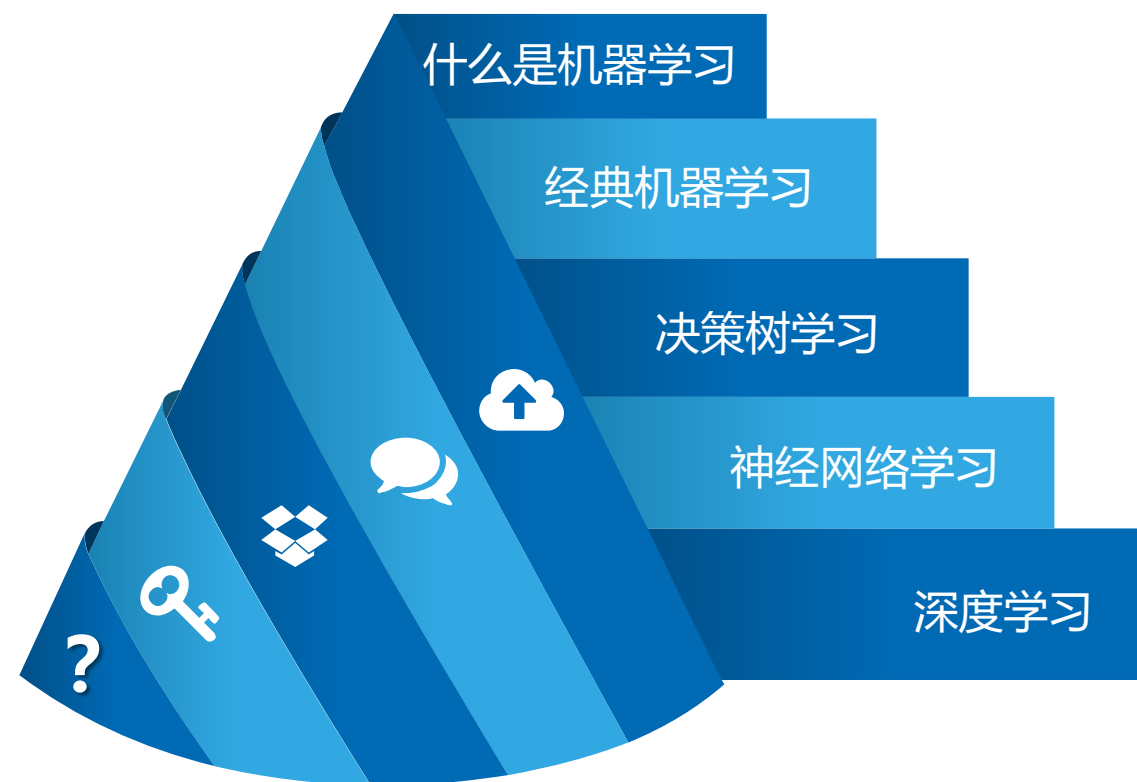


The background of the slide features two wireframe hands, one in the upper right and one in the lower left, reaching towards the center. The hands are composed of a grid of white lines. The background is a solid blue color with faint, concentric circular patterns and small white star-like sparkles.

第6讲 机器学习与知识发现

周文晖

杭州电子科技大学



什么是机器学习?

机器学习原理, 基本过程, 机器学习分类 ...



经典机器学习

简单机器学习、经典机器学习, ...



决策树学习

什么是决策树, 学习步骤, ...



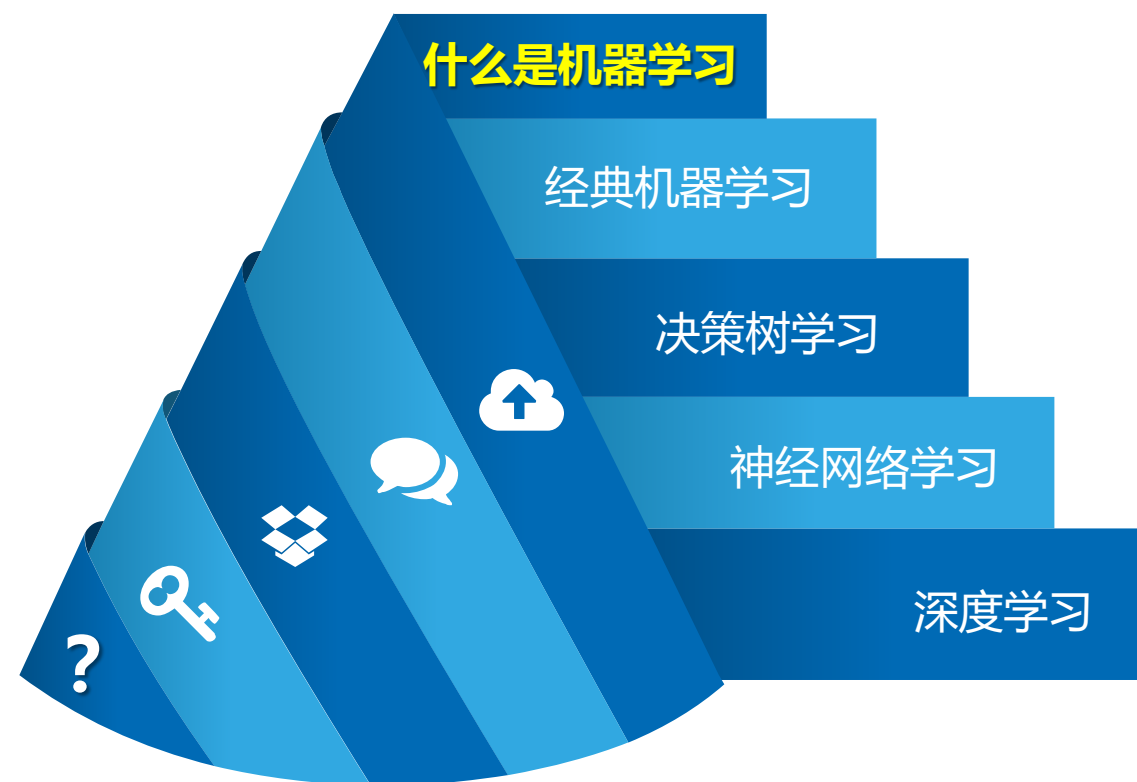
神经网络学习

人工神经元, 人工神经网络发展历史, 感知器, BP, ...



深度学习

什么是深度学习, 深度学习的发展历史, 卷积神经网络, 常见深度学习网络, ...



什么是机器学习?

机器学习原理，基本过程，机器学习分类 ...



经典机器学习

简单机器学习、经典机器学习， ...



决策树学习

什么是决策树，学习步骤， ...



神经网络学习

人工神经元，人工神经网络发展历史，感知器， BP， ...



深度学习

什么是深度学习，深度学习的发展历史，卷积神经网络，常见深度学习网络， ...

为什么要研究机器学习？

人工智能主要是为了研究人的智能，模仿其机理将其应用于工程的科学。

该过程必然涉及“人类怎样做才能获取这种特殊技能（或知识）？”。

当前人工智能研究的主要障碍和发展方向之一就是机器学习，包括学习的计算理论和构造学习系统。

机器学习的核心思想

人工智能的发展经历了两个阶段：

(1) 逻辑推理，知识库与专家系统

1980之前——逻辑推理，人工建立规则、知识库；

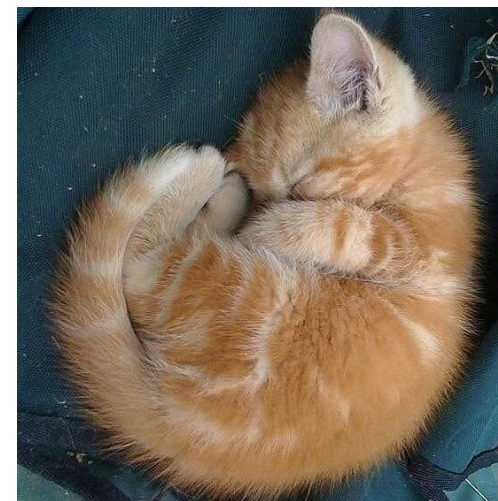
- 可扩展性和通用性差，需要各领域的专家知识，构建专家系统；
- 有些问题无法用精确的规则描述——只可意会不可言传。

(2) 机器学习

1980之后——模仿人的学习能力，从实例中进行学习获得经验和知识。



如何判断一张图像是不是猫？



这一张图像呢？

机器学习的核心思想

与其把知识和经验总结好了告诉计算机，还不如让计算机自己去学习知识和经验。

机器学习的主要动机：与其显式地编写程序来执行某些任务，不如教计算机如何开发一个算法来完成任务。

机器学习思想：使用算法解析数据，从中**学习**，然后对未知事件做出决定或预测。

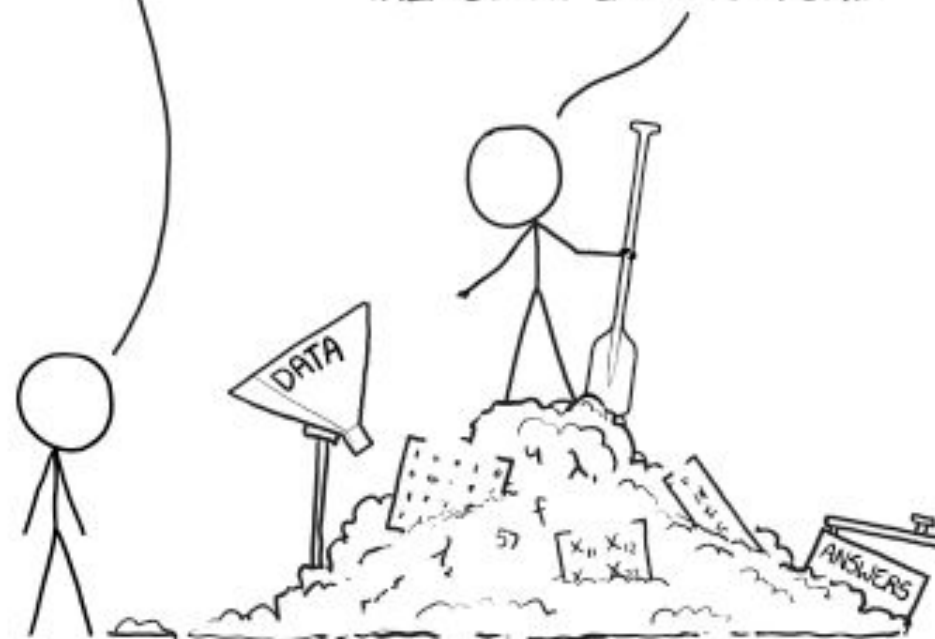
学什么？如何学？怎么学？

THIS IS YOUR MACHINE LEARNING SYSTEM?

YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.



人类的学习

心理学认为：学习是指（人或动物）依靠经验的获得而使行为持久变化的过程。

- Simon（1983）：学习就是系统中的变化，这种变化使系统比以前更有效地去做同样的工作。
- Minsky（1985）：学习是在我们头脑中（心里内部）进行有用的变化。

婴儿通过识图卡或者实物学会了认识物体



初生儿脑皮层
单个神经元易看清



3个月婴儿脑皮层
神经元连接增多



2岁婴儿脑皮层
密集的神经元网络

人类学习与机器学习

人脑有哪些功能？

- 感知 视觉，听觉，嗅觉，触觉
- 计算 数学计算
- 推理 逻辑推理，证明数学定理
- 记忆 长期记忆，短期记忆
- 联想 触景生情
- 控制 控制肢体运动
- 语言 理解语言，说话，写作
- 创作 绘画，写诗，创作音乐
- 感情 情绪，恋爱

典型问题	方法	解决时间
算术运算	CPU指令	1945年
大规模线性方程组	Krylov子空间迭代	1950年
海量数据的排序	快速排序算法	1962年
寻找地图上两点间最短路径	Dijkstra算法	1959年
存储并管理大规模数据	关系型数据库系统	1970年
数学公式推导	符号计算	1970年代
大规模信息检索	搜索引擎	1990年代
感知问题-听觉	机器学习	还未解决
感知问题-视觉	机器学习	还未解决
运动控制	机器人控制算法	还未解决
理解人类语言	机器学习	还未解决
创作，如诗歌和音乐	机器学习	还未解决

有确定性的数学，逻辑模型的问题，计算机远胜人类；无法建立这种精确模型的问题，恰好是机器学习要解决的核心问题。

学习的过程

人类学习→机器学习：

- 获取新的陈述性知识。
- 通过教育或实践发展认知能力。
- 将新知识组织成为通用化和有效的表达形式。
- 借助观察和实验发现新的事实和新的理论。

机器学习：模仿了人的学习能力，从样本数据中学习**经验**，将**经验**用于预测。

机器学习的原理

原理：

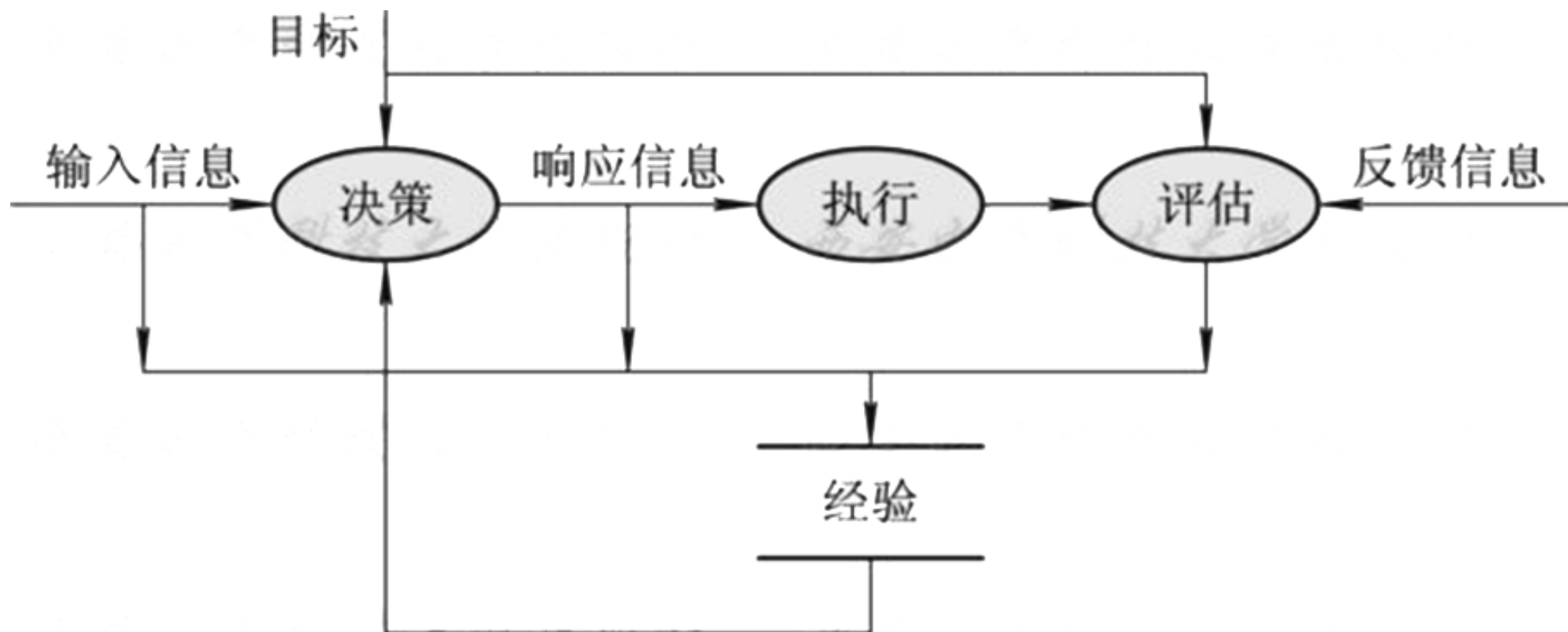
- 学习与经验有关。
- 学习可以改善系统性能。
- 学习是一个有反馈的信息处理与控制过程。

经验是在系统与环境的交互过程中产生的。

经验中应包含系统输入、响应和效果等信息。

因此经验积累、性能完善正是通过重复这一过程而实现的。

机器学习原理图示



该原理称为记忆学习

记忆学习

记忆学习特点:

