复旦大学管理学院 2026

人工智能领创计划

人工智能算法之机器学习

复旦大学 管理学院 统计与数据科学系 刘彬

2026年5月

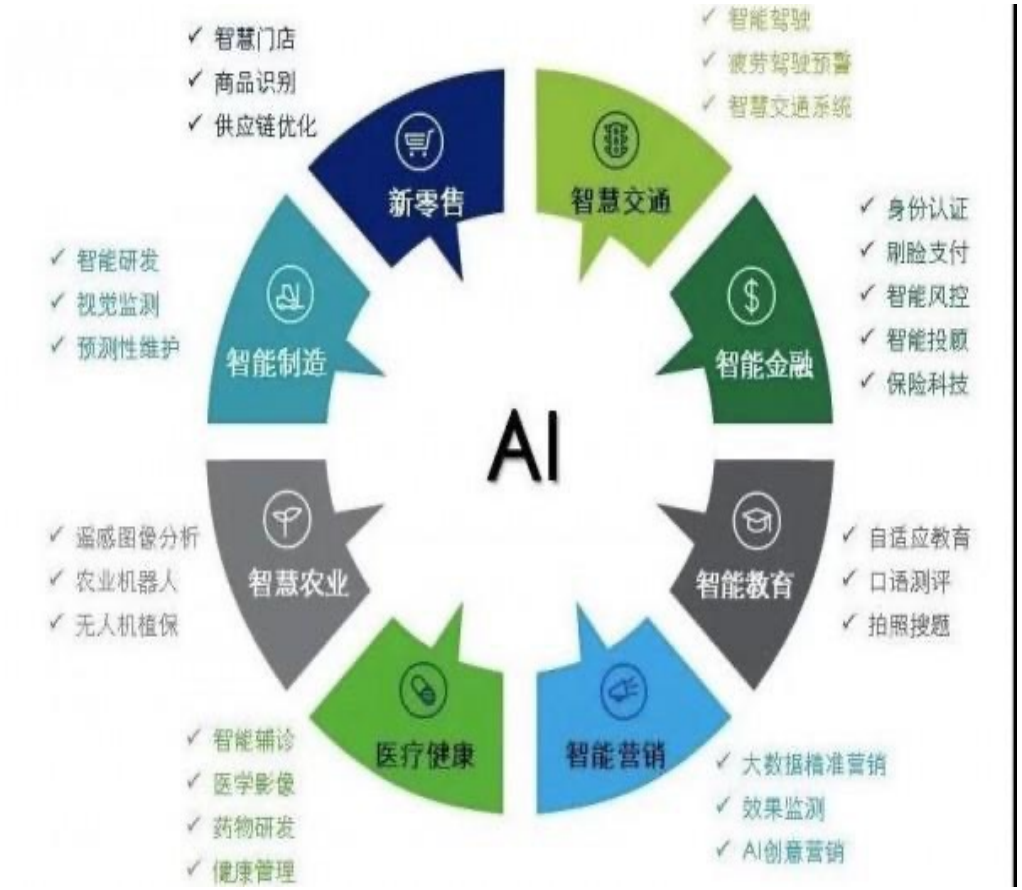
学习目标

- 目标1: 了解什么是机器学习？这包括：
  - 机器学习与人工智能、统计学、大数据、数据挖掘或者数据科学之间的关系是什么？
  - 机器学习的研究方法、研究对象以及研究步骤是什么？能否对机器学习给出一个定义？
  - 机器学习可以解决哪些问题？不能解决哪些问题？
  - 机器学习可以分为哪几个大的类别？
- 目标2: 了解机器学习所涉及到的基本方法和思想
- 理解现有机器学习模型的基本原理和主要思想：例如
  - 线性模型、逻辑回归，决策树，随机森林、Boosting聚类、降维等
  - 掌握机器学习的常见模型选择和评估方法以及理论指导思想
- 目标3: 能结合自身业务需求寻找恰当的机器学习方法解决相关问题

什么是机器学习

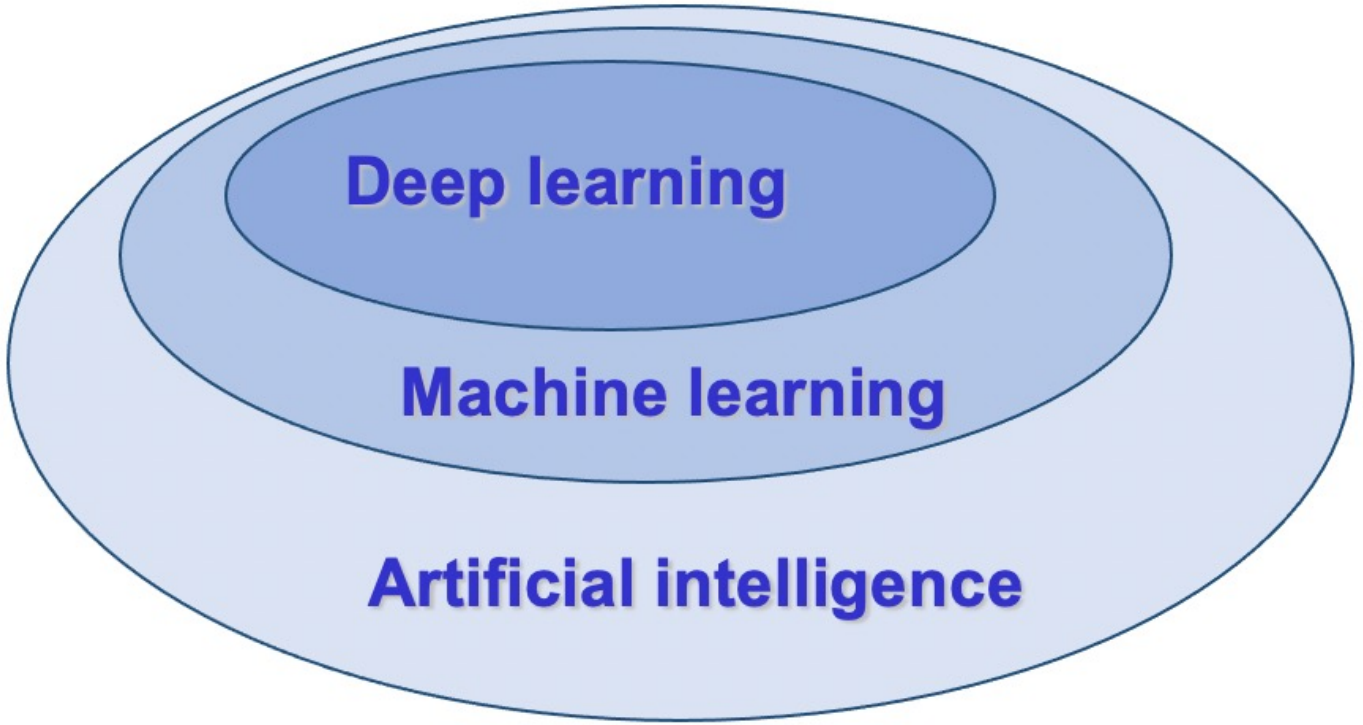
人工智能

- 人工智能：英文缩写为AI，是由一门由计算机科学、控制论、信息论、语言学、神经生理学、心理学、数学、统计学、哲学等多种学科相互渗透而发展的综合性的新学科。
- 在过去的几年里，人工智能(AI)一直是媒体大肆炒作的主题。机器学习、深度学习和人工智能出现在无数的文章中。
- 人工智能入选“2017年度中国媒体十大流行语”。
- 2017年3月，人工智能首次写入政府工作报告；
- 2017年7月，国务院印发了《新一代人工智能发展规划》明确了中国发展人工智能的战略目标，包括建设全球人工智能创新中心，加强核心技术研究，推动人工智能在各领域的应用。
- 2021年，《十四个五年规划和2035年远景目标纲要》，**指出将人工智能作为强化国家战略科技力量。**
- 2025年，国务院发布《关于深入实施“人工智能+”行动的意见》。
- **2026年，**《人工智能+制造“专项行动实施意见》。



什么是机器学习?

