东风

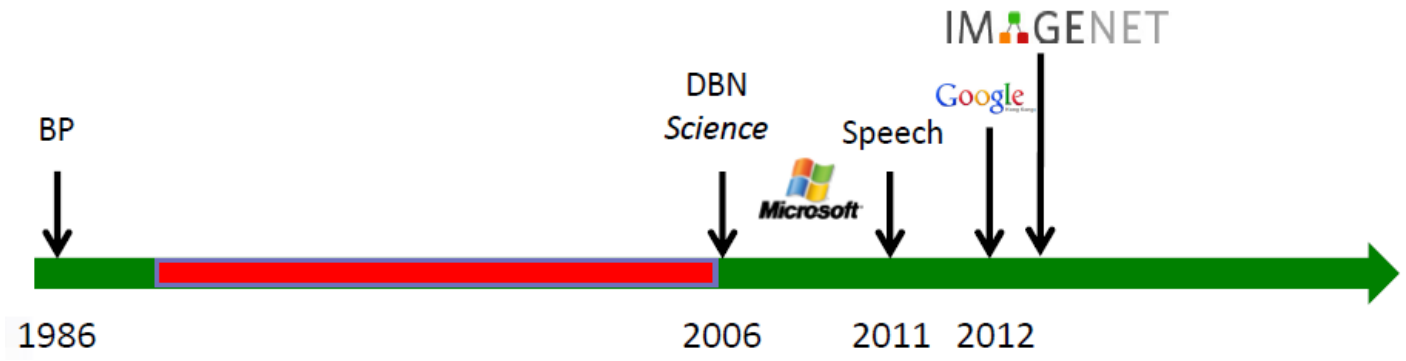
task	hours of training data	DNN-HMM	GMM-HMM with same data
Switchboard (test set 1)	309	18.5	27.4
Switchboard (test set 2)	309	16.1	23.6
English Broadcast News	50	17.5	18.8
Bing Voice Search (Sentence error rates)	24	30.4	36.2
Google Voice Input	5,870	12.3	
Youtube	1,400	47.6	52.3



语音识别（2011）

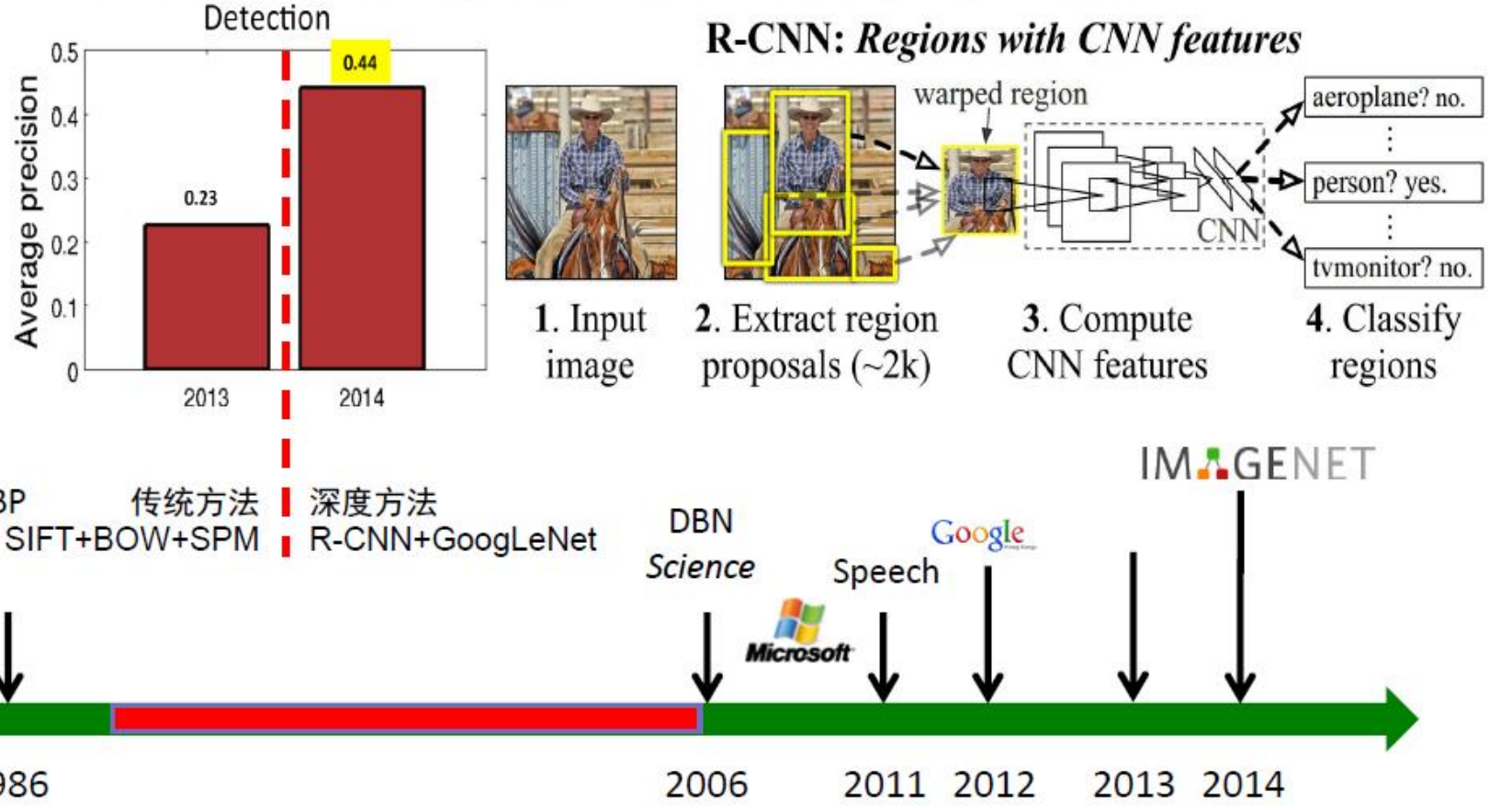
Rank	Name	Error rates(TOP5)	Description
1	U. Toronto	0.153	Deep learning
2	U. Tokyo	0.261	Hand-crafted features and learning models. Bottleneck.
3	U. Oxford	0.270	
4	Xerox/INRIA	0.271	