- 记忆学习是人类和动物的一种基本学习方式。
- 记忆学习依靠经验来提高性能

记忆学习存在严重不足

- 经验积累是一个缓慢过程,致使系统性能的改善也很缓慢;
- 由于经验不是规律, 仅凭经验对系统性能的改善是有限的, 有时甚至是靠不住的。

机器学习完善

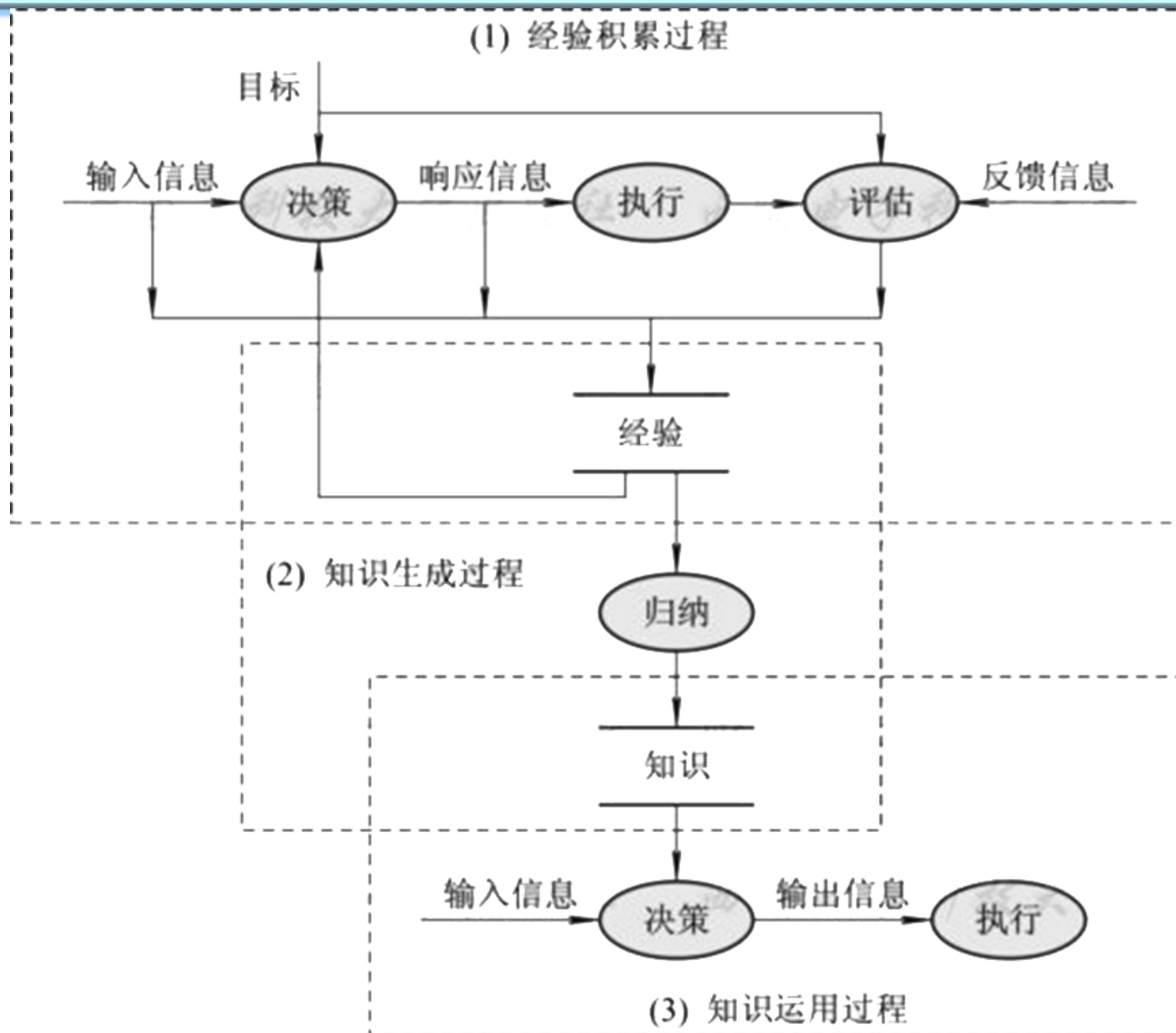
学习方式需要延伸和发展。

机器学习的改善和提高：

- 在积累的经验中进一步发现规律；
- 然后利用所发现的规律即知识来指导系统行为。

机器学习是从**数据**中自动分析获得**规律（模型）**，并利用**规律**对未知数据进行**预测**。

完整的学习过程图示



完整的学习过程

完整的学习过程分为三个子过程：

- 经验积累过程
- 知识生成过程：归纳推导而得出知识
- 知识运用过程：用所得知识去指导行为、改善性能

这种学习方式是人类和动物的技能训练或者更一般的适应性训练过程

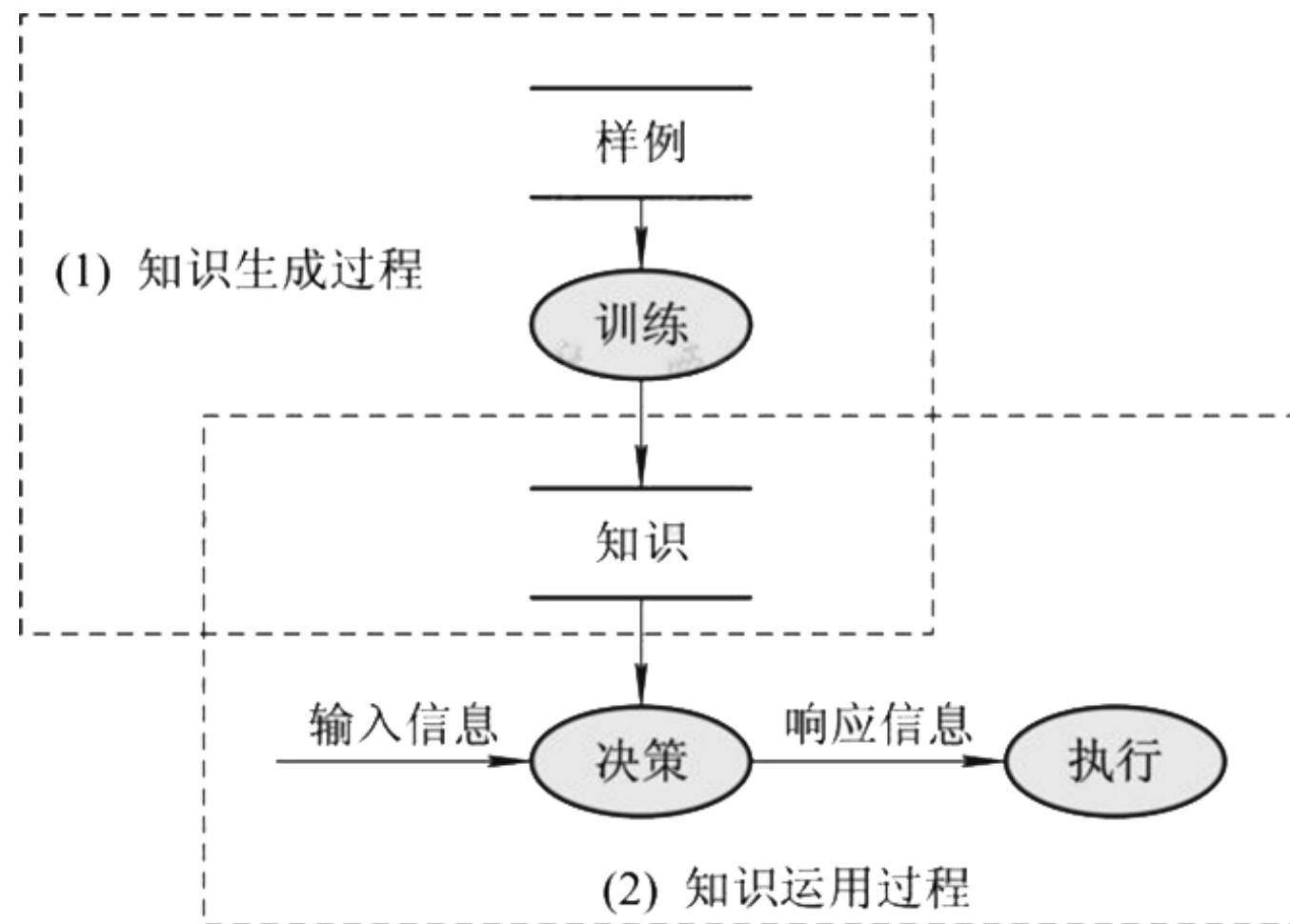
- 如骑车、驾驶、体操、游泳等
- 适合机器的技能训练, 如机器人驾车训练。

通常机器学习过程

通常机器学习研究中省去了经验积累过程。

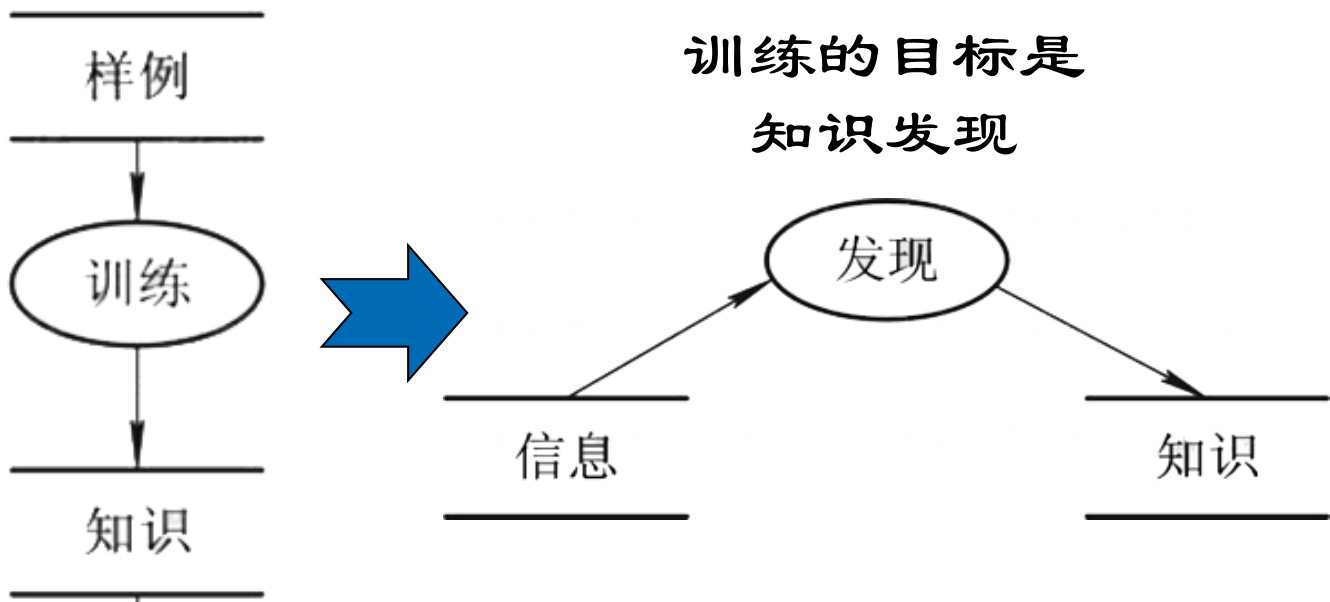
将事先组织好的经验数据（包括实验数据和统计数据）直接作为学习系统的输入。

把组织好的经验数据称为训练样本或样例, 把由样例到知识的转换过程称为学习或训练。

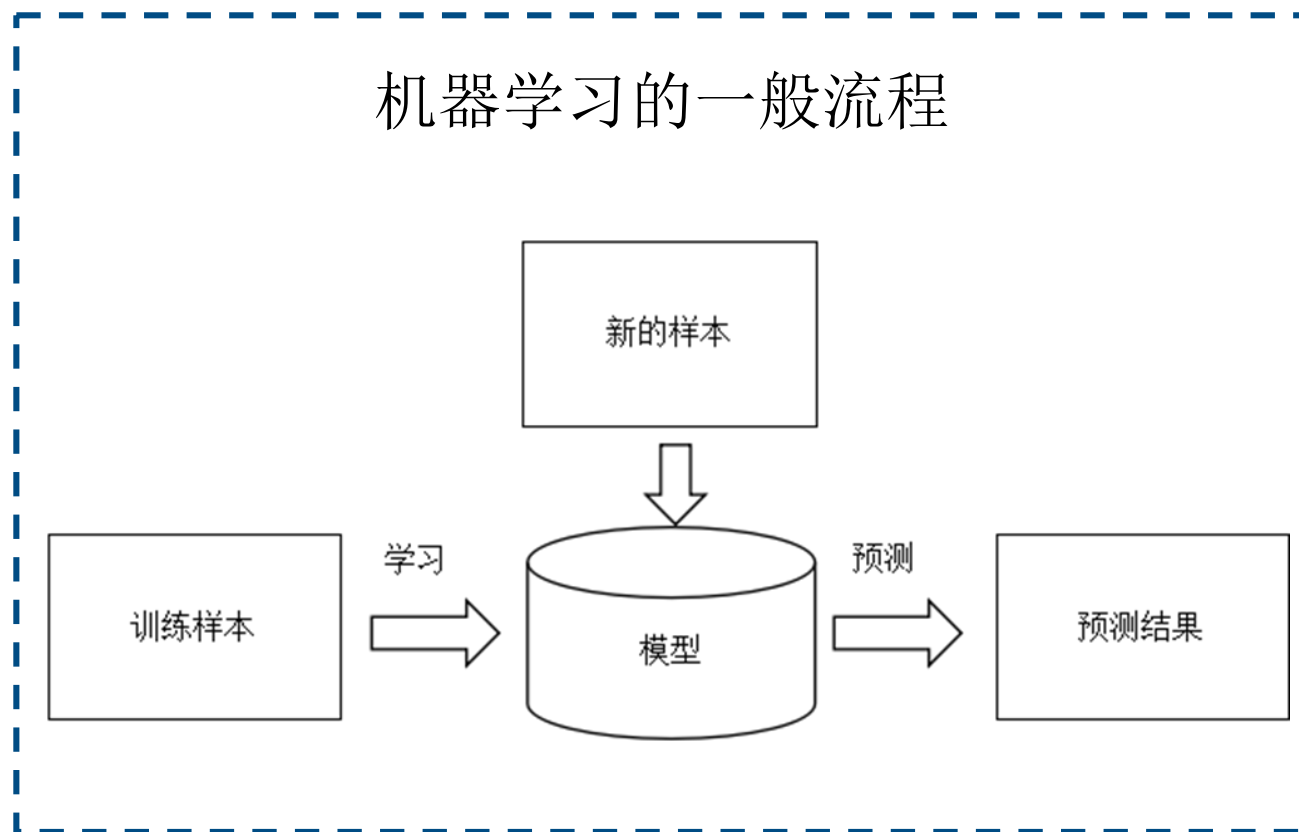


机器学习的关键环节

机器学习的关键环节在于知识发现。



机器学习的一般流程



机器学习的三要素

机器学习的三个要素：

- 信息，发现，知识
- 分别是机器学习的对象、方法和目标。

一个简单的例子——如何区分樱桃和猕猴桃？

选择两个有区分度的特征：颜色 x ，重量 y ，组成特征向量 (x, y)

樱桃的颜色**亮**，重量**小**；猕猴桃的颜色**暗**，重量**大**



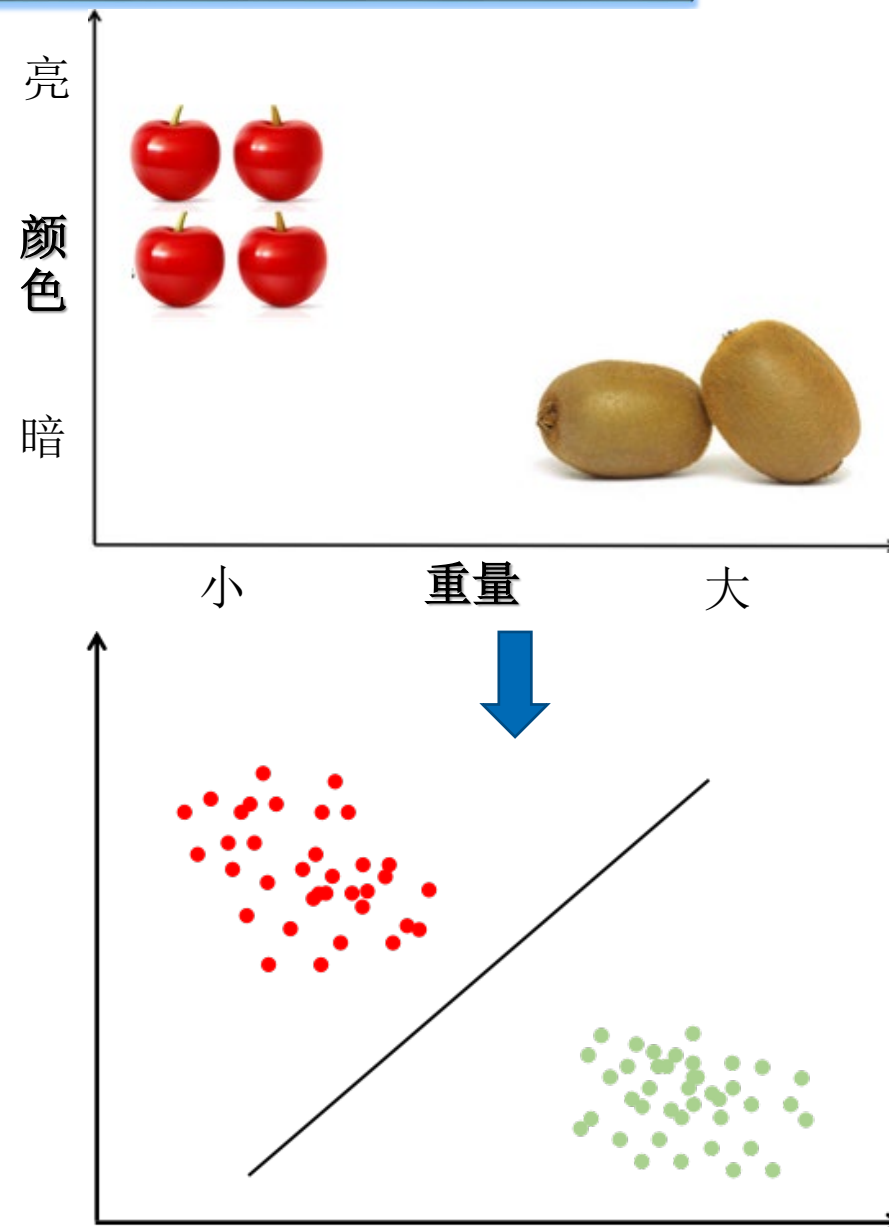
机器学习的一个简单的例子

如何区分樱桃和猕猴桃？

樱桃的颜色**亮**，重量**小**；猕猴桃的颜色**暗**，重量**大**；

收集一些樱桃、猕猴桃，作为训练样本，测量它们的这两个特征；樱桃位于xy平面的左上角，猕猴桃位于右下角。

把樱桃和猕猴桃抽象成平面上的点，
用一条直线可以将这两种类型分开，这条直线称为**线性分类器**。



机器学习的一个简单的例子

如何区分樱桃和猕猴桃？

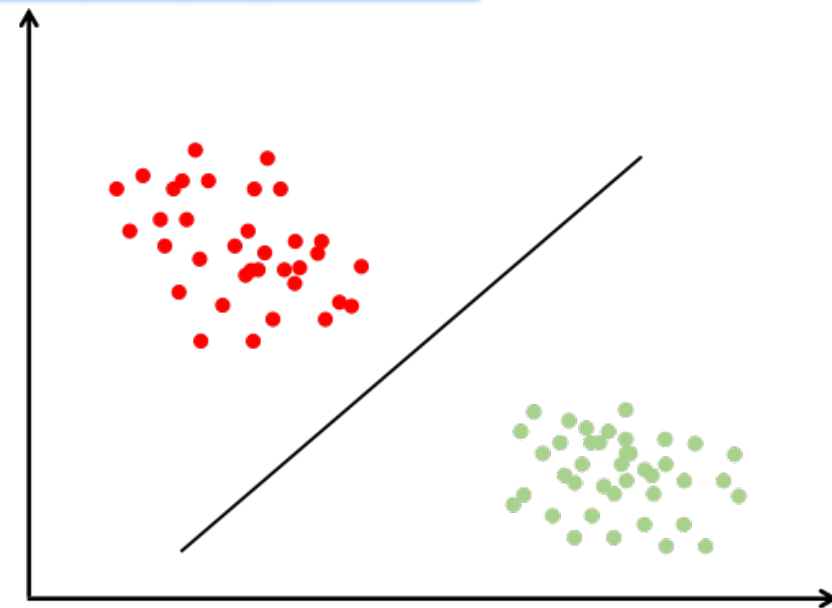
用一条直线可以将这两种类型分开，这条直线称为线性分类器。

直线方程为 $ax + by + c = 0$

落在直线上方的点被判定成樱桃 $ax + by + c < 0$

落在直线下方的点被判定成猕猴桃 $ax + by + c > 0$

这条直线就是训练得到的机器学习模型，直线方程的参数就是机器学习模型的参数



用直线将两类水果分开

机器学习的一个简单的例子

用一条直线可以将这两种类型分开，这条直线称为线性分类器。

一般化：给两类水果编号，称为类别标签，在这里樱桃类别标签为-1，猕猴桃类别标签为+1

判定规则可以写成决策函数
$$f(x, y) = \begin{cases} +1 & ax + by + c > 0 \\ -1 & ax + by + c \leq 0 \end{cases}$$