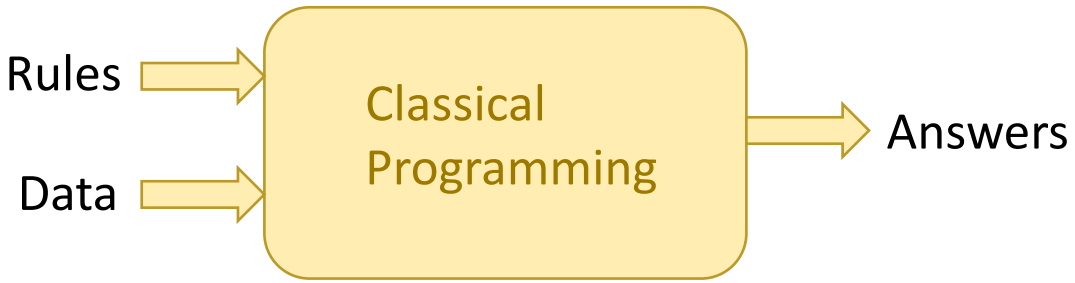
- 机器学习是从人工智能中产生的一个重要学科分支，是实现智能化的关键。机器学习是人工智能中的一步，属于程序范畴。
- 深度学习(Deep Learning)是机器学习发展地比较火热的一部分。其理念非常简单，就是传统的神经网络发展到了多隐藏层的情况，包括CNN、RNN、 LSTM等等。

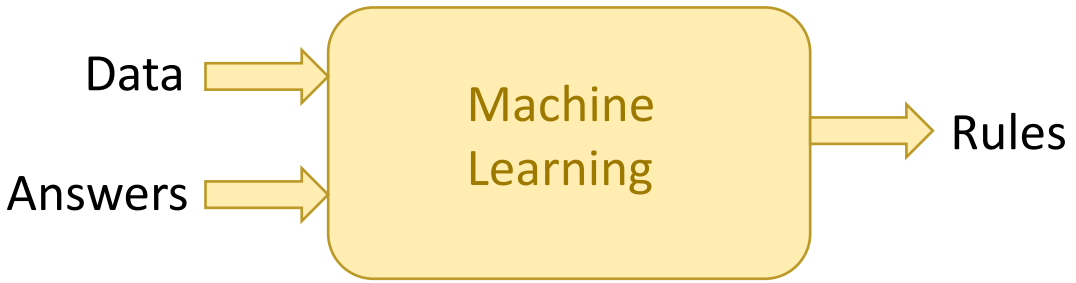


什么是机器学习

- 机器学习的最初想法是能否让计算机具备和人类一样的能力：决策（驾驶）、推理（医生治疗）、预测（股价）、识别（图像、语音）等。
- 在传统编程中，开发人员要对程序进行硬编码：给定算法或者规则，输出结果。
- 在机器学习中，输入数据和“答案”，计算机自己去学习规则或者数据的规律。
- 机器学习是训练数据而不是明确的编程。

数学运算+条件+循环
计算机并不具备理解能力





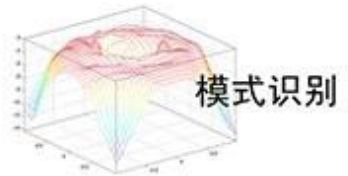
思考：如何判断计算机进行了学习呢?

什么是机器学习

- 机器学习（Mitchell, 1997)
  - 中文定义：假定用P来评估计算机程序在某个任务T上的性能，若一个程序通过利用经验E在T上的任务获得了改善，则我们说关于T和P,该程序对E进行了学习。
  - 英文定义：A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.
    - 任务T : task
    - 性能目标P: Performance Measurement
    - 训练经验E: Experience
- Example: 围棋
  - T : 下围棋
  - P: 比赛中击败对手(的百分比)
  - E: 和自己进行对弈，或看棋谱

Tom M. Mitchell 美国著名的计算机科学家，被誉为“机器学习之父”。他是卡内基梅隆大学（Carnegie Mellon University）计算机学院的创始教授，并创立了世界上第一个机器学习系。

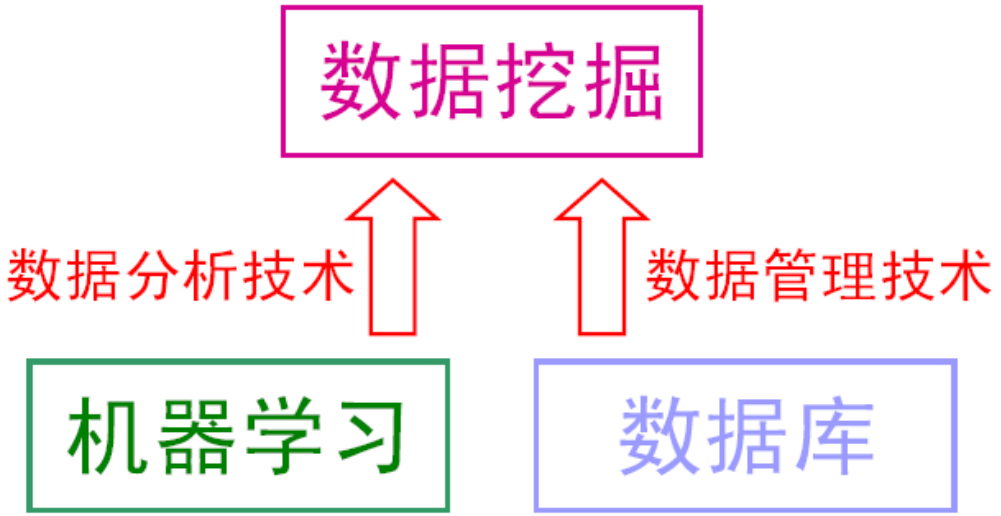
什么是机器学习



- 数据挖掘
- 统计学
- 大数据
- 数据科学
- 计算机视觉
- 语音识别
- 自然语言处理

机器学习与数据挖掘

- 数据挖掘是用人工智能、机器学习、统计学和数据库的交叉方法在相对较大型的数据集中发现模式的计算过程。
- 数据挖掘领域在20世纪90年代形成(KDD， Knowledge Discovery in Databases)，它受到多种科学领域的影响，其中数据库（SQL、Oracle）、机器学习、统计学影响最大。
- 数据挖掘从海量数据中发掘知识，发现数据中的模式、趋势和关联规则，这就必然涉及到对海量数据进行数据分析。
- 数据库为数据挖掘提供数据管理技术（存储、提取、清洗和转换、查询等），而机器学习或者统计学为数据挖掘提供分析技术。



机器学习与数据挖掘

第一步：清洗与整合来自多源的数据

- 数据来自不同业务系统（如销售数据库、日志文件、Excel 报表等）
- 原始数据问题：重复、缺失、格式不同
- 清洗目标：**统一格式、填补缺失、去除异常**
- **将整合后的数据打包进入数据仓库**