- ImageNet 物体分类任务上
  - 物体分类任务：1000 类， 1,431,167 幅



计算机视觉（2012）

- 200类，456, 567 幅图像，检测率



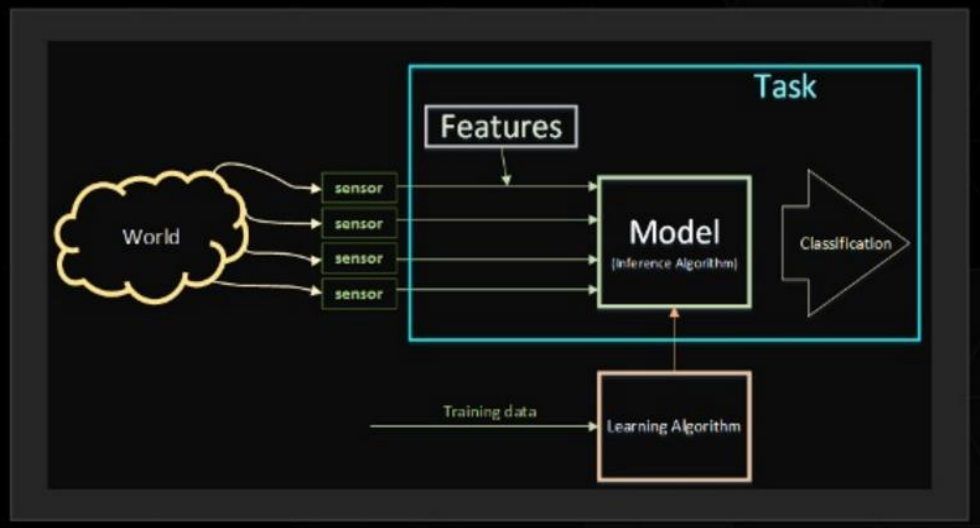
ImageNet物体检测任务

计算机视觉（2012）

WHAT IS DEEP LEARNING?

Systems that learn to recognize objects that are important, without us telling the system explicitly what that object is ahead of time