这个函数可以简写成符号函数 $\text{sgn}(ax + by + c)$ ，其中
$$\text{sgn}(x) = \begin{cases} +1 & x > 0 \\ -1 & x \leq 0 \end{cases}$$

问题：如何找到这条直线？

通过训练样本训练得到，保证对训练样本都能正确的分类。

机器学习的分类

- 1、基于机器学习三要素的分类
- 2、基于学习策略的分类
- 3、基于学习方法的分类
- 4、基于学习方式的分类
- 5、基于数据形式的分类
- 6、基于学习目标的分类

机器学习的分类——基于三要素的分类

根据机器学习三要素, 对机器学习进行分类:

机器学习的三个要素:

信息, 发现, 知识

- 基于信息: 机器学习可分为符号学习和数值学习;
- 基于知识的形式: 机器学习又可分为规则学习和函数学习等;
- 基于发现的逻辑方法: 则机器学习可分为归纳学习、演绎学习和类比学习等等。

机器学习的分类——基于学习策略的分类

模拟人脑的机器学习

- 符号学习
- 神经网络学习(或连接学习)

直接采用数学方法的机器学习

- 统计机器学习
 - 广义统计机器学习
 - 狭义统计机器学习

符号学习

符号学习

- 模拟人脑的宏观心理级学习过程。
- 以认知心理学原理为基础，以符号数据为输入，以符号运算为方法，用推理过程在图或状态空间中搜索，学习的目标为概念或规则等。

符号学习的典型方法：

- 记忆学习、示例学习、演绎学习、类比学习、解释学习等。

神经网络学习(或连接学习)

神经网络学习(或连接学习)

- 模拟人脑的微观生理级学习过程。
- 以脑和神经科学原理为基础,以人工神经网络为函数结构模型,以数值数据为输入,以数值运算为方法,用迭代过程在系数向量空间中搜索,学习的目标为函数。

典型的连接学习:

- 权值修正学习、拓扑结构学习等。

广义统计机器学习

广义统计机器学习

- 以样本数据为依据,以概率统计理论为基础,以数值运算为方法的一类机器学习。

广义统计学习分类:

- 以概率表达式函数为目标: 如贝叶斯学习、贝叶斯网络学习等。
- 以代数表达式函数为目标两大: 几何分类学习方法和支持向量机(SVM)。

机器学习的分类——基于学习方法的分类

归纳学习 □

- 符号归纳学习: 示例学习、决策树学习等。
- 函数归纳学习(发现学习): 神经网络学习、示例学习、发现学习、统计学习等

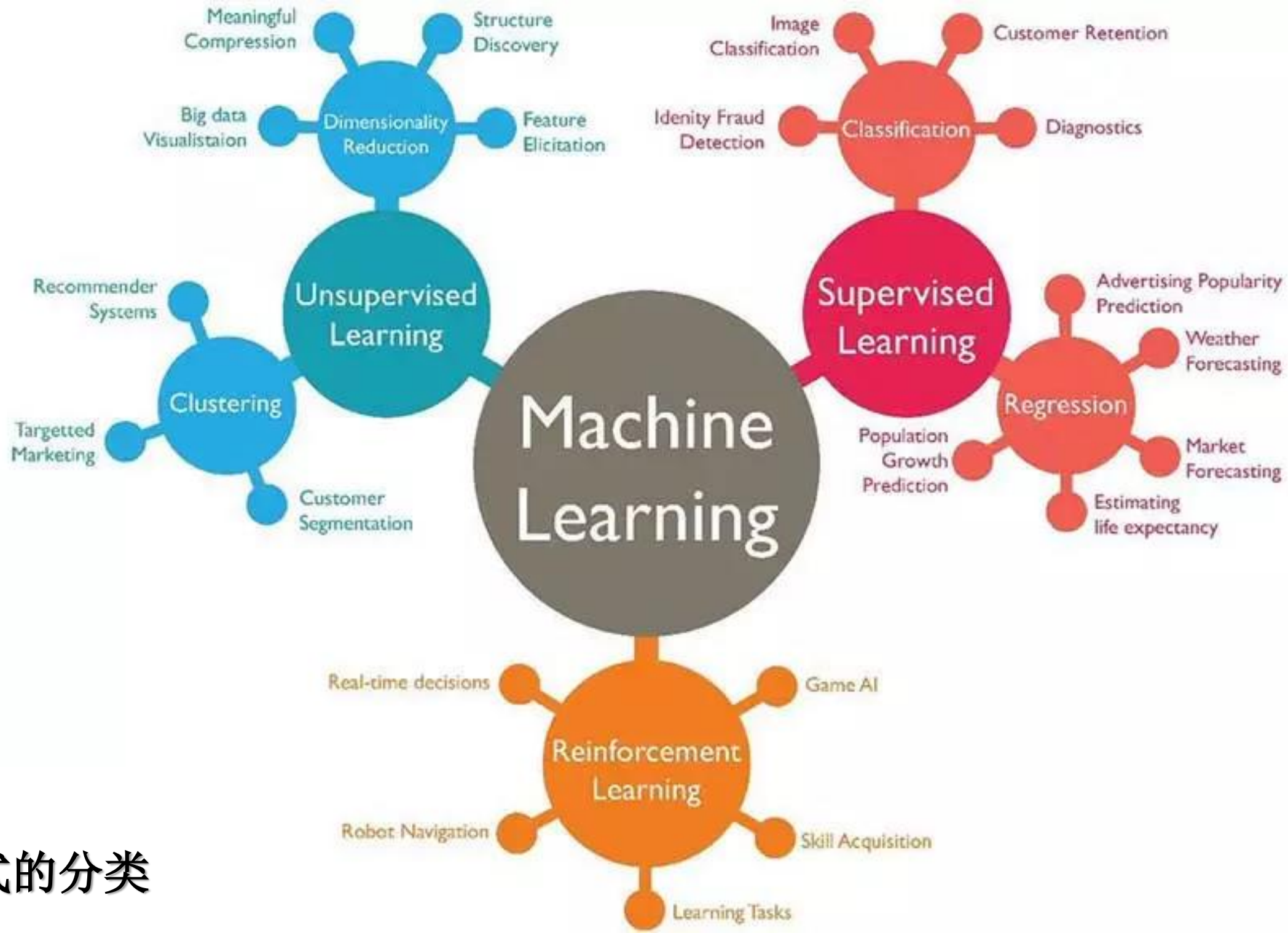
演绎学习

类比学习

- 案例(范例)学习。

分析学习

- 案例(范例)学习和解释学习等。



机器学习的分类

——基于学习方式的分类

有监督学习

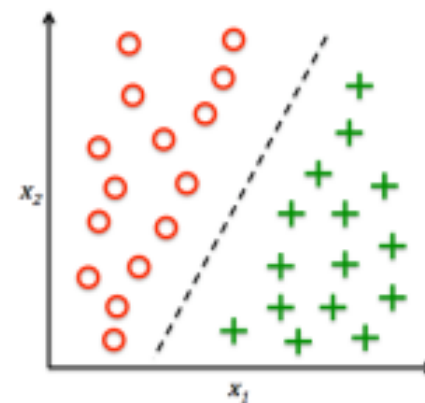
有导师学习(监督学习)

- 输入数据中有导师信号;
- 以概率函数、代数函数或人工神经网络为基函数模型;
- 采用迭代计算方法,学习结果为函数。



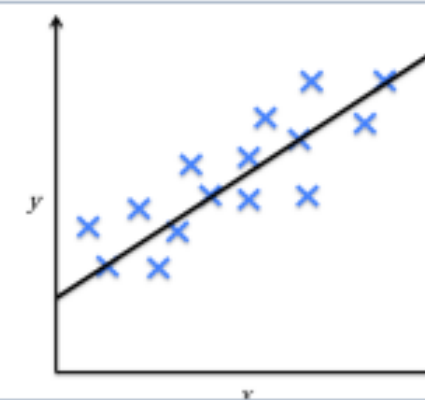
分类问题 (监督学习) :

- 根据数据样本上抽取出的特征, 判定其属于有限个类别中的哪一个
- 垃圾邮件识别 (结果类别: 1、垃圾邮件 2、正常邮件)
- 文本情感褒贬分析 (结果类别: 1、褒 2、贬)
- 图像内容识别识别 (结果类别: 1、喵星人 2、汪星人 3、人类 4、草泥马 5、都不是)



回归问题 (监督学习) :

- 根据数据样本上抽取出的特征, 预测连续值结果
- 《芳华》票房值
- 魔都房价具体值
- 刘德华和吴彦祖的具体颜值得分



无监督/强化学习

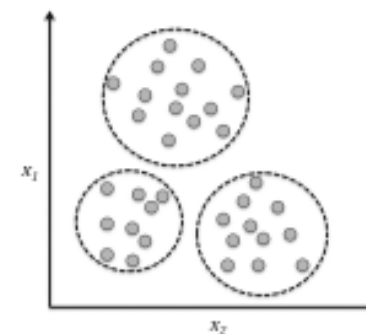
无导师学习(非监督学习)

- 输入数据中无导师信号;
- 采用聚类方法,学习结果为类别。
- 典型: 发现学习、聚类学习、竞争学习等。



聚类问题(无监督学习):

- 根据数据样本上抽取出的特征,挖掘数据的关联模式
- 相似用户挖掘/社区发现
- 新闻聚类



强化问题:

- 研究如何基于环境而行动,以取得最大化的预期利益
- 游戏(“吃鸡”)最高得分
- 机器人完成任务

强化学习(增强学习)

- 以环境反馈(奖/惩信号)作为输入;
- 以统计和动态规划技术为指导。

机器学习的分类——基于数据形式的分类

结构化学习

- 以结构化数据为输入；
- 以数值计算或符号推演为方法。
- 典型：神经网络学习、统计学习、决策树学习和规则学习。

非结构化学习

- 以非结构化数据为输入
- 典型：类比学习、案例学习、解释学习、文本挖掘、图像挖掘、Web挖掘等。

机器学习的分类——基于学习目标的分类

概念学习

- 学习的目标和结果为概念, 或者说是为了获得概念的一种学习。
- 典型的概念学习有示例学习。

规则学习

- 学习的目标和结果为规则, 或者说是为了获得规则的一种学习。
- 典型的规则学习有决策树学习。

函数学习

- 学习的目标和结果为函数, 或者说是为了获得函数的一种学习。
- 典型的函数学习有神经网络学习。

机器学习的分类——基于学习目标的分类2

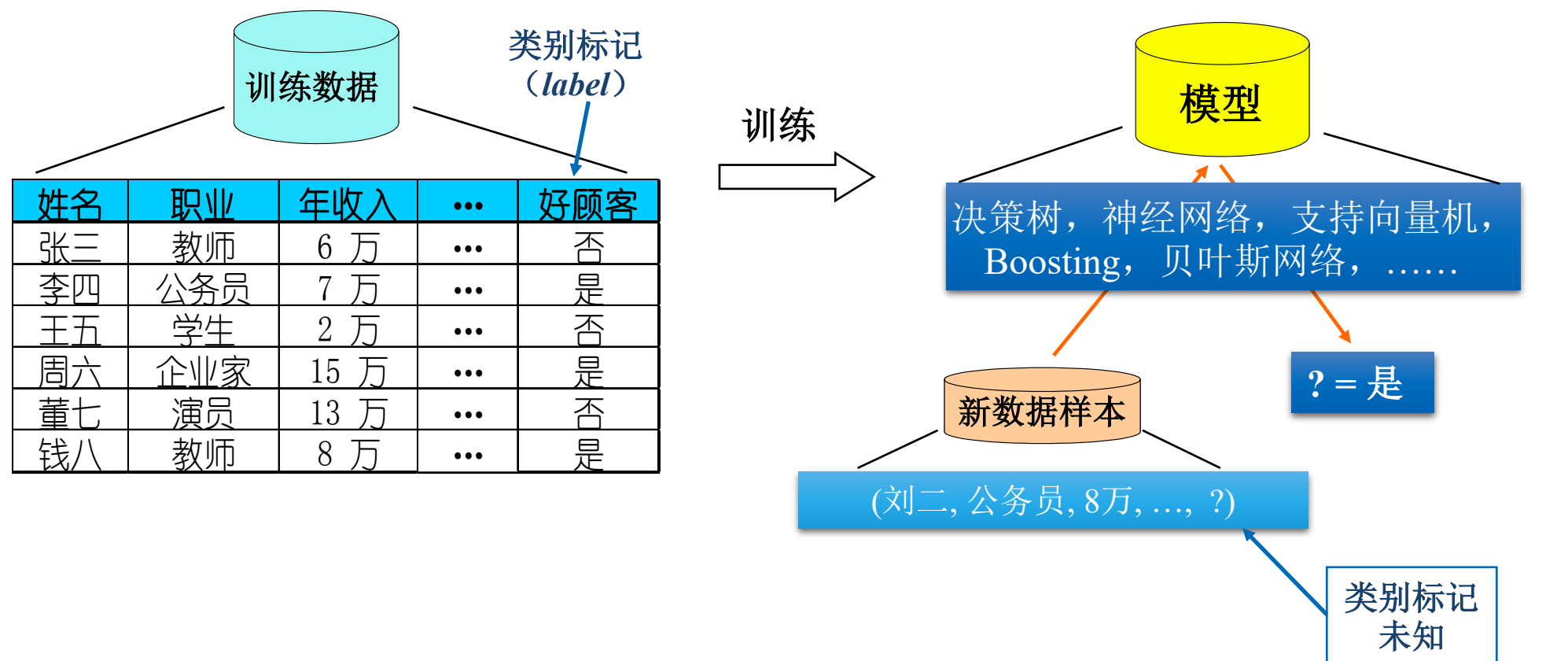
类别学习

- 学习的目标和结果为对象类, 或者说是为了获得类别的一种学习。
- 典型的类别学习有聚类分析。

贝叶斯网络学习

- 学习的目标和结果是贝叶斯网络, 或者说是为了获得贝叶斯网络的一种学习。
- 可分为结构学习和参数学习。

小结





什么是机器学习?

机器学习原理, 基本过程, 机器学习分类 ...



经典机器学习

简单机器学习、经典机器学习, ...



决策树学习

什么是决策树, 学习步骤, ...



神经网络学习

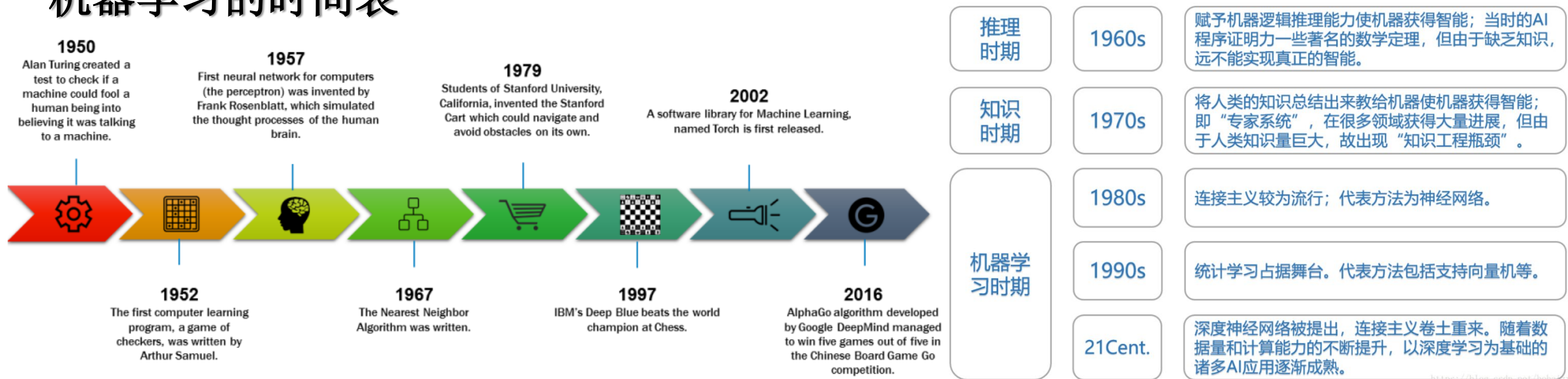
人工神经元, 人工神经网络发展历史, 感知器, BP, ...



深度学习

什么是深度学习, 深度学习的发展历史, 卷积神经网络, 常见深度学习网络, ...