第二步：数据选择与变换

- 选择与问题相关的子数据集合或者变量
- 特征工程构建：创建新变量、数据离散化、类别编码、主成分构造等

第三步：应用算法提取模式或者模型

- 分类、聚类、关联规则、异常检测
- 统计学、机器学习

第四步：模型评估

- 技术评估：RMSE、准确性、F1度量等
- 业务评估：模型解释是否合理、是否达到业务要求等

第五步：发现知识、部署到业务平台

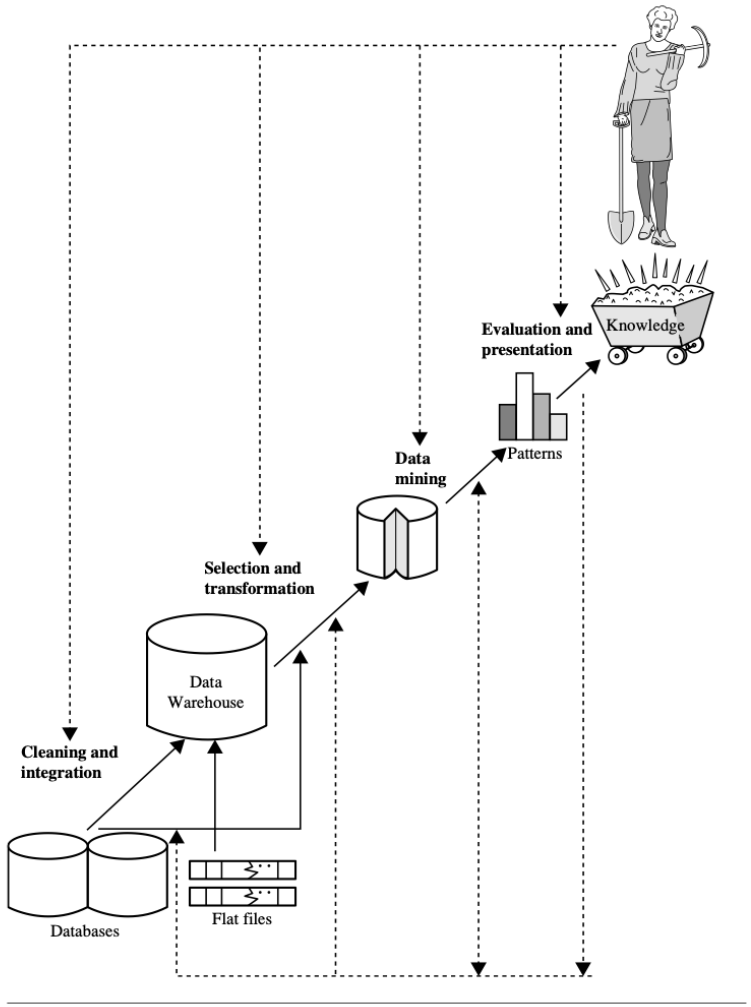


Figure 1.4 Data mining as a step in the process of knowledge discovery.

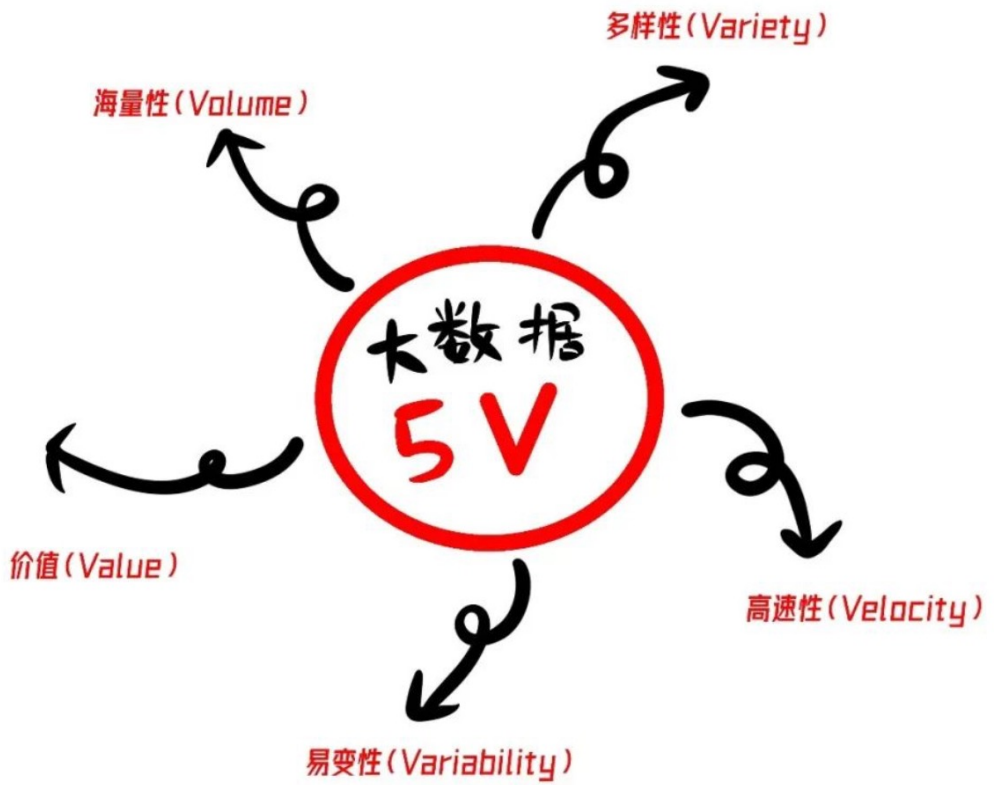
图片来源：《Data Mining: Concepts and Techniques》

机器学习与大数据

- 什么是大数据?

我国在2015年8月出台的《促进大数据发展行动纲要》中指出：大数据是以容量大、类型多、存取速度快、应用价值高为主要特征的数据集合，正快速发展为对数量巨大、来源分散、格式多样的数据进行采集、存储和关联分析，从中发现新知识、创造新价值、提升新能力的新一代信息技术和服务业态

- 业内较多的描述是大数据普遍具有“5V”特征:
- 根据国际数据公司IDC公布的《数据时代2025》报告显示，2025年人类的大数据量将达到163ZB（1ZB（太字节）=1024EB（艾字节），1EB=1024PB，1PB（拍字节）=1024TB，1TB=1024GB），相当于65亿年时长的高清视频内容。



大数据已经成为各行各业关键的生产要素！

机器学习与大数据

- 数据量必须与参数量（模型复杂度）相匹配，不仅适用于机器学习，深度学习以及大语言模型同样适用。
- 我们到底应该用多大的模型？又该配多少数据？这并不是一个凭经验拍脑袋的决策，而是可以被理论所指导的-**尺度定律**。
- OpenAI 在其经典论文《Scaling Laws for Neural Language Models》中，提出了一组精确描述**模型性能**随**规模**变化规律的幂律公式。