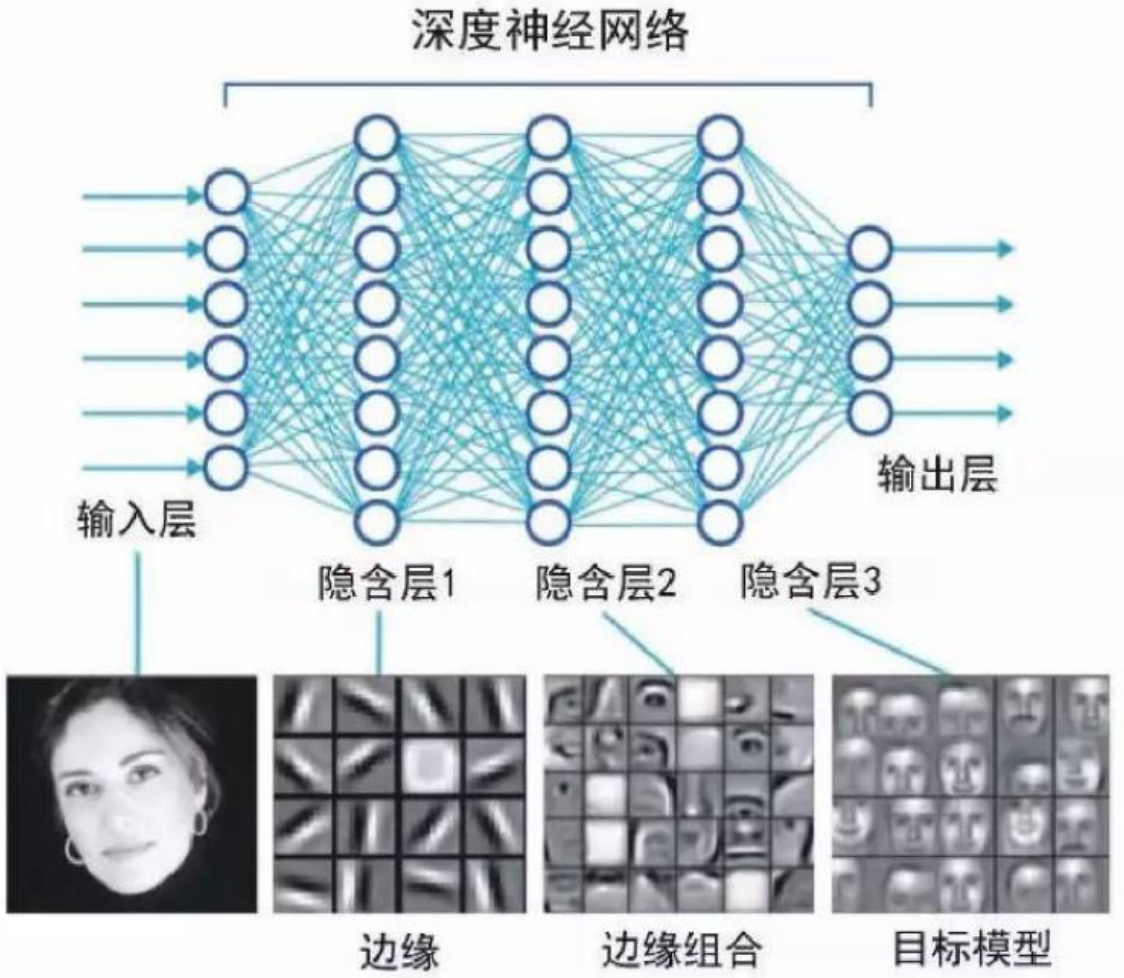
- Key components
  - Task
  - Features
  - Model
  - Learning Algorithm