机器学习的时间表

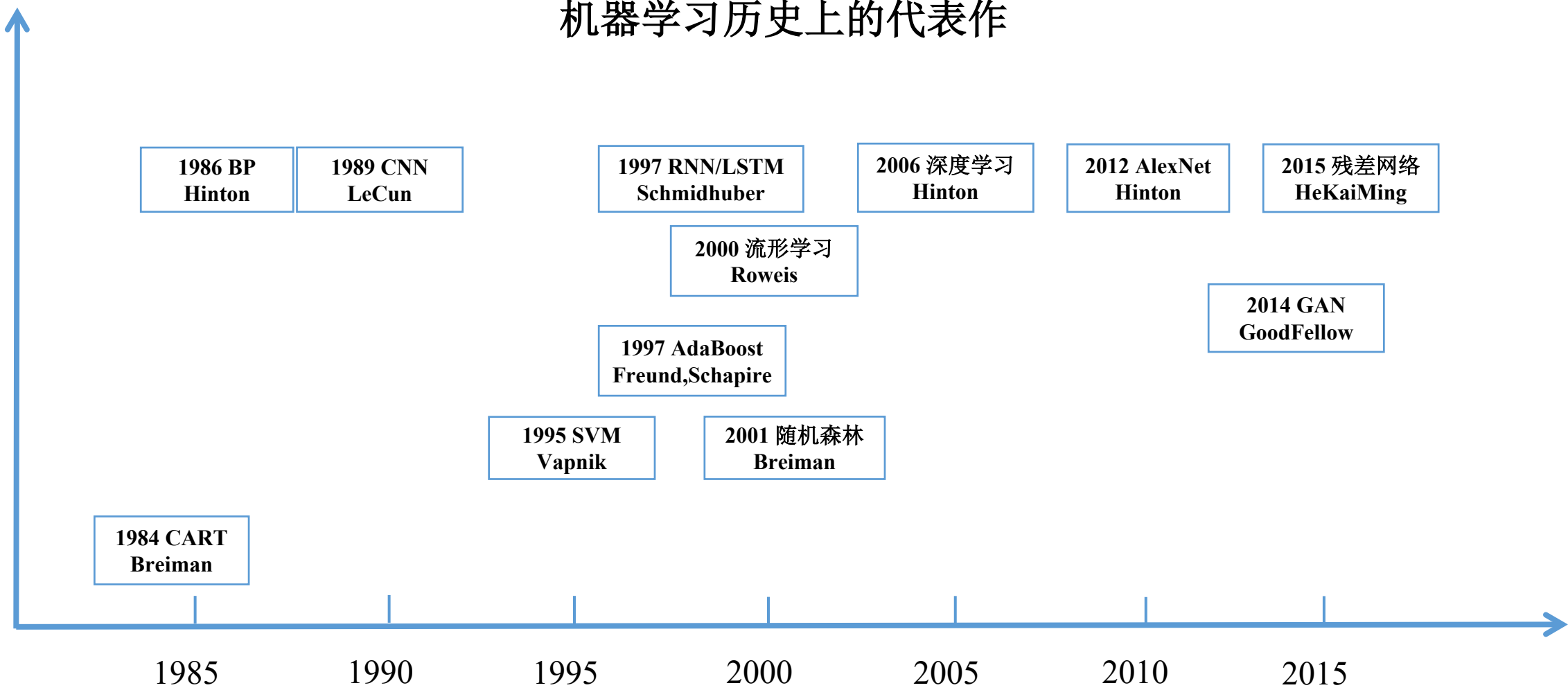


机器学习的发展历程大致可以分为三个阶段：

- 1980s 早期，成为独立的学科，诞生了决策树、反向传播算法，浅层神经网络等经典算法。
- 1990s 蓬勃发展期，2次获得图灵奖。经典算法:支持向量机，随机森林，Boosting，RNN/LSTM，流形学习，聚类。
- 2012-至今 深度学习时代，卷积神经网络，深度RNN，强化学习，GAN，图神经网络，自动化机器学习/NAS等

机器学习的时间表

机器学习历史上的代表作



简单机器学习——记忆学习

记忆学习也称死记硬背学习或机械学习。

该学习方法不要求系统具有对复杂问题求解的能力, 即没有推理能力。

例: 对于某数据 x ,

- 经过某种计算过程得到结果 y ;
- 系统将 (x, y) 作为联想对存储起来;
- 以后再要对 x 作同样的计算时, 就可通过查询(而不是计算)直接得到 y 。

简单机器学习——记忆学习

记忆学习基于记忆和检索

学习方法很简单, 但学习系统需要几种能力:

- 能实现有组织的存储信息: 应利于检索
- 能进行信息综合: 减少冗余数据
- 能控制检索方向: 减少检索次数

简单机器学习——示例学习

示例学习（实例学习）是一种归纳学习。

从若干实例(包括正例和反例)中归纳出一般概念或规则的学习方法。

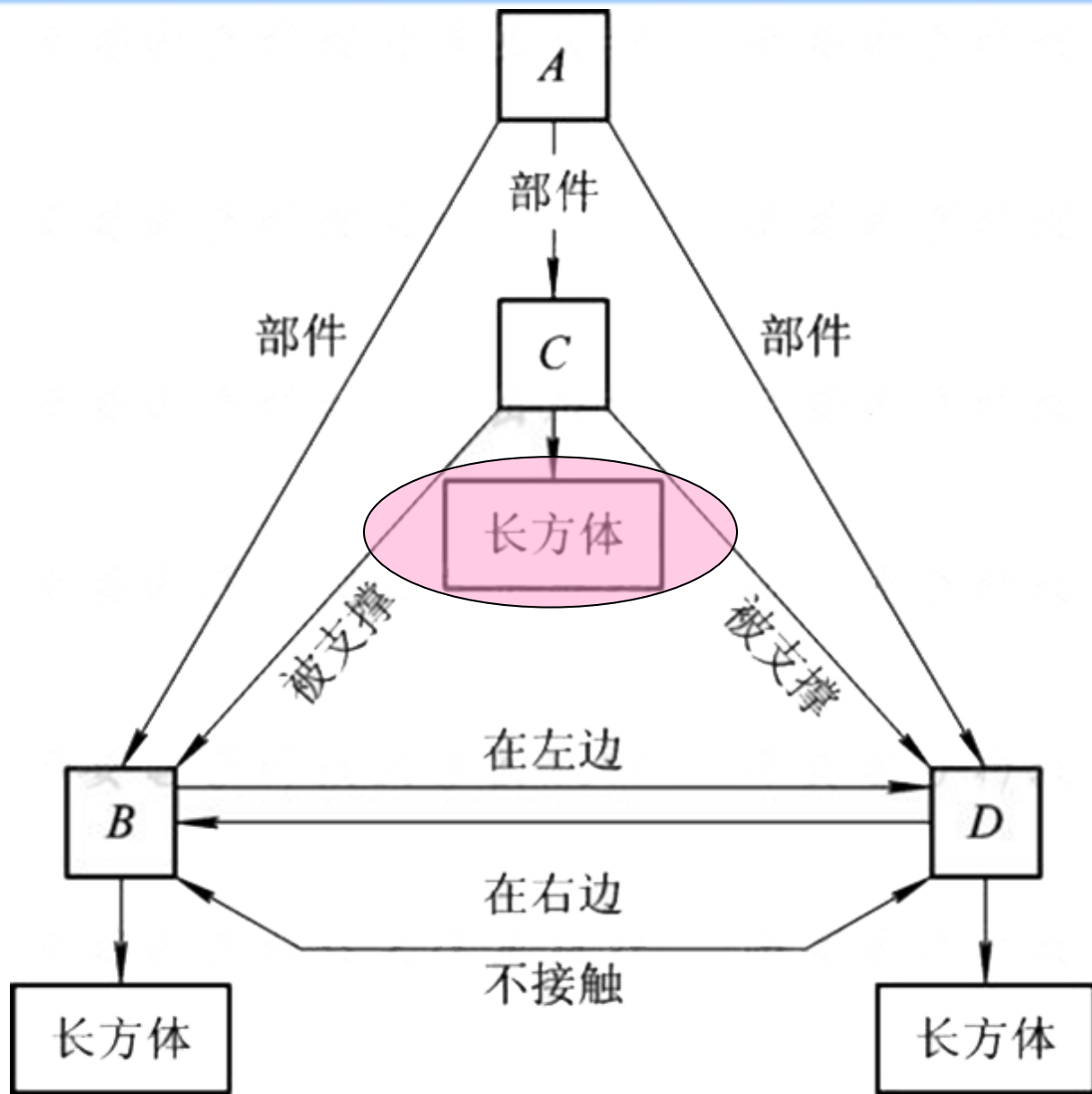
例：学习“狗”的概念。

可先提供给程序各种动物, 并告知程序哪些动物是“狗”, 哪些不是“狗”, 系统学习后便概括出“狗”的概念模型或类型定义, 利用这个类型定义作为动物世界中识别“狗”的标准的准则。

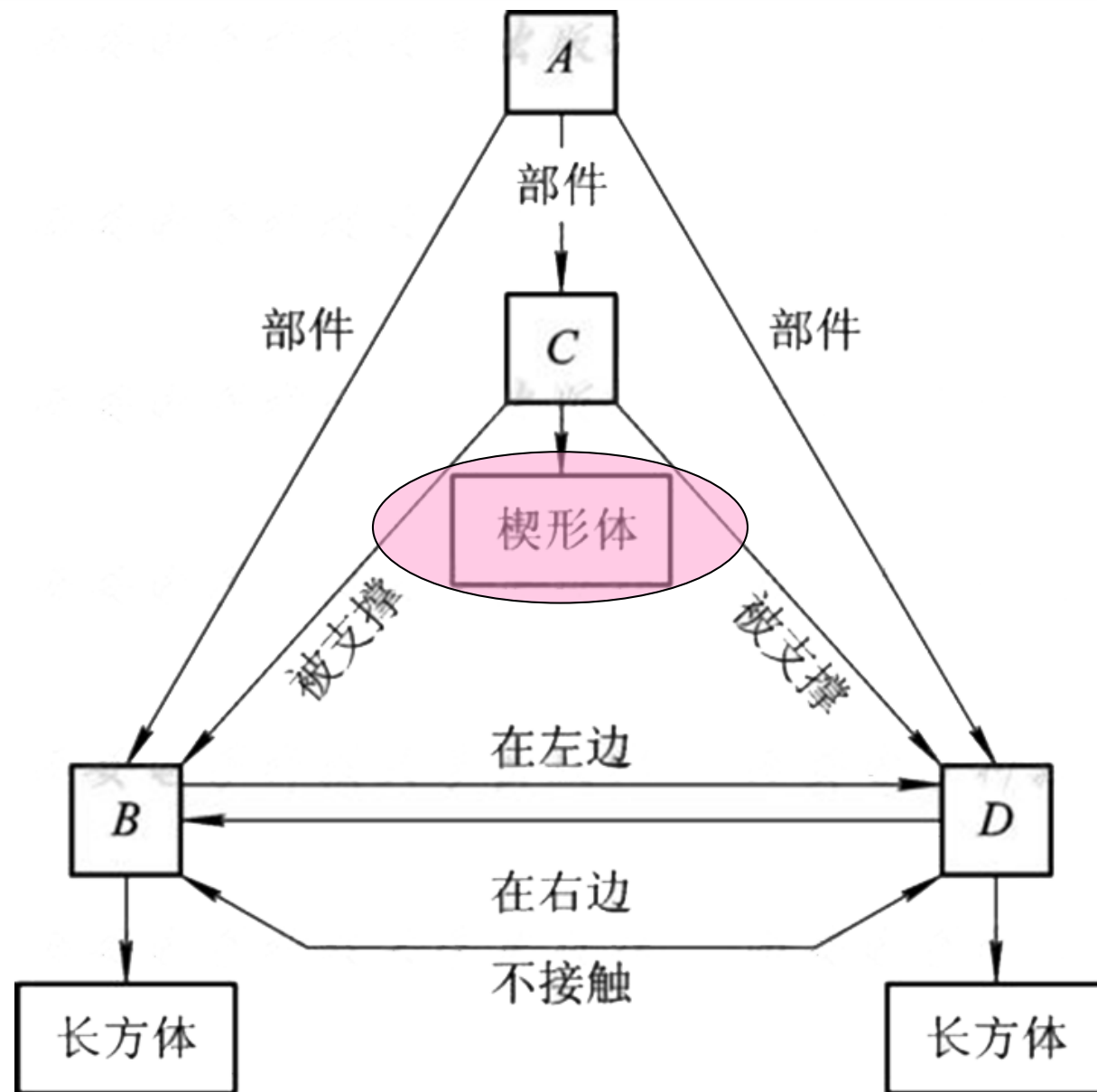
Winston在1975年提出的结构化概念学习方法。

系统需要学习**拱桥**概念, 通过搭积木实现示例学习。

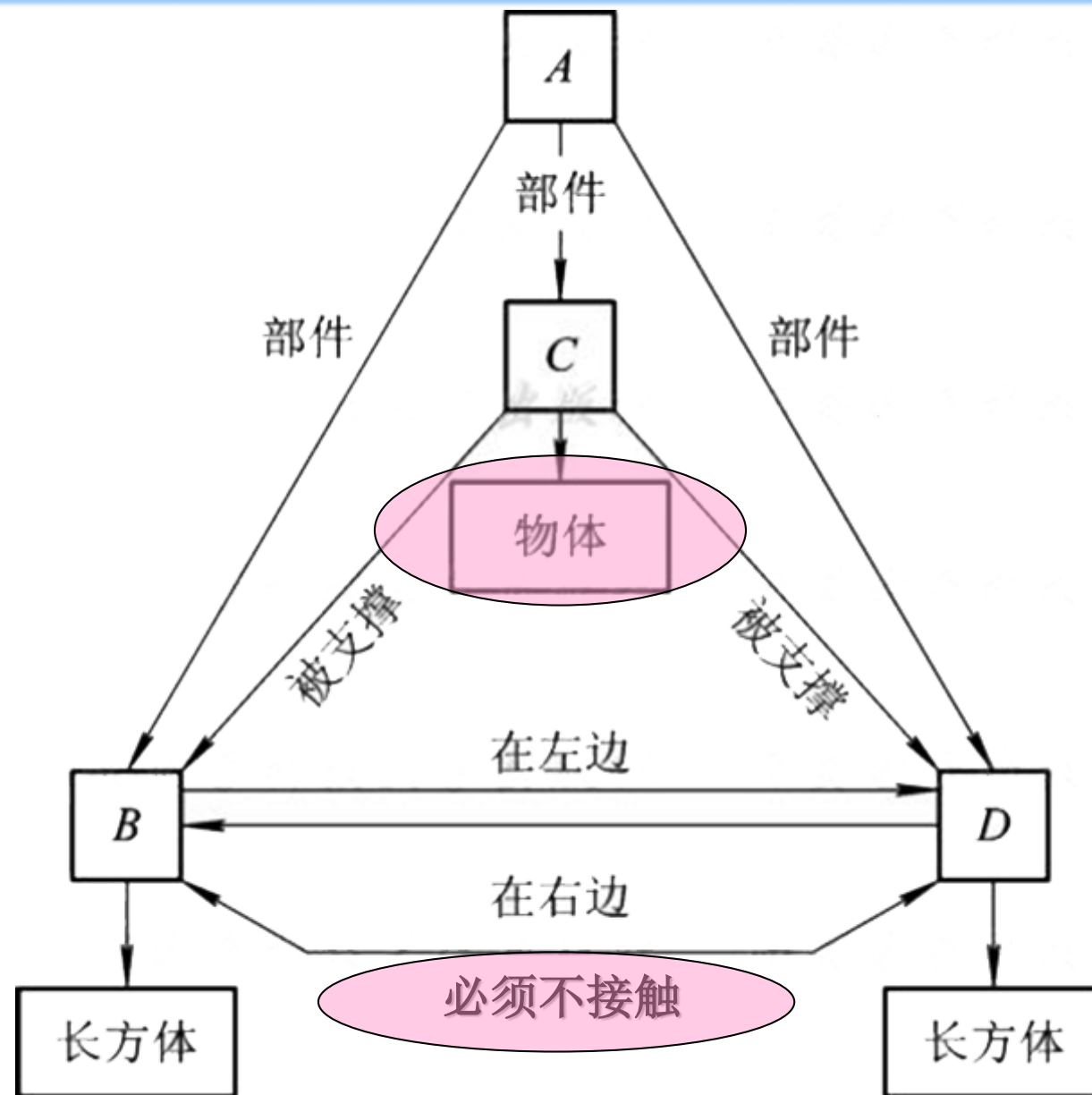
第一个拱桥的语义网络



第二个拱桥的语义网络



学习程序归纳出的语义网络



例“同花”概念学习

示例1:

花色(c1, 梅花) \wedge 花色(c2, 梅花) \wedge 花色(c3, 梅花) \wedge 花色(c4, 梅花) \rightarrow 同花(c1, c2, c3, c4)

示例2:

花色(c1, 红桃) \wedge 花色(c2, 红桃) \wedge 花色(c3, 红桃) \wedge 花色(c4, 红桃) \rightarrow 同花(c1, c2, c3, c4)

...

归纳推理后规则:

花色(c1, x) \wedge 花色(c2, x) \wedge 花色(c3, x) \wedge 花色(c4, x) \rightarrow 同花(c1, c2, c3, c4)

例 曲线拟合

假设示例空间存放有如下的三个示例：

示例1: $(0, 2, 7)$ 示例2: $(6, -1, 10)$ 示例3: $(-1, -5, -10)$

这三个3维向量表示空间中的三个点。

求出过这三点的曲线。

归纳出规则：

- $(x, y, 2x+3y+1); z=2x+3y+1$

经典机器学习

- 1、线性回归
- 2、逻辑回归
- 3、线性判别分析
- 4、决策树
- 5、朴素贝叶斯学习
- 6、K均值聚类
- 7、支持向量机
- ...