L：模型的性能（越小越好）

N：模型的参数量大小

D：训练模型的数据大小

C：训练模型使用的计算量 之间的关系

- 模型性能与数据量

$$L(D) = (\frac{D_c}{D})^{\alpha_D}$$

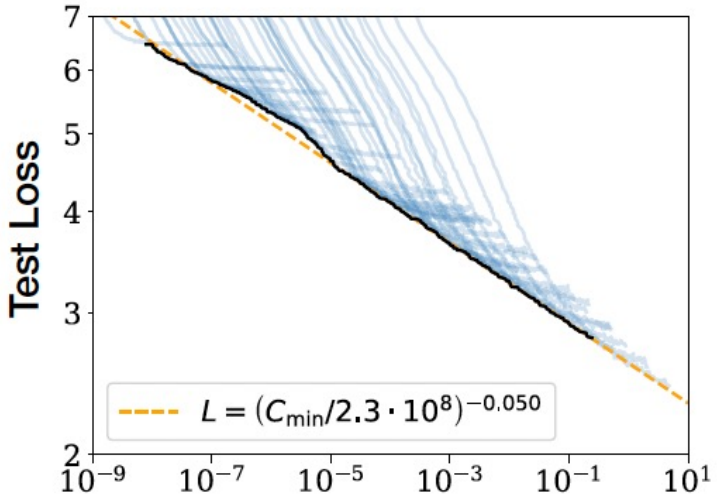
- 模型性能与参数量

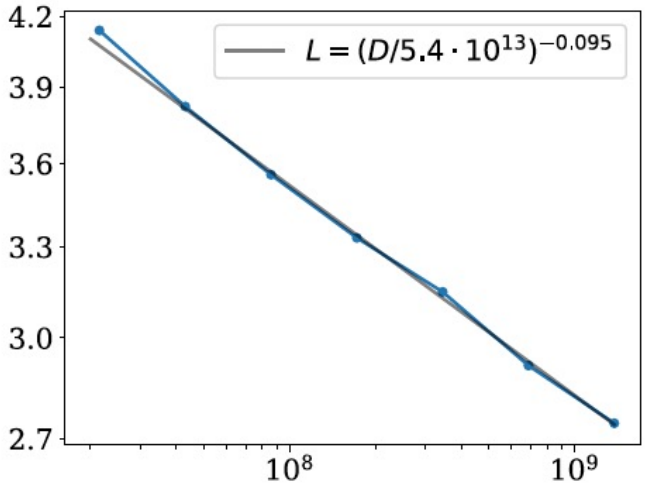
$$L(N) = (\frac{N_c}{N})^{\alpha_N}$$

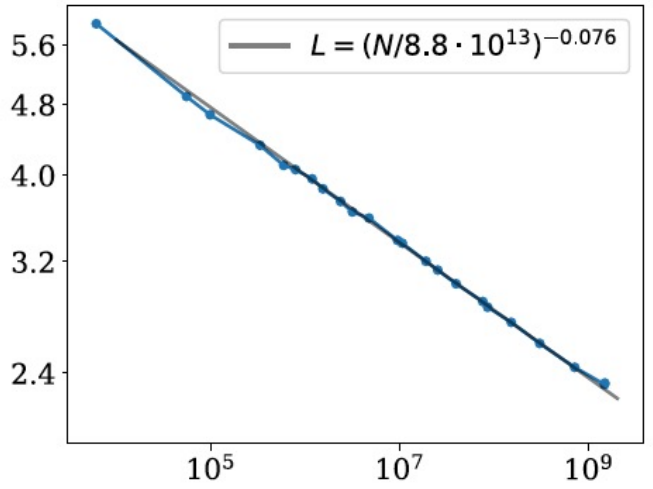
- 模型性能与计算量

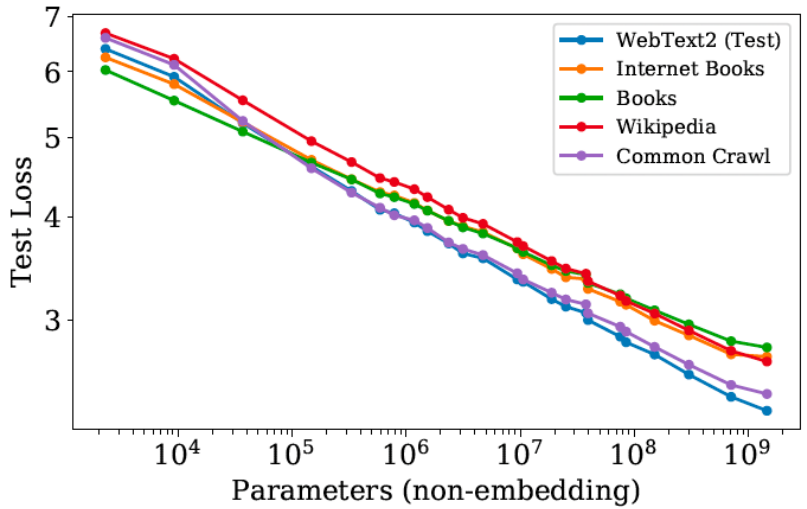
$$L(C) = (\frac{C_c}{C})^{\alpha_c}$$

机器学习与大数据


Test Loss
$L = (C_{min}/2.3 \cdot 10^8)^{-0.050}$
Compute
PF-days, non-embedding


$L = (D/5.4 \cdot 10^{13})^{-0.095}$
Dataset Size


$L = (N/8.8 \cdot 10^{13})^{-0.076}$
Parameters
non-embedding


WebText2 (Test)
Internet Books
Books
Wikipedia
Common Crawl
Test Loss
Parameters (non-embedding)

机器学习与大数据

- OpenAI 在其经典论文《Scaling Laws for Neural Language Models》中，提出了一组精确描述模型性能随规模变化规律的幂律公式。

L：模型的性能（越小越好）

N：模型的参数量大小

D：训练模型的数据大小

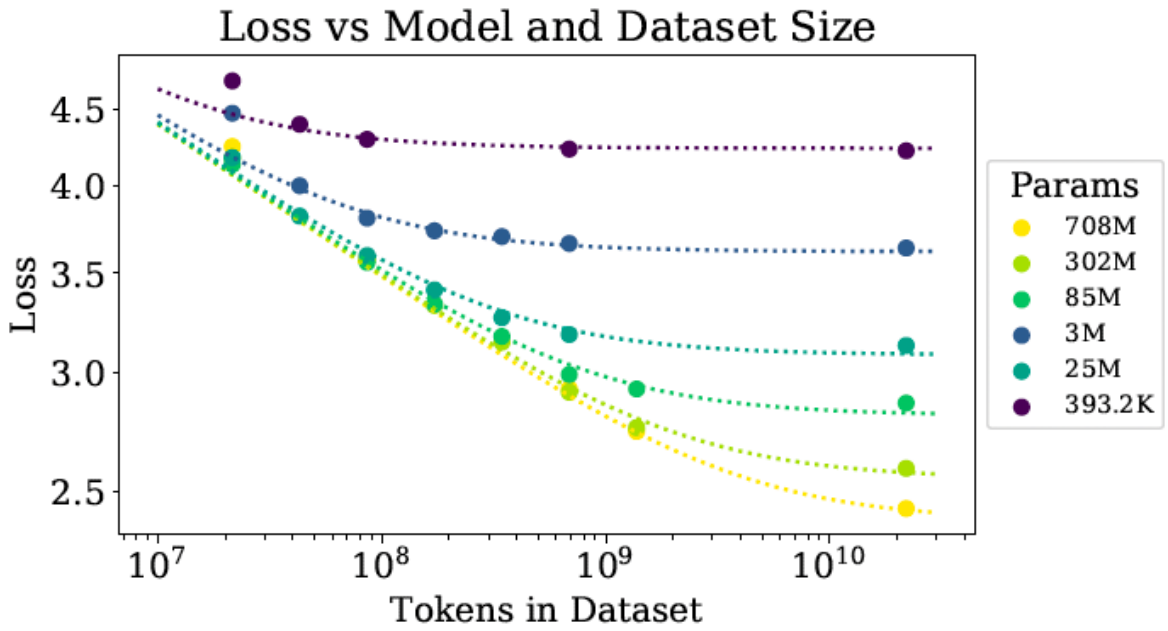
C：训练模型使用的计算量 之间的关系

- **模型性能与数据量、参数量经验关系**

$$L(N, D) = [(\frac{N_C}{N})^{\frac{\alpha_N}{\alpha_D}} + \frac{D_C}{D}]^{\alpha_D}$$

为了防止过拟合，要求参数和样本量满足

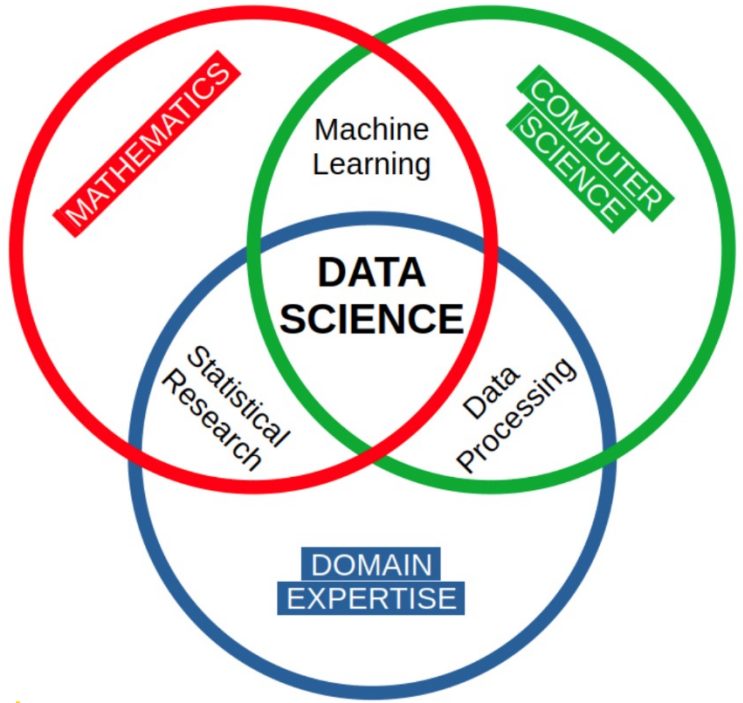
$$D \gtrsim (5 \times 10^3) N^{0.74}$$



机器学习与数据科学

- Data science is a "concept to unify statistics, data analysis, informatics, and their related methods" in order to "understand and analyse actual phenomena" with data. It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, information science, and domain knowledge. However, data science is different from computer science and information science.