深度学习

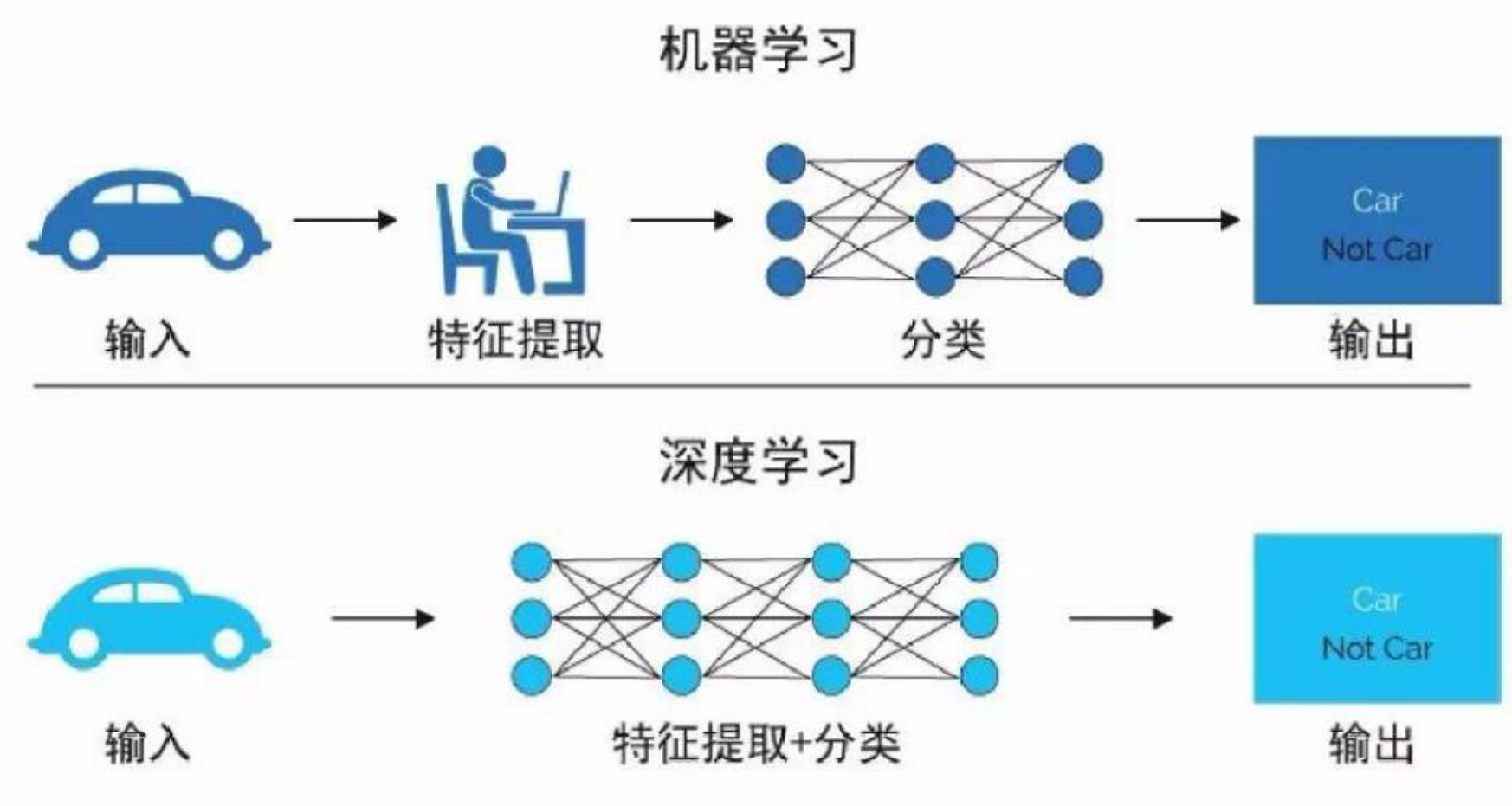
深度学习、机器学习、人工智能三者关系

深度学习



典型深度学习工作流程

深度学习



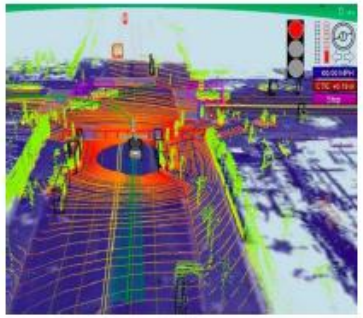
机器学习与深度学习的区别

深度学习

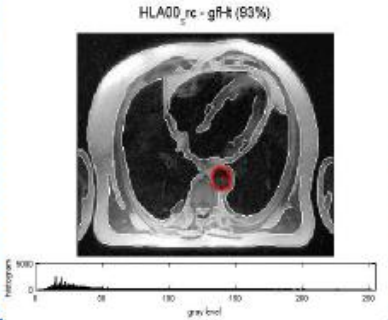
人脸认证

智慧城市
（安保、交运）

辅助驾驶



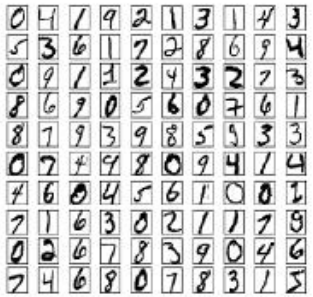
医学诊断



智能硬件

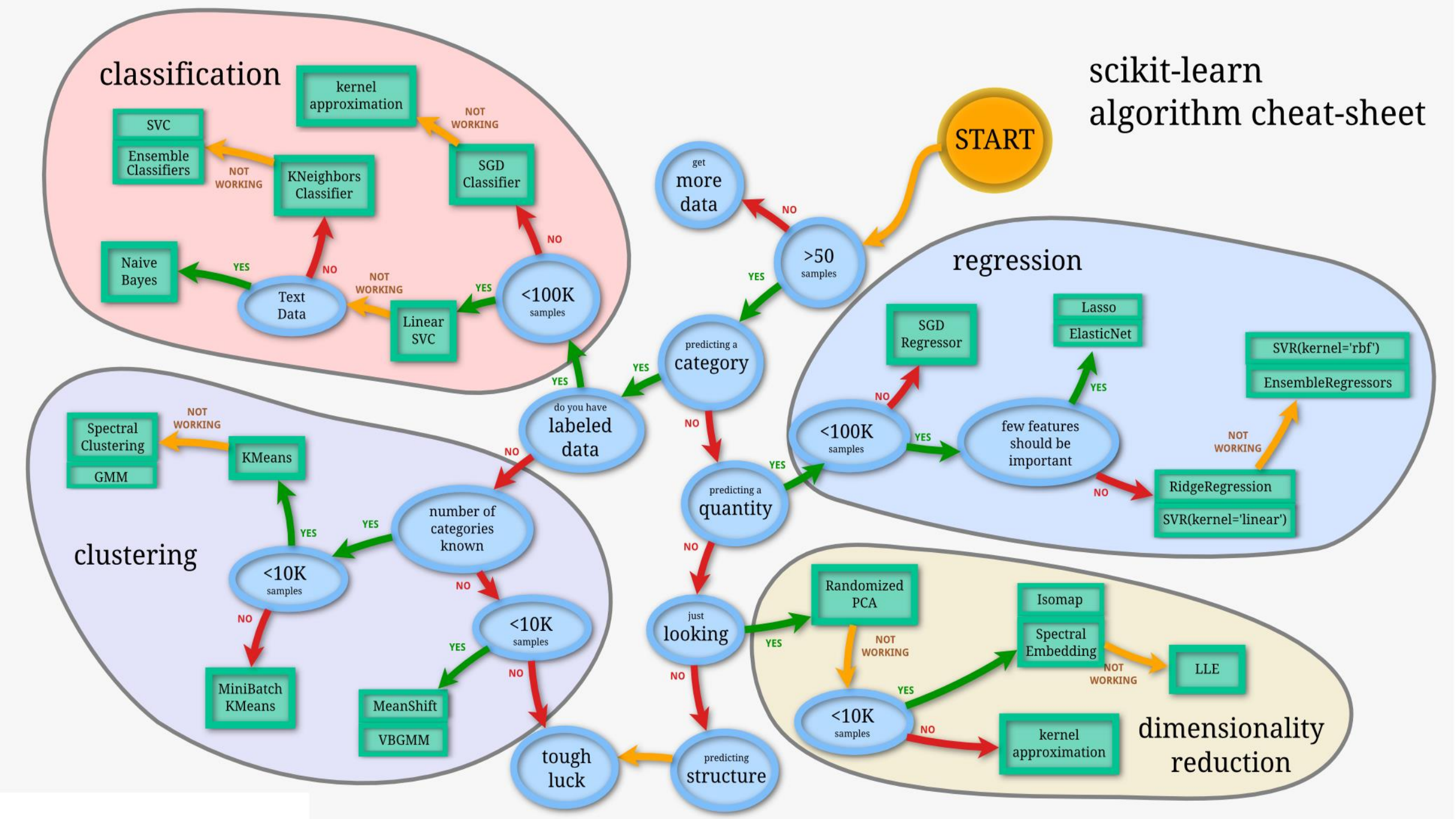
人机交互

输入法



广告营销

机器学习的今天与未来


scikit-learn
algorithm cheat-sheet
START
>50 samples
NO
get more data
YES
predicting a category
YES
do you have labeled data
NO
predicting a quantity
NO
just looking
NO
predicting structure
tough luck
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
Text Data
NO
NOT WORKING
Linear SVC
YES
<100K samples
NO
clustering
Spectral Clustering
GMM
NOT WORKING
KMeans
YES
number of categories known
YES
<10K samples
NO
MiniBatch KMeans
NO
<10K samples
YES
MeanShift
VBGMM
NO
regression
SGD Regressor
NO
<100K samples
YES
few features should be important
YES
Lasso
ElasticNet
NO
RidgeRegression
SVR(kernel='linear')
NOT WORKING
SVR(kernel='rbf')
EnsembleRegressors
YES
dimensionality reduction
Randomized PCA
YES
NOT WORKING
<10K samples
YES
Isomap
Spectral Embedding
NOT WORKING
LLE
NO
kernel approximation


THE MASTER ALGORITHM
HOW THE QUEST FOR THE ULTIMATE LEARNING MACHINE WILL REMAKE OUR WORLD
终极算法
机器学习和人工智能如何重塑世界
(Pedro Domingos)
[美] 佩德罗·多明戈斯 著
中信出版集团


周志华 著
MACHINE LEARNING
机器学习
清华大学出版社

书籍推荐

第11章
机器学习方法

1 机器学习的发展历史

2 机器学习的方法

3 机器学习最新发展