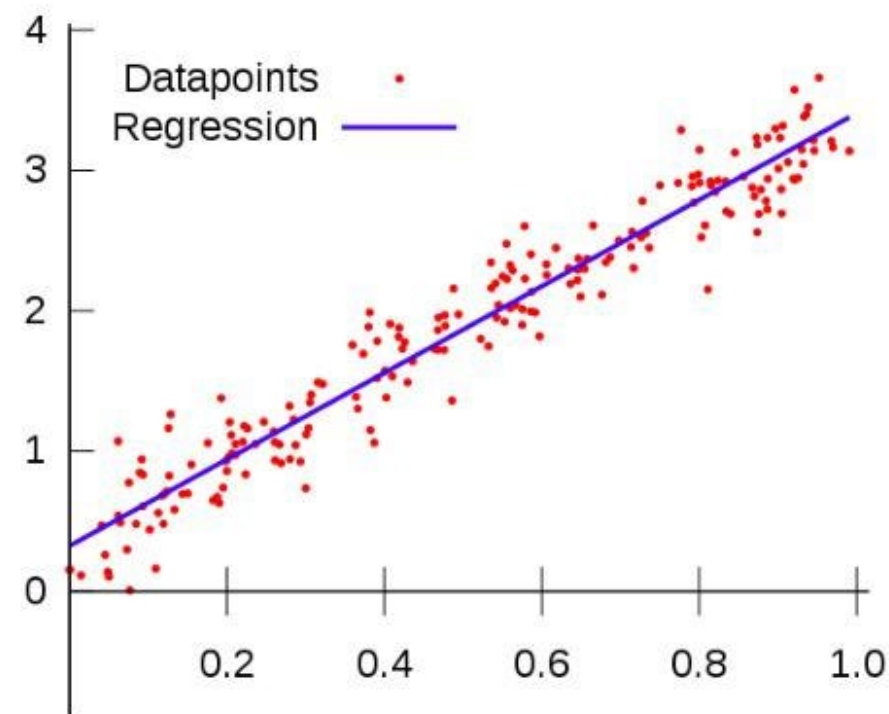
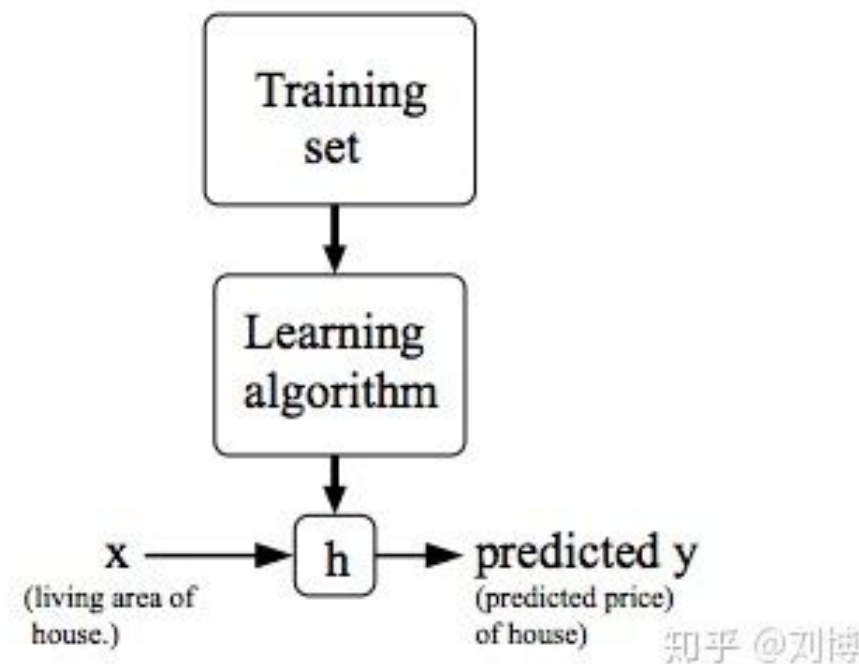
经典机器学习——线性回归

在统计学和机器学习领域，线性回归是最广为人知也最易理解的算法之一。

我们的目标是“学习”得到一个函数 $h: x \rightarrow y$, 使 $h(x)$ 是真实值 y 的一个“好的”预测值。这里 h 叫做模型，也叫做**假设 (hypothesis)**。

如果我们要预测的输出值是**连续的**，那么该问题就称作**回归问题**。

回归的过程就是要**找到最优的模型来描述数据**。



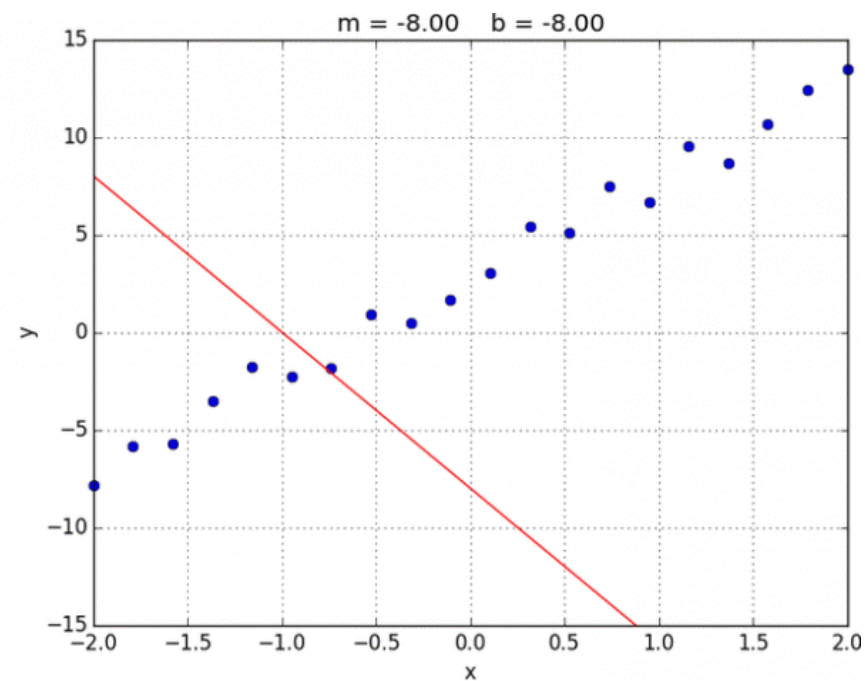
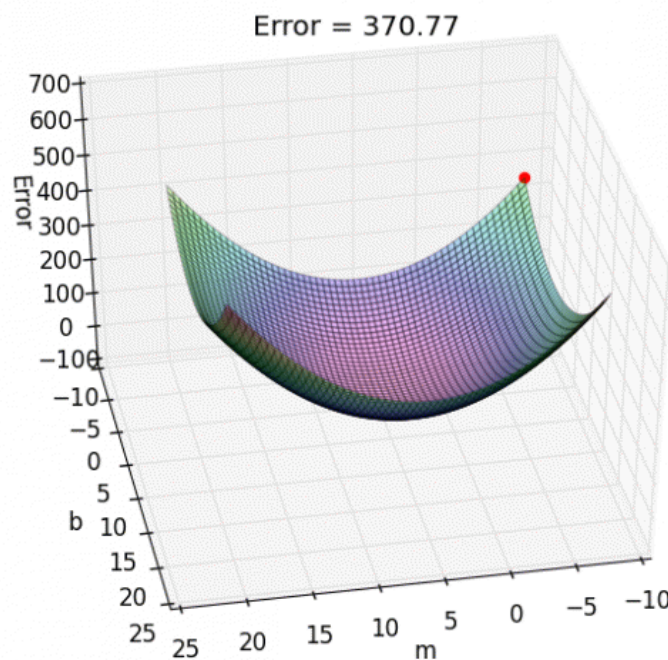
经典机器学习——线性回归

一元线性回归：简单线性回归模型
模型中只含有一个自变量，线性数学表达式

$$y=a+bx+r$$

a 、 b 为回归系数， r 为模型误差项。

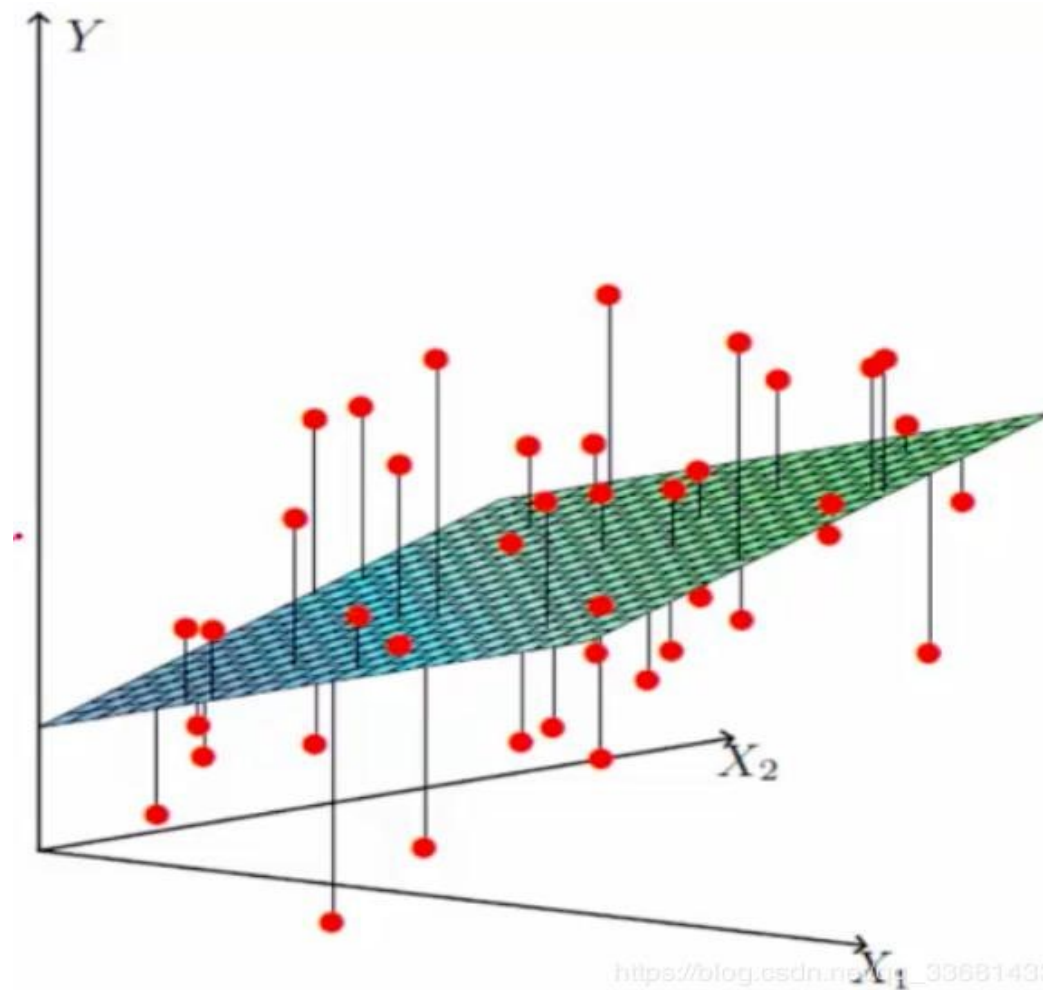
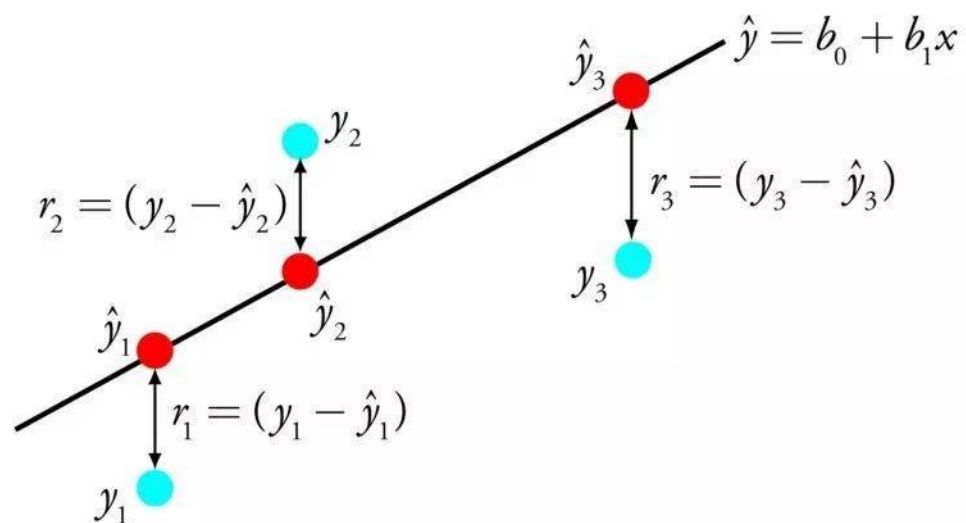
要得到理想的拟合线，则要使误差，总体上最小，于是转换成了误差平方和最小的问题，此方法就是“最小二乘法”。



经典机器学习——线性回归

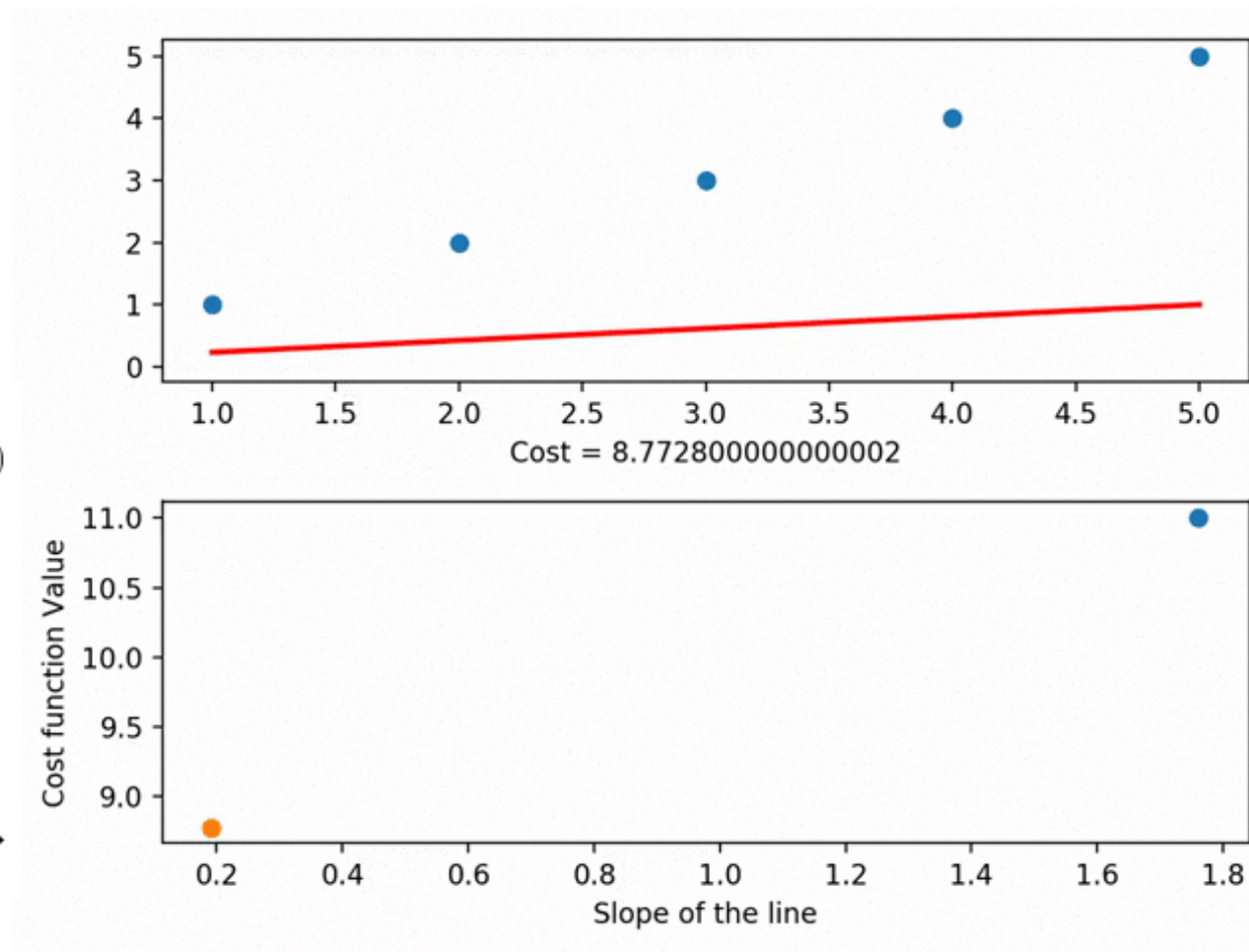
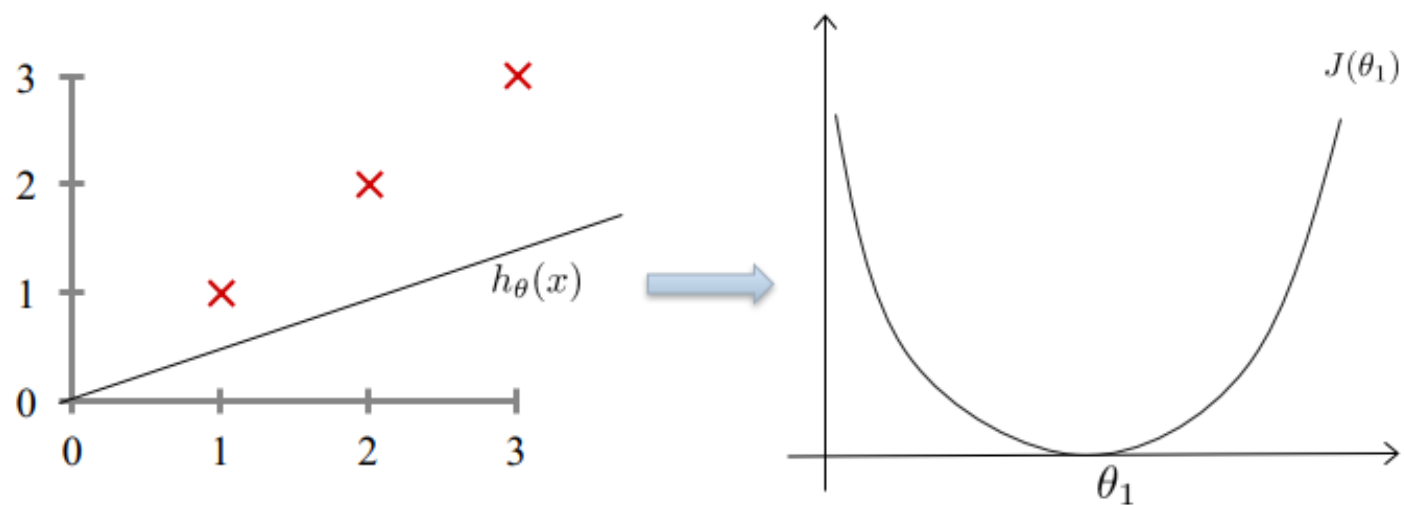
损失函数定义