—From Wikipedia, the free encyclopedia

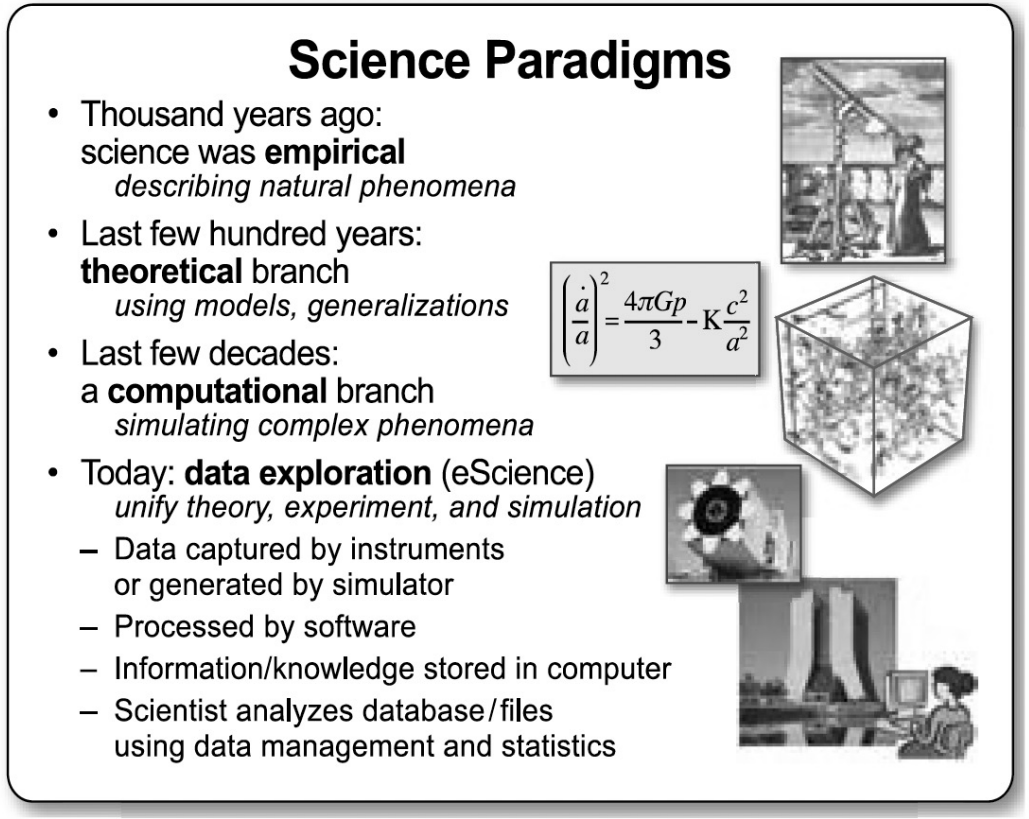


- 数据科学结合了诸多领域中的理论和技术，包括应用数学、统计、模式识别、机器学习、数据可视化、数据仓库以及高性能计算。

- 数据科学解决数据的高效获取、存储、计算、分析及应用问题

机器学习与数据科学

- **实验归纳**：钻木取火，24节气，伽利略发现行星运动规律等等。
- **模型推演**：牛顿三大定律，麦克斯韦理电磁学理论等等。
- 计算机仿真：模拟核试验、天气预报、新冠疫情传播等等。
- **数据密集范式**：通过搜集大量数据，利用计算机和人工智能技术发现规律。
- **以前研究模式**：提出可能的理论，搜集数据，然后通过计算来验证。
- **基于大数据的第四范式**：先有了大量的已知数据，然后通过计算得出之前未知的理论



机器学习可以做什么

机器学习可以解决什么

- 给定数据的预测问题
- 数据清洗/特征选择
- 确定算法模型/参数优化
- 发现数据的背后模式

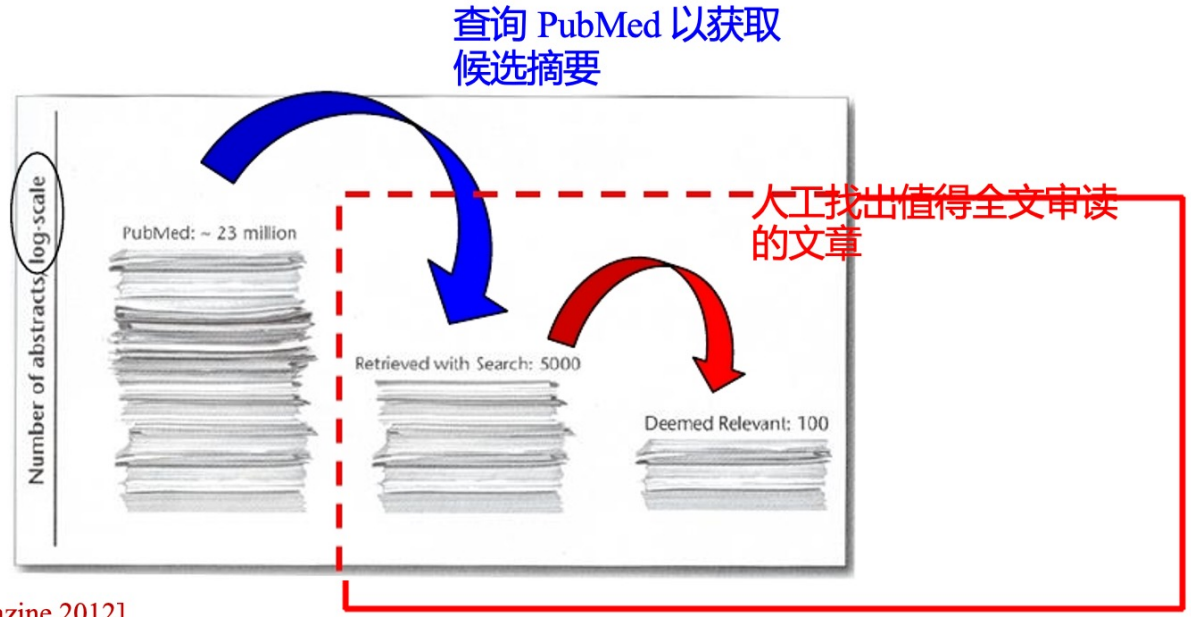
不能解决什么

- 大数据存储
- 做一个机器人

- 根据电子邮件中最常见的单词和标点符号的出现的相对频率，判断一封电子邮件是否为垃圾邮件。
- 根据数字化的图片文件，在图像中识别动物(如狗或猫)。
- 根据公司业绩指标和经济数据，预测6个月后的股价。
- 将大量的消费者分成若干组，每组的消费者都有相似的偏好。

“文献筛选”的故事

在“循证医学”（evidence-based medicine）中，针对特定的临床问题，先要对相关研究报告进行详尽评估



[C. Brodley et al., AI Magazine 2012]

“文献筛选”的故事

- 在一项关于婴儿和儿童残疾的研究中，美国Tufts医学中心筛选了约 33,000 篇摘要
- 尽管Tufts医学中心的专家效率很高，对每篇摘要只需 30 秒钟，
- 但该工作仍花费了 250 小时

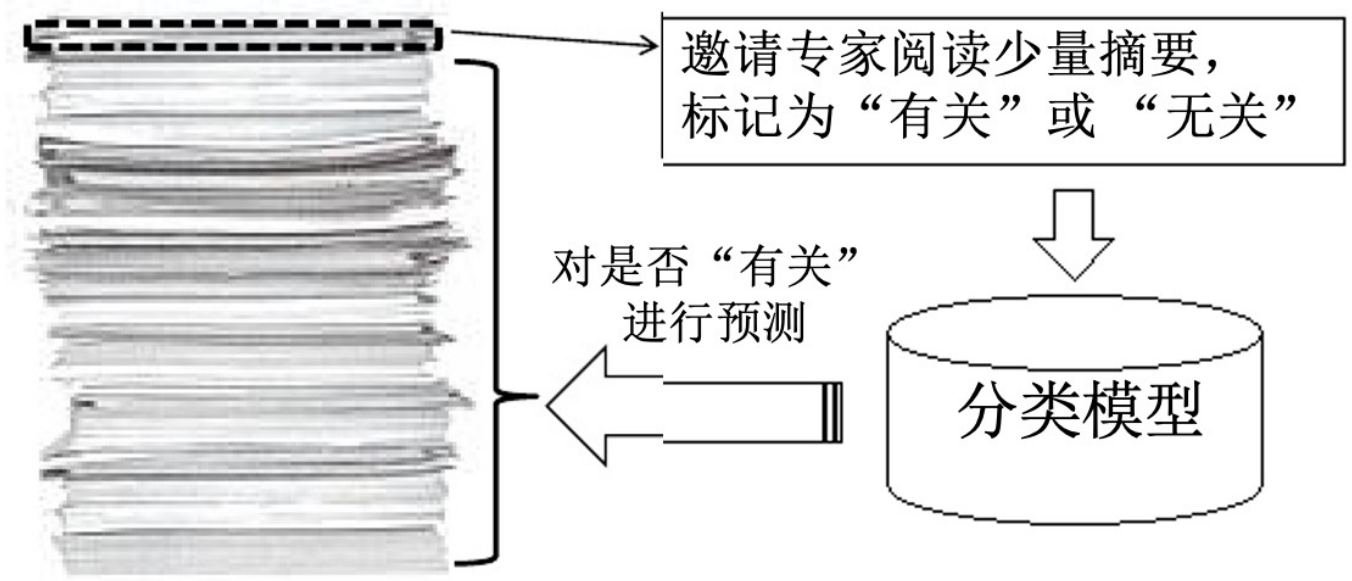
a portion of the 33,000 abstracts

每项新的研究都要重复这个麻烦的过程！

需筛选的文章数在不断显著增长！

“文献筛选”的故事

为了降低昂贵的成本, Tufts医学中心引入了机器学习技术



人类专家只需阅读 **50** 篇摘要，系统的自动筛选精度就达到 **93%**

人类专家阅读 **1,000** 篇摘要，则系统的自动筛选敏感度达到 **95%**

机器学习：NETFLIX的个性化推荐系统

如何推荐电影?

张三	李四	王五	陈六	
1	?	5	4	悬崖之上
?	1	4	5	飘
4	5	2	?	我不是药神
5	4	2	1	你好, 李焕英
4	5	1	2	长津湖
1	2	?	5	中国机长

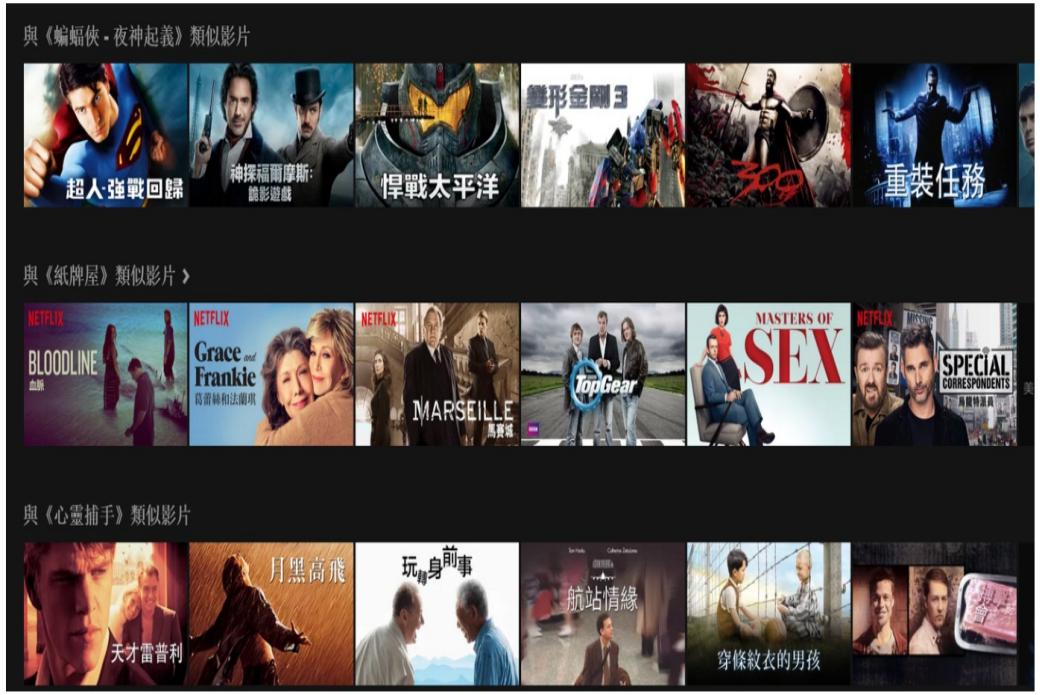
机器学习：NETFLIX的个性化推荐系统

- Netflix成立于1997年，最初是一家在线DVD租赁服务提供商。
- Netflix 百万美元大奖赛（Netflix Prize）是由Netflix公司于2006年10月发起的公开竞赛，目的是为了改进其电影推荐系统，并奖励能够显著提升推荐系统性能的团队或个人
- 为了更好地服务用户，Netflix使用推荐系统根据用户的观看历史和评分推荐内容。当时，Netflix的推荐系统叫做Cinematch，它基于协同过滤算法。Netflix意识到其推荐系统还可以提升，特别是在如何更准确地预测用户可能喜欢的内容上，因此希望通过竞赛的方式利用全球开发者的智慧来改进其算法。
- **竞赛内容：**参赛者需要设计新的算法，基于Netflix提供的用户评分数据集，来改进Cinematch算法的预测准确性。Netflix提供了一个大规模的匿名数据集，包含了1000多万条用户对电影的评分数据。
- **任务**：参赛者的任务是构建模型来预测用户对未看过的电影的评分。参赛者提交的预测结果会与Netflix的实际用户评分进行比较，系统通过均方根误差（Root Mean Square Error，RMSE）来评估预测精度。
- Netflix设定的目标是：团队必须至少比Netflix现有的Cinematch算法提高 **10%的预测精度**

2009年，经过三年的激烈竞争，名为 **BellKor's Pragmatic Chaos** 的团队首先达到了目标，成功提高了10.06%的预测准确度，赢得了100万美元大奖。