$$\sum_i |\hat{y}_i - y_i| \quad \text{或} \quad \sum_i (\hat{y}_i - y_i)^2$$



经典机器学习——线性回归

损失函数是一个凸函数，优化的目标是达到最低点，也就是使得损失函数最小。



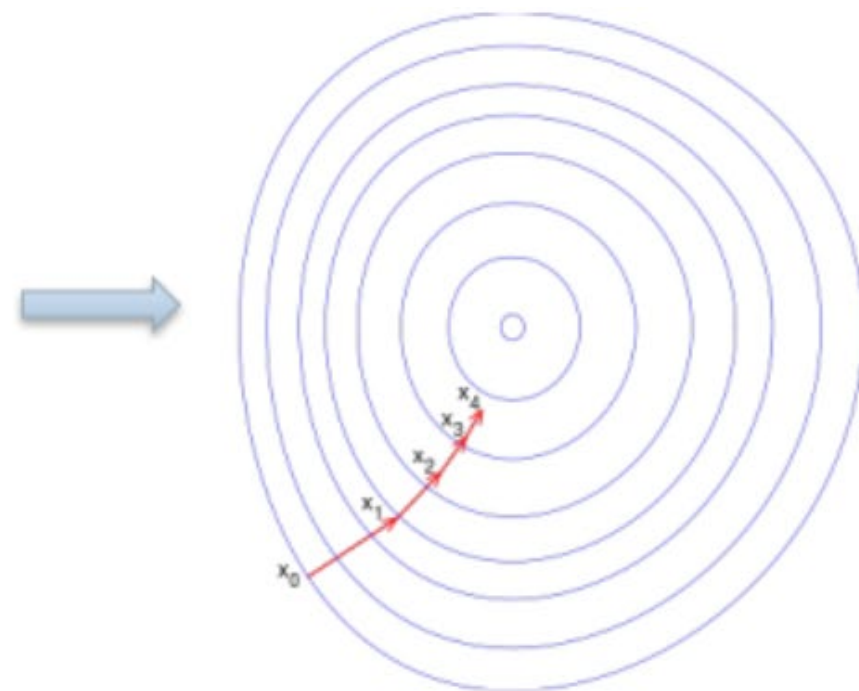
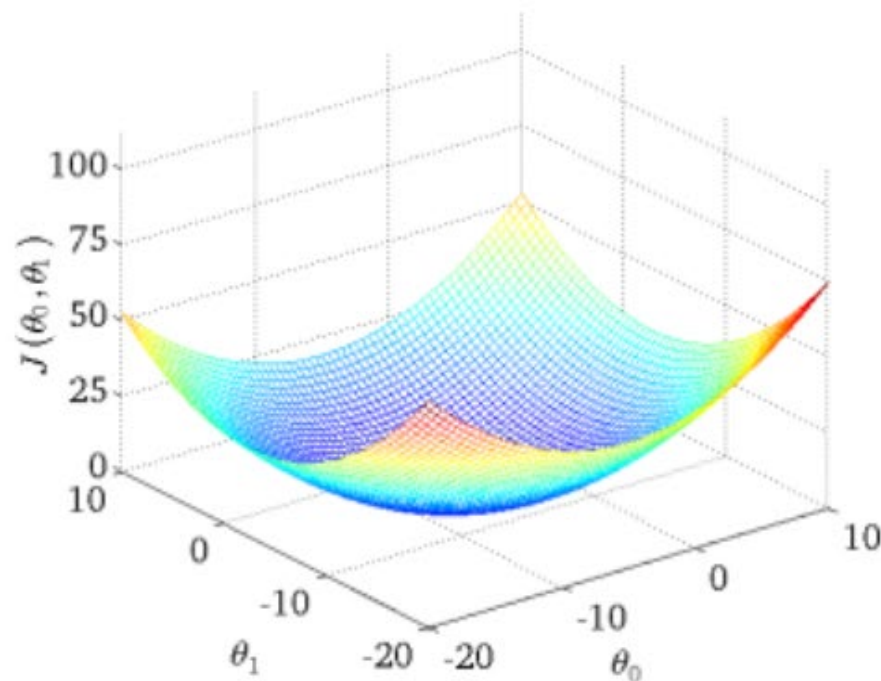
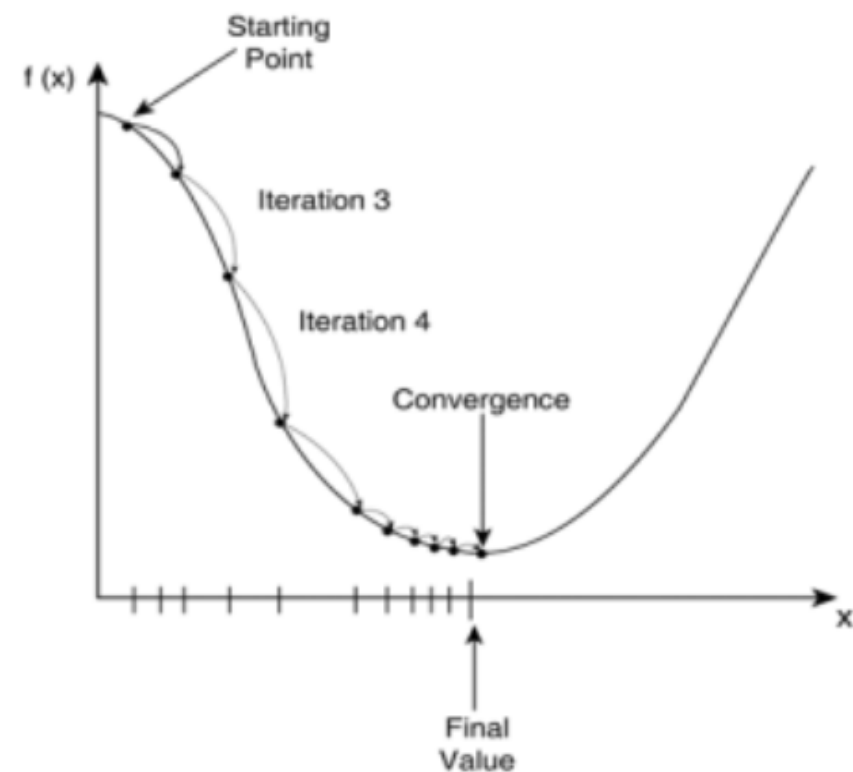
经典机器学习——线性回归

梯度下降：

逐步最小化损失函数的过程

如同下山，找准方向（斜率），每次迈进一小步，直至山底。

$$\theta_1 := \theta_1 - \alpha \frac{d}{d\theta_1} J(\theta_1)$$



经典机器学习——线性回归

多元线性回归

$$y_i = \beta_0 1 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

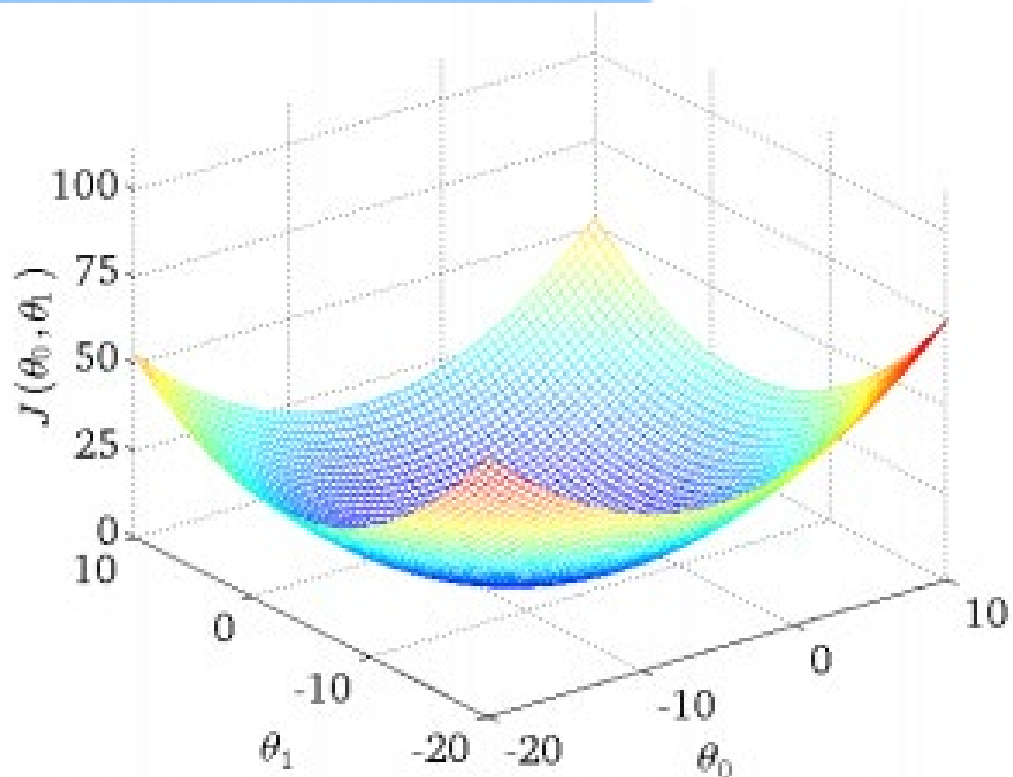
$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n \times 1} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}_{n \times p} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}_{p \times 1} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{n \times 1}$$

自变量个数为 p ，观测值个数为 n ，也可简化表示为：

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$\boldsymbol{\beta}$ 为偏回归系数向量， $\boldsymbol{\varepsilon}$ 为误差向量。找到最优的 $\boldsymbol{\beta}$ 和 $\boldsymbol{\varepsilon}$ ，就找到了最优的模型。

$$(\beta^*, \varepsilon^*) = \operatorname{argmin} \sum_{i=1}^m (f(x_i) - y_i)^2$$



经典机器学习——线性回归

从线性回归推广到广义线性回归

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\ln(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

左边是线性回归模型，右边是对数线性回归模型（log-Linear Regression）

对数线性回归实质上是完成输入空间 \mathbf{X} 到输出空间 \mathbf{y} 的非线性映射。

广义线性回归模型（GLM, Generalized Linear Model）

$$\mathbf{y} = g^{-1}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})$$

经典机器学习——Logistic回归

逻辑回归（Logistic Regression）是一种广义的线性回归分析模型。

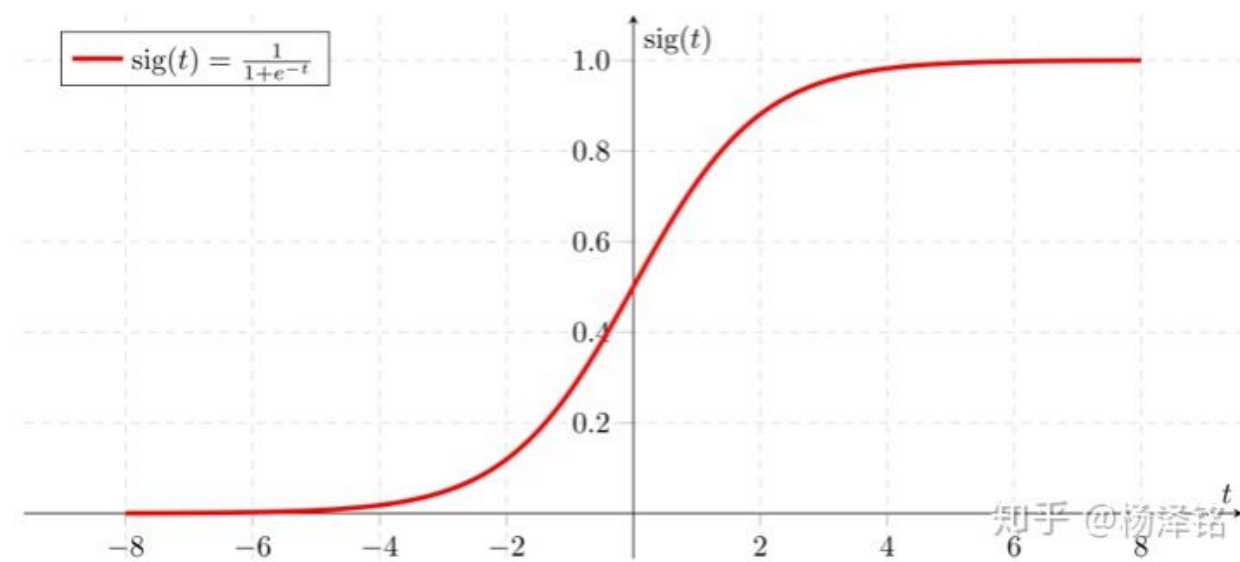
逻辑回归虽然名字里有“回归”，但它是最为经典的分类算法之一

逻辑回归使用Sigmoid函数（S形函数），即对数几率函数（Logistic Function）

函数原型： $y = \frac{1}{1 + e^{-z}}$ 或： $\ln \frac{y}{1 - y} = z$

逻辑回归的数学原型：

$$y = \frac{1}{1 + e^{-(\beta X + \epsilon)}}$$



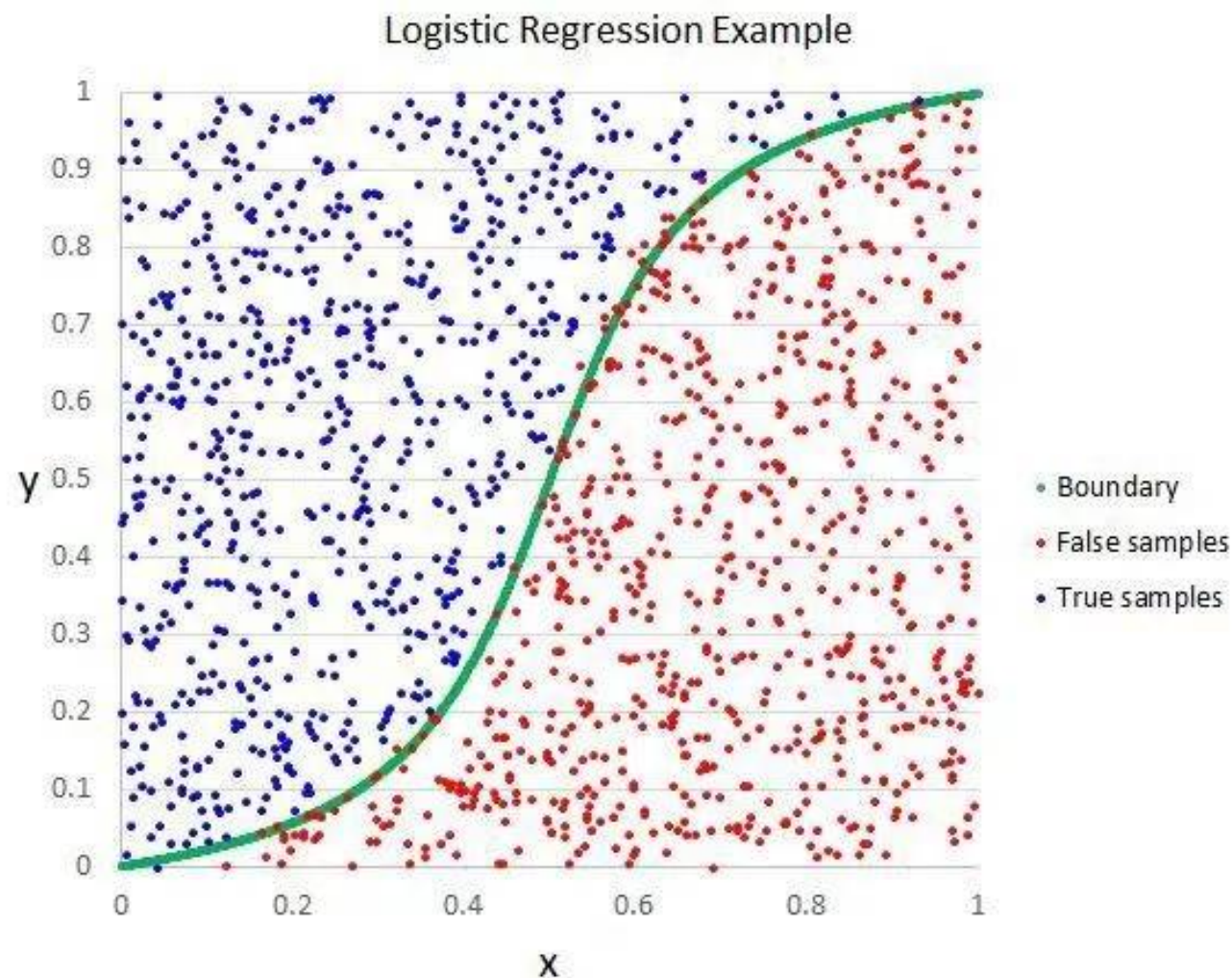
经典机器学习——Logistic回归

Logistic回归虽然名字里有“回归”，但实际上是一种分类方法，用于0/1二分类问题。

分类的本质：在空间中找到一个决策边界来完成分类的决策。

线性回归可以预测连续值，但是不能解决分类问题。

逻辑回归就是将线性回归的 $(-\infty, +\infty)$ 结果，通过Sigmoid函数映射到 $(0,1)$ 之间，从而判定其属于正类还是负类。



经典机器学习——Logistic回归

为什么用**Sigmoid**函数？

① 可将 $(-\infty, +\infty)$ 结果，映射到 $(0, 1)$ 间，作为概率。

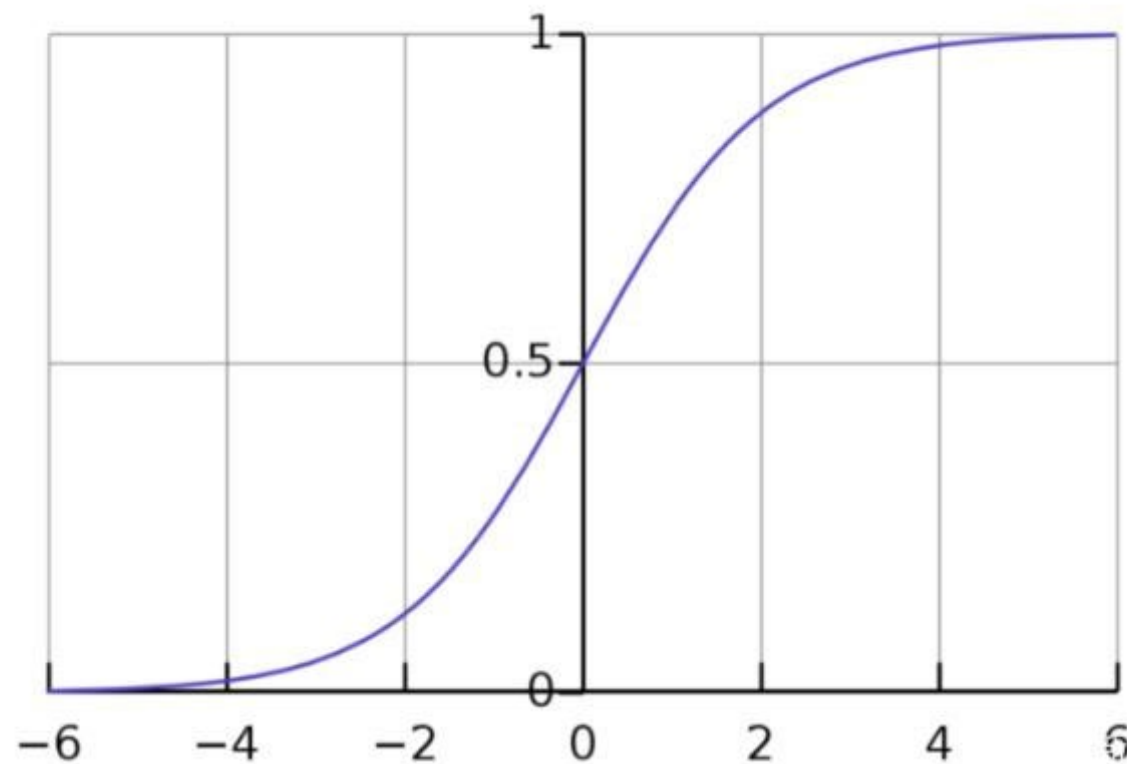
② 当 $x < 0$ 时， $\text{sigmoid}(x) < 0.5$

当 $x > 0$ 时， $\text{sigmoid}(x) > 0.5$

因此可将 0.5 作为分类的决策边界

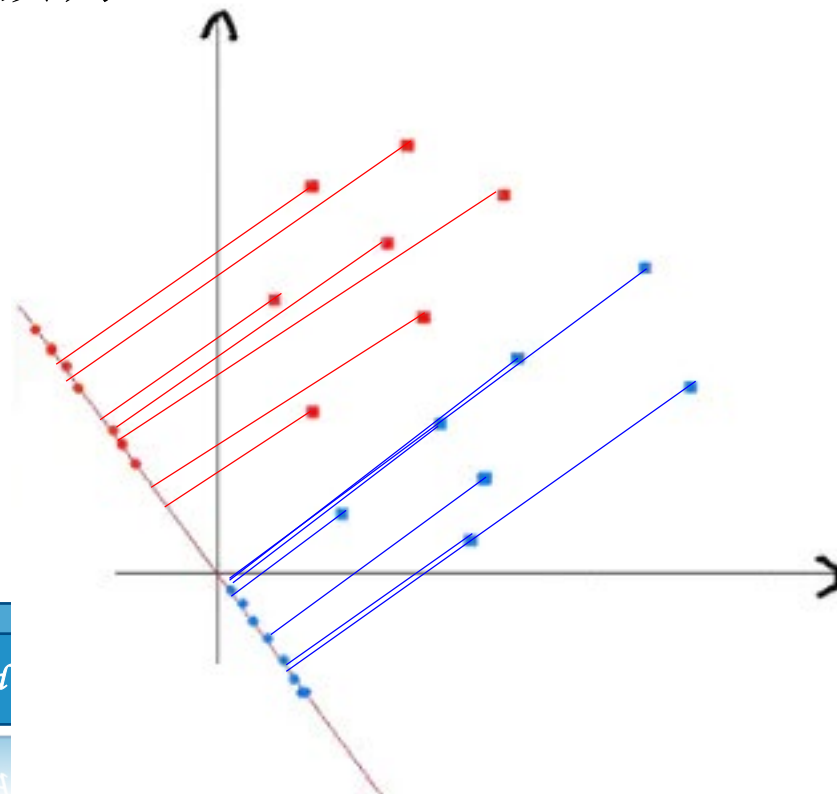
③ 数学特性好，求导容易：

$$s'(x) = s(x) \cdot (1 - s(x))$$



经典机器学习——线性判别分析(LDA)

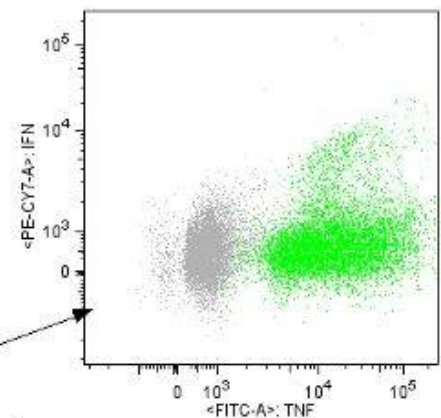
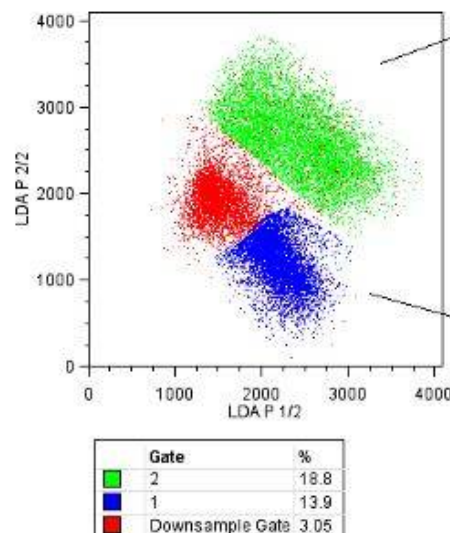
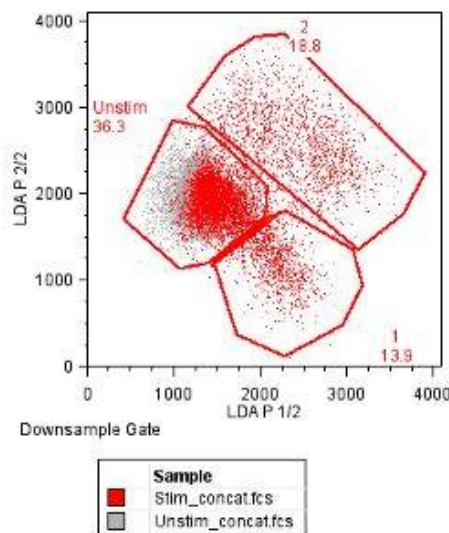
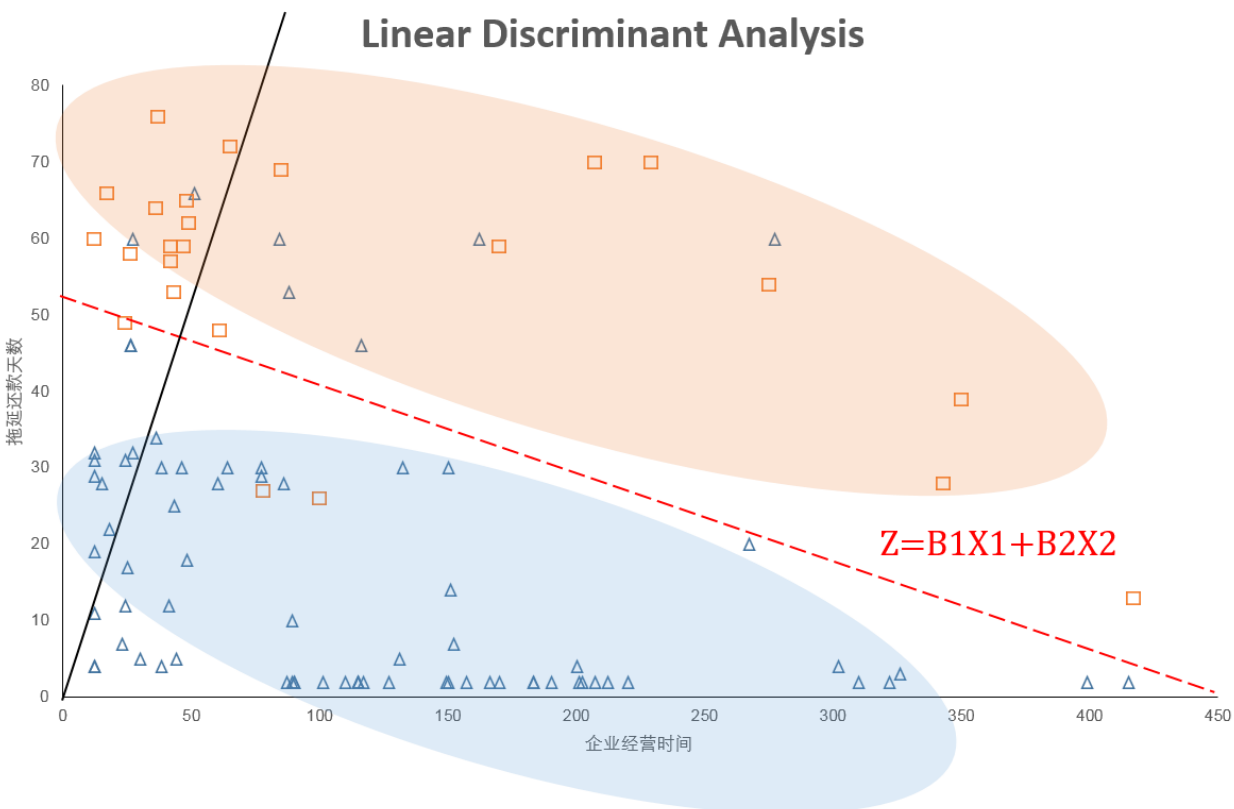
- ※ Logistic 回归是一种传统的分类算法，它的使用场景仅限于二分类问题。
- ※ 对于两个以上的类，线性判别分析算法（LDA）是首选的线性分类技术。
- ※ LDA通过对历史数据进行投影，以保证投影后同一类别的数据尽量靠近，不同类别的数据尽量分开。并生成线性判别模型对新生成的数据进行分离和预测。
- ※ LDA目标是给出一个标注了类别的数据集，投影到了一条直线之后，能够使得点尽量按类别区分开。
- ※ 红色方形点和蓝色方形点为2类数据点，经过原点的线就是投影直线，可以看到，红色点和蓝色点被原点分开。



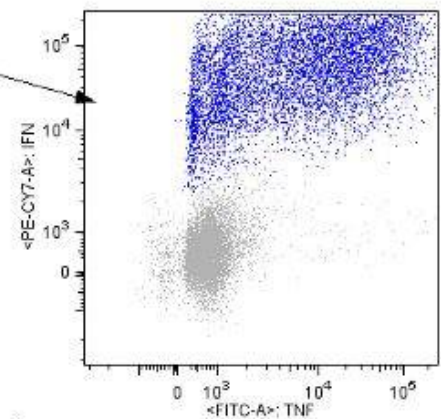
经典机器学习——线性判别分析(LDA)

LDA 表示方法非常直接，即每个类的数据统计属性，如均值和方差。

预测结果是通过计算每个类的判别值、并将类别预测为判别值最大的类而得出的。



Sample	Gate	%
Unstim_concat.fcs	Downsample Gate	2.54
Stim_concat.fcs	1	13.9

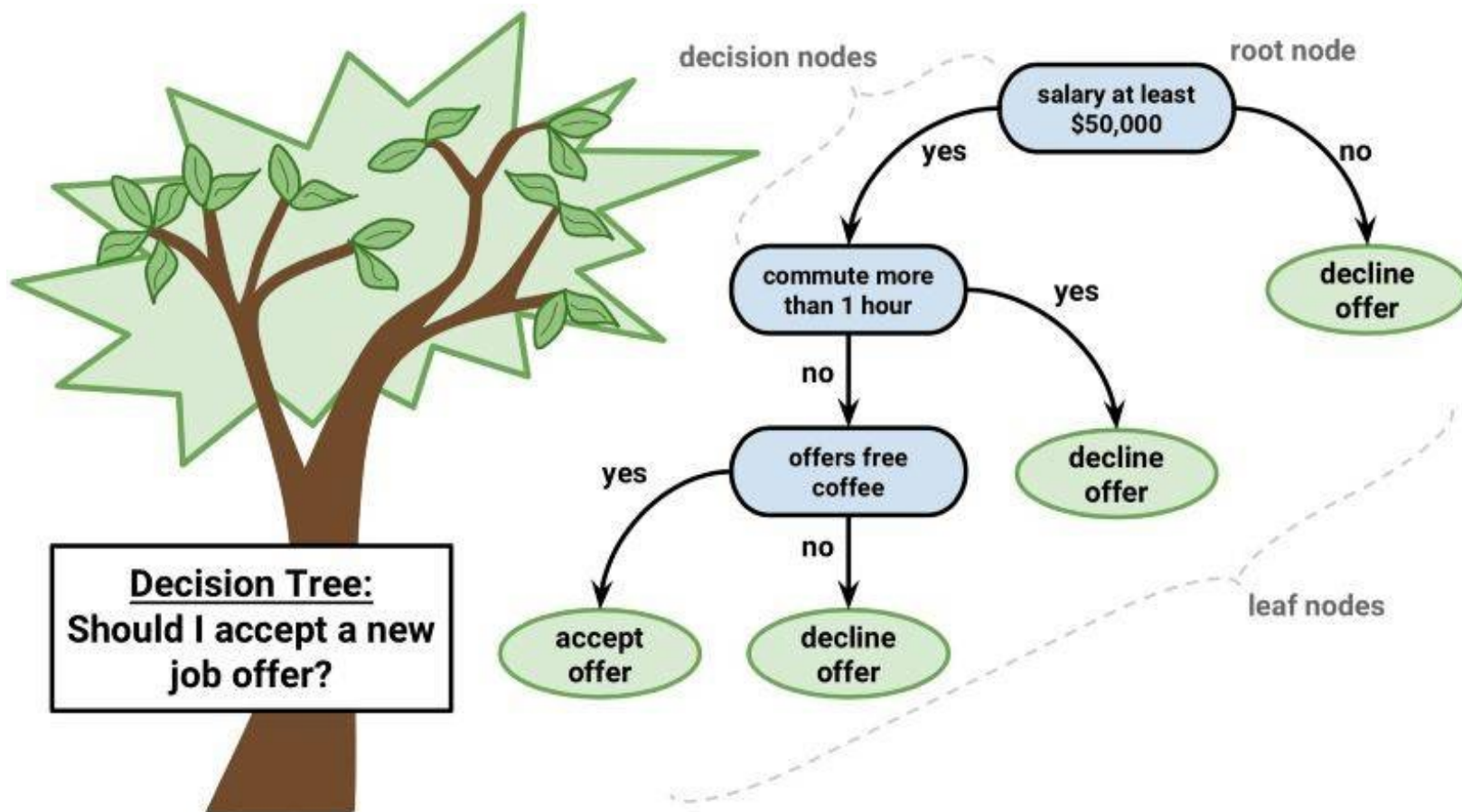


Sample	Gate	%
Unstim_concat.fcs	Downsample Gate	2.54
Stim_concat.fcs	2	18.8

经典机器学习——决策树学习

决策树(Decision Tree)

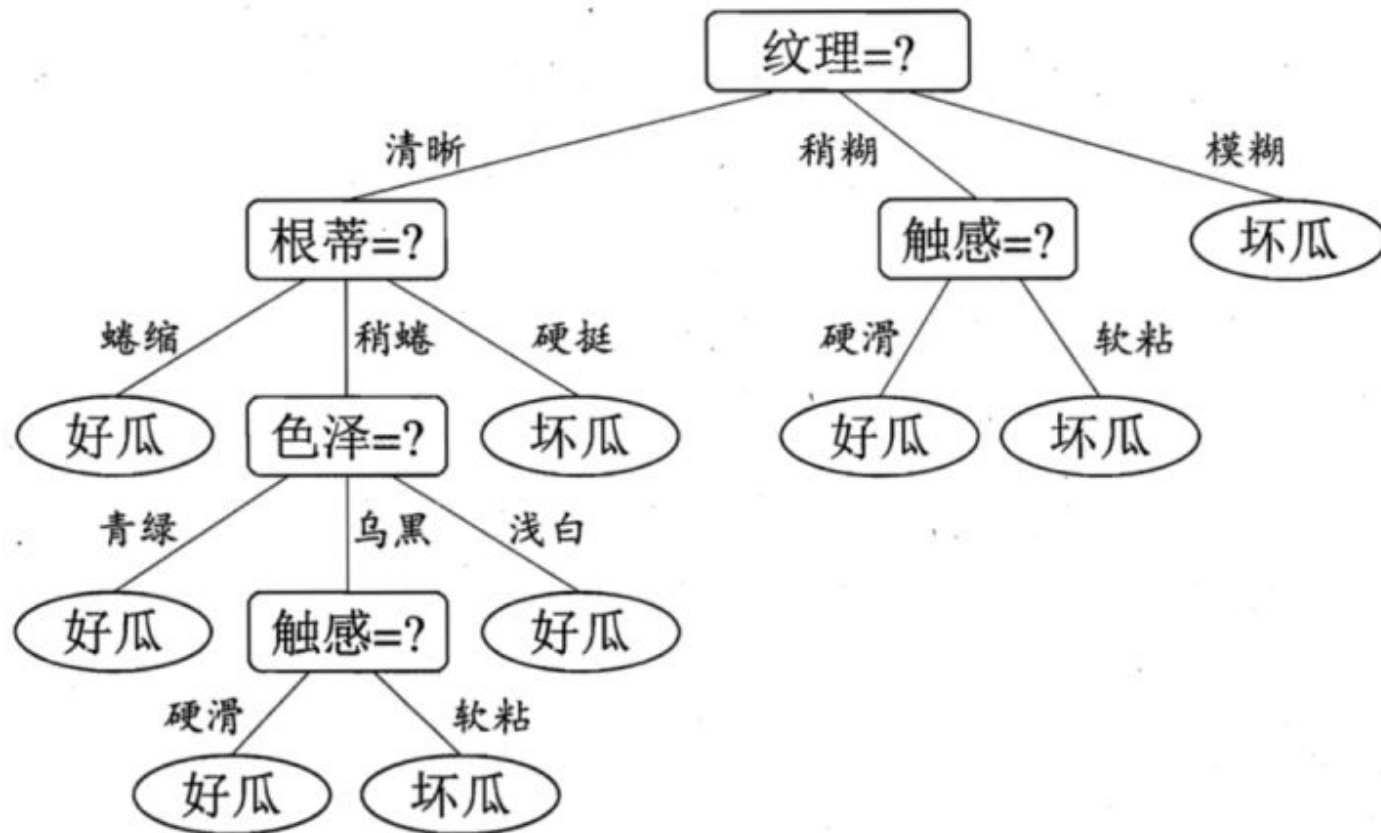
- 决策树模型呈树形结构。
- 在分类中，可看作是if-then规则的集合。
- 也可看作是定义在特征空间与类空间上的条件概率分布。
- 主要优点是模型具有可读性，分类速度快。