机器学习：NETFLIX的个性化推荐系统

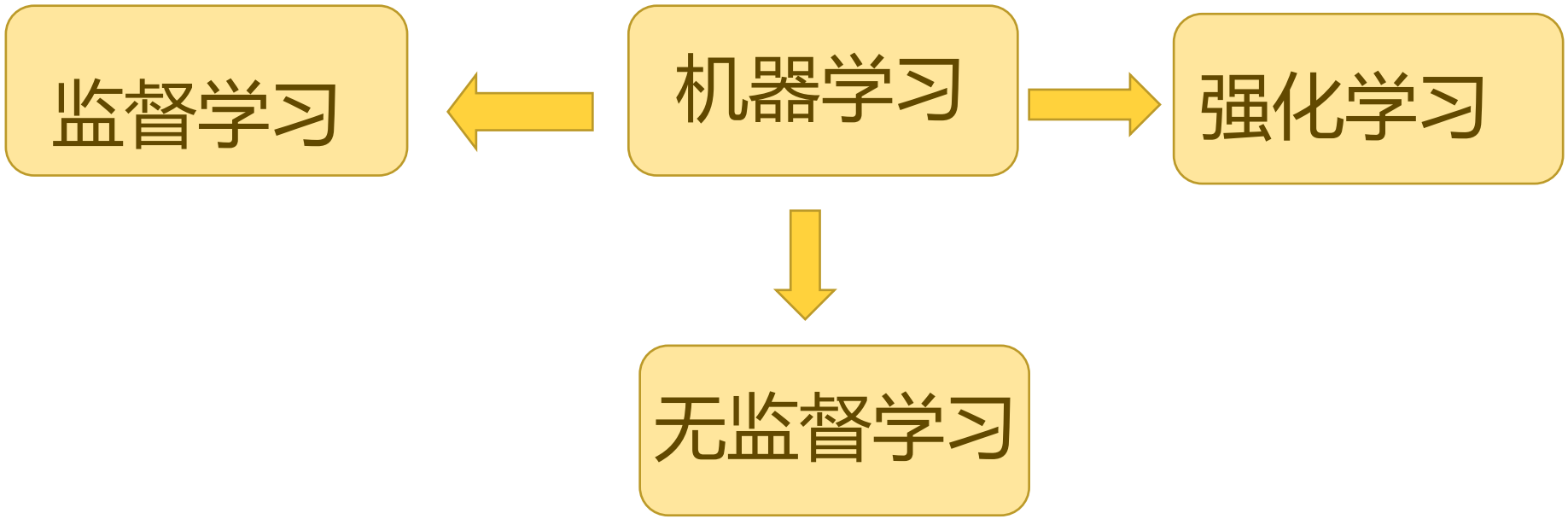
- 评分数据：上百亿的用户对视频的评分数据，用户对视频的主观评分，反映用户的喜好。
- 播放数据：每天上千万的播放数据，这些数据包括用户的播放时长、播放时间点、设备类型等。
- 喜好数据：如将视频添加进我的片单、将视频添加进播放列表等操作数据，反映着用户的喜好。
- 社交数据：Netflix与Facebook等社交网站打通，可以取到用户关联的Facebook账号的社交数据，如取到他们好友的播放记录，可实现基于好友的推荐。
- 其它比如用户在观看节目时暂停、倒带、跳过的时间节点



- Netflix成立于1997年，最初是一家在线DVD租赁服务提供商。随着时间的推移，公司转向了在线流媒体，并于2007年推出了其在线流媒体服务。
- 2006-2009年，推出百万美元大奖赛。
- Netflix积累了庞大的用户数据，这些数据让他成为世界上最了解用户的电影公司，也让Netflix从影片租赁、视频流媒体服务走上了自制剧的道路，纸牌屋等。

机器学习的分类

根据训练数据是否有标注（Label），机器学习可以分为**监督学习（Supervised Learning）**和**无监督学习 (Unsupervised Learning)**，以及**强化学习 (Rainforcement Learning)**。



机器学习的分类

根据训练数据是否有标注（Label），机器学习可以分为**监督学习（Supervised Learning）**和**无监督学习 (Unsupervised Learning)**，以及**强化学习 (Rainforcement Learning)。**

- 在一个典型的场景中，我们感兴趣的输出结果，通常是定量的(例如，股票价格)或分类的(例如，购买/不购买)，我们希望根据一组输入/特征 (例如，消费者的特征)来预测输出。
- 现在我们有多组输入与输出的样例，又称为训练数据集，其中每组样例是输出和与输入对。
- 使用这些数据，我们建立/训练一个从输入到输出的规则，或预测模型。
- 通过这组规则，当给定新的输入或者特征以后，能够预测新的看不见的输出。
- 这就是所谓的监督学习问题。

监督学习

垃圾邮件分类

- 数据由来自4601个电子邮件消息的信息组成，这些信息试图预测电子邮件是正常邮件还是“垃圾邮件”。
- 目标是设计一个自动垃圾邮件检测器(一条规则)。
- 对于所有4601电子邮件信息，我们有:
  - 电子邮件中57个最常用的单词和标点符号的相对频率——输入数据/特征;
  - 真实结果(电子邮件类型)：正常电子邮件或垃圾邮件——预期输出的示例。

Table： Average percentage of words and characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between spam and email.

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

监督学习

垃圾邮件分类

- 机器学习方法决定使用哪些特性以及如何使用: 例如

$$\text{if } (\%\texttt{george} < 0.6) \ \& \ (\%\texttt{you} > 1.5) \quad \text{then spam else email.}$$

$$\text{if } (0.2 \cdot \%\texttt{you} - 0.3 \cdot \%\texttt{george}) > 0 \quad \text{then spam else email.}$$

——规则/学习器/预测模型

- 这是一个监督学习问题; 也叫分类问题
- 对于该问题，并不是所有的错误都是相同的:
  - 将正常邮件过滤成了垃圾邮件
  - 将垃圾邮件错认为正常邮件
  - 我们想避免过滤掉好的电子邮件，虽然让垃圾邮件通过是不可取的，但其后果不那么严重

性能度量问题

监督学习的数学表述

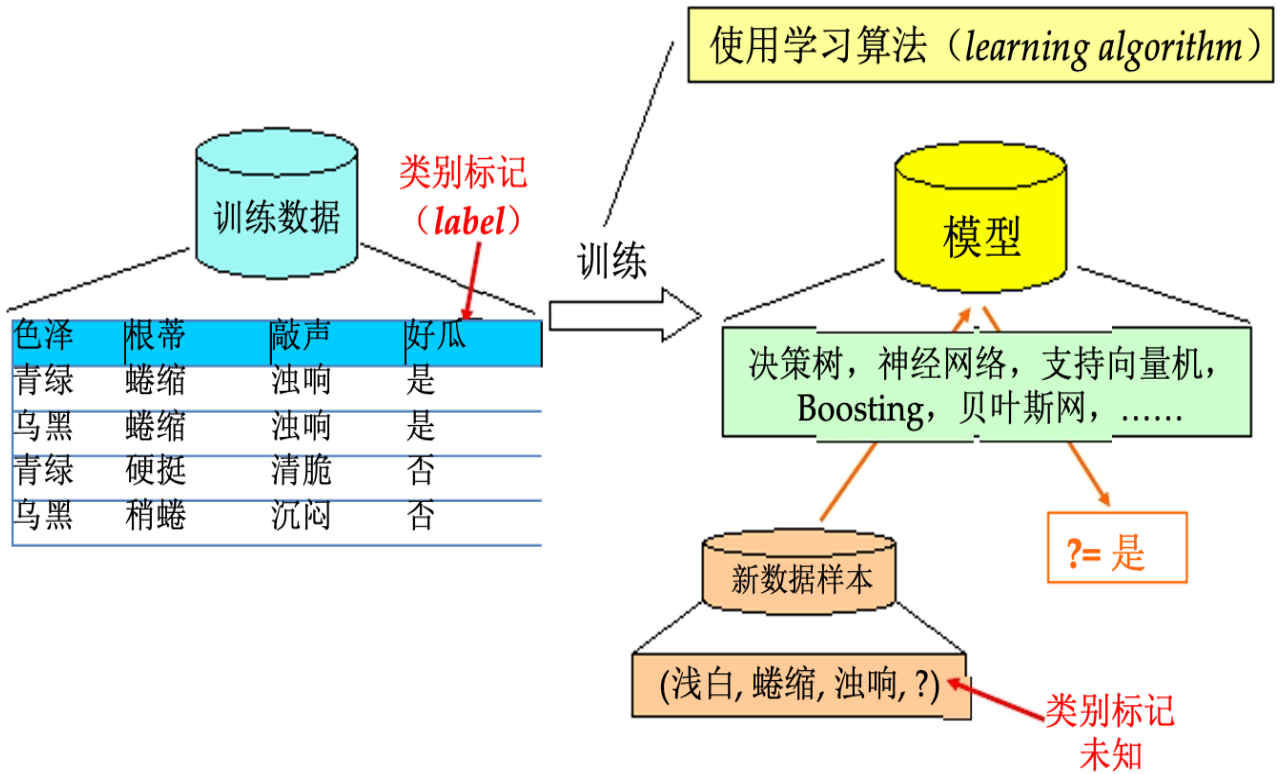
- 在监督学习中，数据通常由输入变量和输出变量组成。
- 输入变量：$X = (X^{(1)}, \ldots, X^{(n)}) \in \mathbb{R}^n$，其中输入变量又称作特征向量，输入变量取值的空间称为特征空间
- 输出变量：$Y \in \mathbb{R}^1$，其中输出变量取值的空间称为输出空间
- 假设总共有 N 个实例，对第 i 个实例，对应的样本记为