经典机器学习——决策树学习

决策树结构

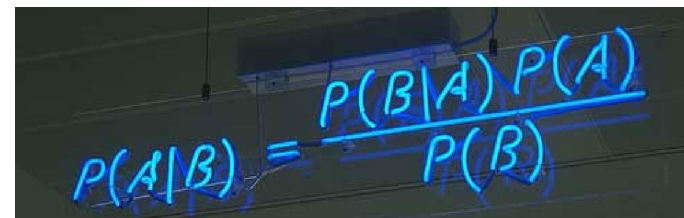
- 每个结点表示在一个属性上的测试。
- 每个分支代表一个测试输出。
- 每个叶结点代表一种类别。



决策树学习采用的是自顶向下的递归方法，其基本思想是以信息熵为度量构造一棵熵值下降最快的树。到叶子节点的处的熵值为零，此时每个叶结点中的实例都属于同一类。

经典机器学习——朴素贝叶斯学习 (Naive Bayes)

1) 朴素贝叶斯算法 (Naive Bayes) 是一种简单而强大的预测建模算法，广泛应用于文本分类、垃圾邮件过滤器、医疗诊断等。


$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

2) 朴素贝叶斯学习基于概率论的贝叶斯定理。

3) 之所以被称为「朴素」，是因为它假设每个输入变量相互之间是独立的。

优点：简单易懂、学习效率高、在某些领域的分类问题中能与决策树、神经网络相媲美。

缺点：由于该算法以自变量之间的独立（条件特征独立）性和连续变量的正态性假设为前提，就会导致算法精度在某种程度上受影响。

经典机器学习——朴素贝叶斯学习 (Naive Bayes)

贝叶斯定理:

$$P(A_k | B_1 B_2 \dots B_n) = \frac{P(A_k) P(B_1 B_2 \dots B_n | A_k)}{P(B_1 B_2 \dots B_n)}$$

对于 $P(A_k | B_1 B_2 \dots B_n)$ ，分母 $P(B_1 B_2 \dots B_n)$ 为固定值。

因为我们只需要比较 $P(A_k | B_1 B_2 \dots B_n)$ 的大小，去掉分母不影响比对结果，因此有：

$$P(A_k | B_1 B_2 \dots B_n) = P(A_k) P(B_1 B_2 \dots B_n | A_k)$$

根据朴素贝叶斯学习的假定，即 $B_1 B_2 \dots B_n$ 相互独立，有

$$\begin{aligned} P(A_k | B_1 B_2 \dots B_n) &= P(A_k) P(B_1 | A_k) P(B_2 | A_k) \cdots P(B_n | A_k) \\ &= P(A_k) \prod_i P(B_i | A_k) \end{aligned}$$

经典机器学习——朴素贝叶斯学习（Naive Bayes）

某个医院早上来了六个门诊的病人，他们的情况如下表所示：

症状	职业	疾病
打喷嚏	护士	感冒
打喷嚏	农夫	过敏
头痛	建筑工人	脑震荡
头痛	建筑工人	感冒
打喷嚏	教师	感冒
头痛	教师	脑震荡

现在又来了第七个病人，是一个打喷嚏的建筑工人。请问他患上感冒的概率有多大？

经典机器学习——朴素贝叶斯学习 (Naive Bayes)

$$P(\text{感冒} | \text{打喷嚏, 建筑工人}) = \frac{P(\text{感冒})P(\text{打喷嚏, 建筑工人} | \text{感冒})}{P(\text{打喷嚏, 建筑工人})}$$

根据朴素贝叶斯条件独立性的假设可知, "打喷嚏"和"建筑工人"这两个特征是独立的, 因此,

$$P(\text{感冒} | \text{打喷嚏, 建筑工人}) = \frac{P(\text{感冒})P(\text{打喷嚏} | \text{感冒})P(\text{建筑工人} | \text{感冒})}{P(\text{打喷嚏})P(\text{建筑工人})}$$

经典机器学习——朴素贝叶斯学习 (Naive Bayes)

根据表格估计概率，

$$P(\text{感冒})=3/6=0.5$$

$$P(\text{打喷嚏}|\text{感冒})=2/3=0.66$$

$$P(\text{建筑工人}|\text{感冒})=1/3=0.33$$

$$P(\text{打喷嚏})=3/6=0.5$$

$$P(\text{建筑工人})=2/6=0.33$$

症状	职业	疾病
打喷嚏	护士	感冒
打喷嚏	农夫	过敏
头痛	建筑工人	脑震荡
头痛	建筑工人	感冒
打喷嚏	教师	感冒
头痛	教师	脑震荡

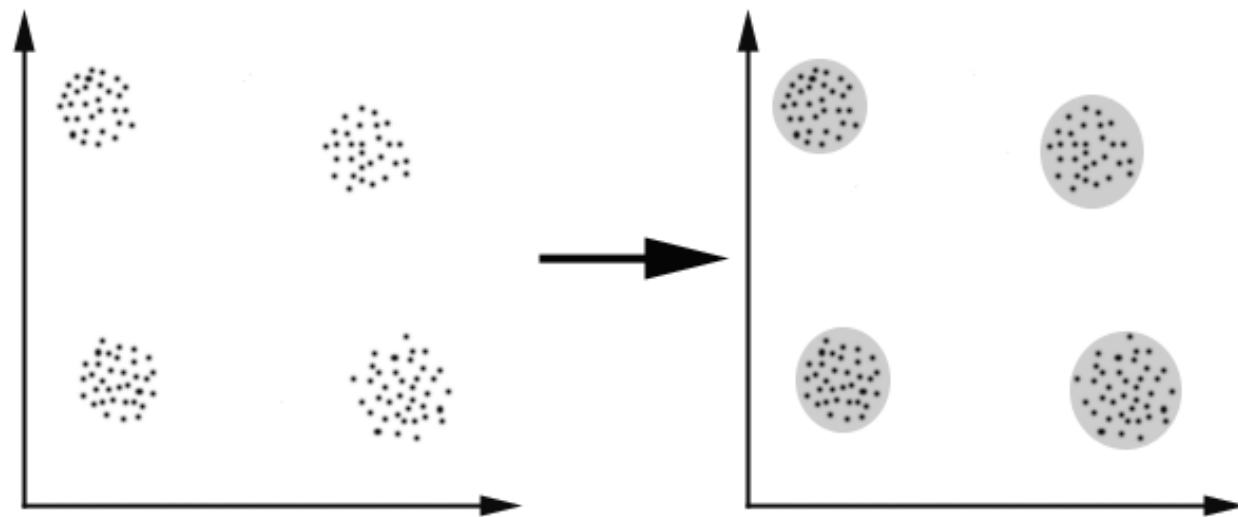
$$\begin{aligned}P(\text{感冒}|\text{打喷嚏}, \text{建筑工人}) &= \frac{P(\text{感冒})P(\text{打喷嚏}|\text{感冒})P(\text{建筑工人}|\text{感冒})}{P(\text{打喷嚏})P(\text{建筑工人})} \\&= \frac{0.5 \times 0.66 \times 0.33}{0.5 \times 0.33} = 0.66\end{aligned}$$

经典机器学习——K均值聚类 (K-Means)

K-Means算法是一种无监督学习算法，为聚类问题提供了一种解决方案。

K-Means 算法把 n 个点划分到 k 个类 (cluster)，使得每点都属于离它最近的种子点 (即均值，聚类中心，centroid) 所对应的类。

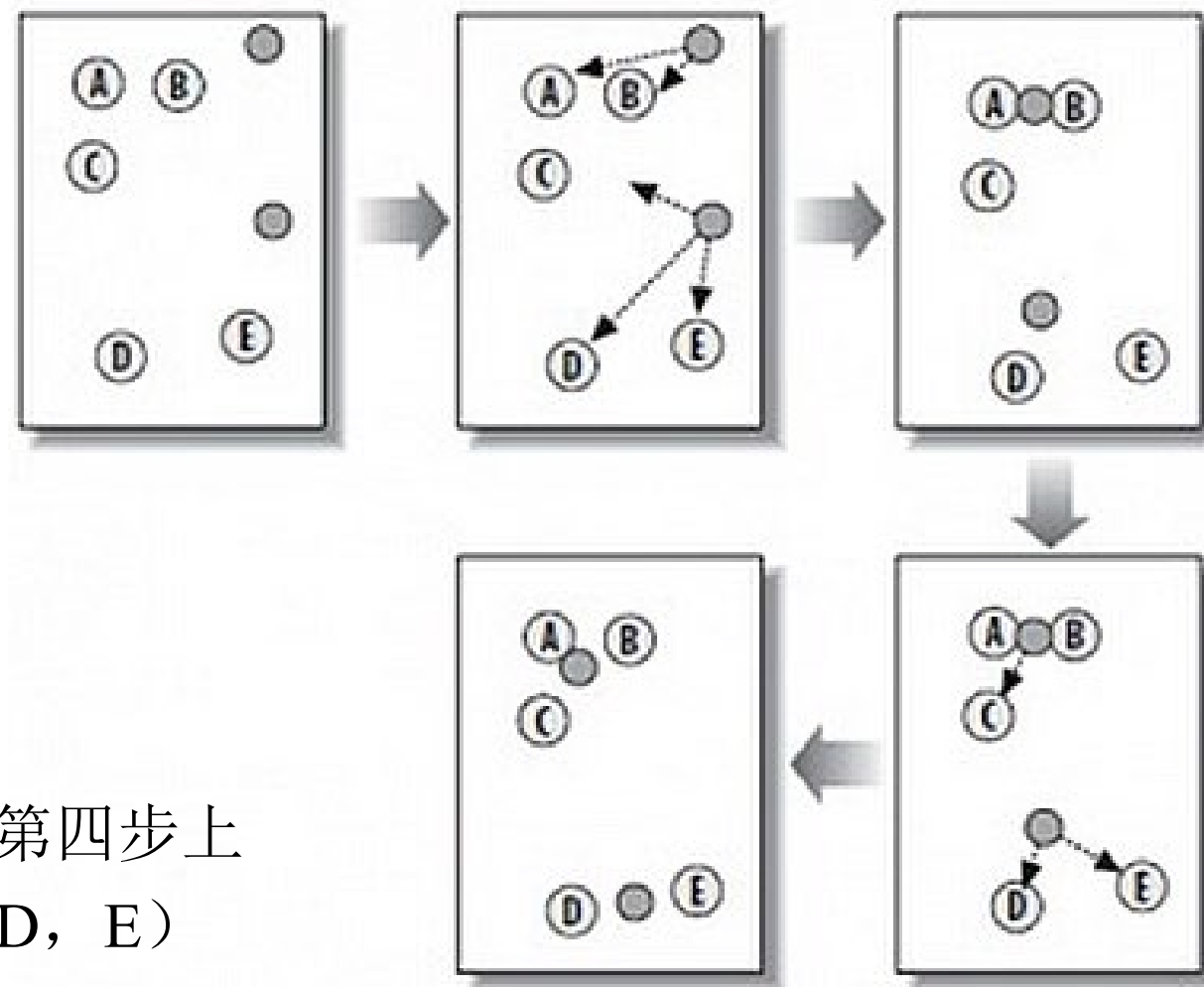
通过不断地更新种子点均值的算法。



经典机器学习——K均值聚类 (K-Means)

算法概要:

- 1) 随机在图中取 K 个种子点。
- 2) 求图中所有点到这 K 个种子点的距离，假如点 P_i 离种子点 S_i 最近，那么 P_i 属于 S_i 点群。（上图中，我们可以看到 A, B 属于上面的种子点， C, D, E 属于下面中部的种子点）
- 3) 移动种子点到所属“点群”的质心。
- 4) 重复第2)和第3)步，直到种子点不移动（图中的第四步上面的种子点聚合了 A, B, C ，下面的种子点聚合了 D, E ）



经典机器学习——支持向量机（SVM）

支持向量机（support vector machines）是一种二分类模型。

目的是寻找一个超平面来对样本进行分割。

分割的原则是间隔最大化，最终转化为一个凸二次规划问题来求解。

由简至繁的模型包括：

当训练样本线性可分时，通过硬间隔最大化，学习一个线性可分支持向量机；

当训练样本近似线性可分时，通过软间隔最大化，学习一个线性支持向量机；

当训练样本线性不可分时，通过核技巧和软间隔最大化，学习一个非线性支持向量机；

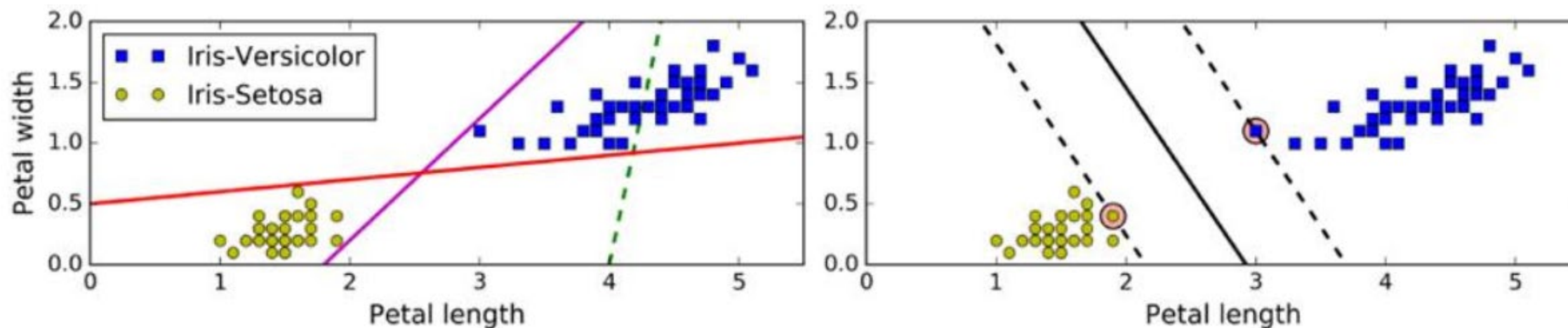
经典机器学习——支持向量机（SVM）

线性可分支持向量机

如果一个线性函数能够将样本分开，称这些数据样本是线性可分的。

显然不只有一条直线可将样本分开，而是有无数条。

线性可分支持向量机就对应着能将数据正确划分并且**间隔最大**的直线。

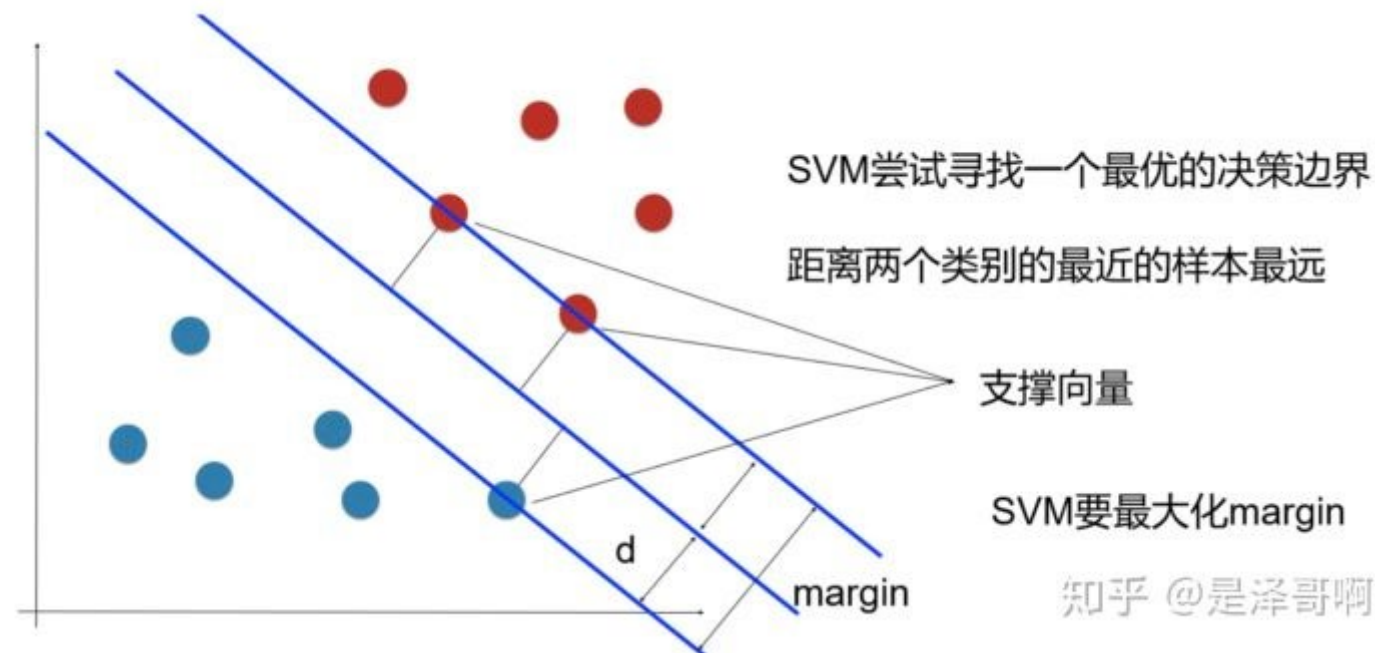


经典机器学习——支持向量机（SVM）

支持向量

超平面与最近数据点之间的距离叫作间隔（margin）。

能够将两个类分开的最佳超平面是具有最大间隔的直线。



只有这些点与超平面的定义和分类器的构建有关，这些点叫作支持向量，它们支持或定义超平面。在实际应用中，人们采用一种优化算法来寻找使间隔最大化的系数值。

其他——演绎学习

演绎学习基于演绎推理。

演绎推理是一种保真变换，即前提真则推出的结论也真。

在演绎学习中，学习系统由给定的知识进行演绎的保真推理，并存储有用的结论。

演绎学习包括知识改造、知识编译、产生宏操作、保持等价的操作和其他保真变换。

其他——类比学习

类比学习是基于类比推理的学习方法。

寻找和利用事物间可类比的关系,从已有的知识推导出未知的知识。

例如,学生在做练习时,往往在例题和习题之间进行对比,以发现相似之处,然后利用这种相似关系解决习题中的问题。

其他——类比学习的主要步骤

主要步骤:

- 回忆与联想
 - 通过回忆与联想,寻找出与之相似的已经解决了的有关问题,以获得有关知识。
- 建立对应关系
 - 建立相似问题知识和求解问题之间的对应关系,以获得求解问题的知识。
- 验证与归纳
 - 检验所获知识的有效性,如发现有错,就重复上述步骤进行修正。对于正确的知识,经过推广、归纳等过程取得一般性知识。

Any Questions?