  $x_i = (x_i^{(1)}, \ldots, x_i^{(n)})^\top$ 以及 y_i，其中，$\mathrm{i} = 1, \ldots, N$.
- 监督学习的目的是在训练数据集合中学习模型，其中训练数据表示为

$$T = \{(x_1, y_1), \ldots, (x_N, y_N)\}$$

- 机器学习的目的是建立从输入值 X到输出值Y的之间映射 (又称决策函数)
- 一般通过训练数据集，通过某种模型和策略，学习得到一个模型(又称学习器)，通常记为 $Y = \hat{f}(X)$
- 当输入新的变量 x_{N+1}，得到相应输出值的预测值为 $y_{N+1} = \hat{f}(x_{N+1})$

监督学习



图：典型的机器学习过程,西瓜分类为例

好瓜还是坏瓜?

机器学习应用举例-回归的例子

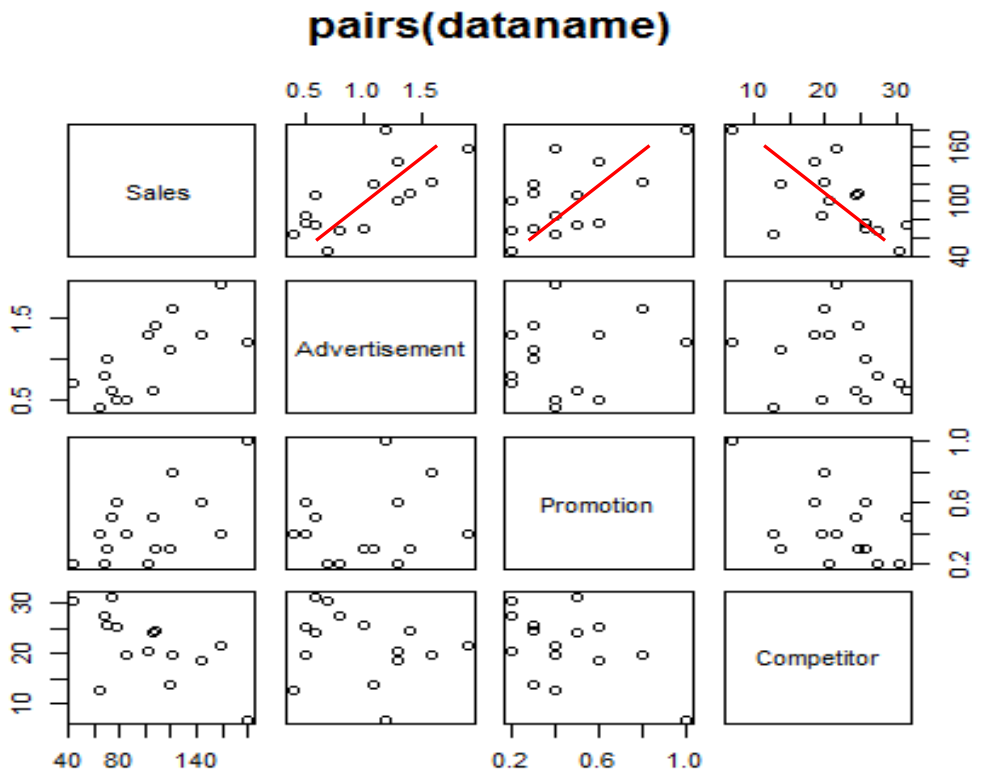
- 例：Country Kitchen公司零食年销售额

序号	地区	年销售额（百万美元）	广告投放额（百万美元）	促销投放额（百万美元）	竞争者年销售额（百万美元）
1	Selkirk	101.8	1.3	0.2	20.4
2	Susquehanna	44.4	0.7	0.2	30.5
3	Kittery	108.3	1.4	0.3	24.6
4	Acton	85.1	0.5	0.4	19.6
5	Finger Lakes	77.1	0.5	0.6	25.5
6	Berkshire	158.7	1.9	0.4	21.7
7	Central	180.4	1.2	1.0	6.8
8	Providence	64.2	0.4	0.4	12.6
9	Nashua	74.6	0.6	0.5	31.3
10	Dunster	143.4	1.3	0.6	18.6
11	Endicott	120.6	1.6	0.8	19.9
12	Five-Towns	69.7	1.0	0.3	25.6
13	Waldeboro	67.8	0.8	0.2	27.4
14	Jackson	106.7	0.6	0.5	24.3
15	Stowe	119.6	1.1	0.3	13.7
16	Columbus	?	1.2	0.5	25
17	Bloomington	?	1	0.6	10
18	Buffalo	?	1.4	0.7	8

根据历史数据，如何利用机器学习方法预测新的三个地区的销售额？

机器学习应用举例-回归的例子

- 首先回答几个问题
- 第一，这是一个分类问题还是回归问题？
  - 由于考虑的是年销售额的预测，销售额取值范围连续，因此是回归问题
- 第二，这里的输入变量和目标变量分别是什么？
  - 目标变量为年销售额，即
    Y=年销售额
  - 输入变量有三个，因此
    X=(广告投放额度，促销投放额，竞争者年销售额)
- 第三，应该考虑什么模型呢？
  - 观察右侧两两散点图，可以看到广告、促销与目标变量呈现正相关关系，竞争对手与目标变量为负相关关系。
  - 因此可以考虑建立线性回归模型来对销售额进行预测。
- 第四，什么是线性模型？



$$Y \approx f_\beta(\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

⇩

决策函数

机器学习应用举例-回归的例子

- 首先回答几个问题
- 第四，什么是线性模型？

$$Y \approx f_\beta(\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

- 第五，这个线性形式的决策函数好还是不好？如何度量？
  - 采用L-2损失，即

$$\begin{aligned} L(Y, f_\beta(\mathrm{X})) &= (Y - f_\beta(\mathrm{X}))^2 \\ &= (Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2 - \beta_3 X_3)^2 \\ &= \left(Y - \boldsymbol{X}^\top \boldsymbol{\beta}\right)^2 \end{aligned}$$

- 第六，怎么把最优的参数学习出来？

$$T = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$$

序号	地区	年销售额 (百万美元)	广告投放额 (百万美元)	促销投放额 (百万美元)	竞争者年销售额 (百万美元)
1	Selkirk	101.8	1.3	0.2	20.4
2	Susquehanna	44.4	0.7	0.2	30.5
3	Kittery	108.3	1.4	0.3	24.6
4	Acton	85.1	0.5	0.4	19.6
5	Finger Lakes	77.1	0.5	0.6	25.5
6	Berkshire	158.7	1.9	0.4	21.7
7	Central	180.4	1.2	1.0	6.8
8	Providence	64.2	0.4	0.4	12.6
9	Nashua	74.6	0.6	0.5	31.3
10	Dunster	143.4	1.3	0.6	18.6
11	Endicott	120.6	1.6	0.8	19.9
12	Five-Towns	69.7	1.0	0.3	25.6
13	Waldeboro	67.8	0.8	0.2	27.4
14	Jackson	106.7	0.6	0.5	24.3
15	Stowe	119.6	1.1	0.3	13.7

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \quad \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_N \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{pmatrix}$$

机器学习应用举例-回归的例子

- 第六，怎么把最优的参数学习出来？
  - 构造基于训练数据的经验损失：

$$R_{\text{emp}}(\boldsymbol{\beta}) = \frac{1}{2N}\sum_{i=1}^{N}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2$$

  - 最优的参数应该是能够拟合训练数据最小的参数组合

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\hat{\beta}_1, \ldots, \hat{\beta}_p) \\ &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2N}\sum_{i=1}^{N}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2\end{aligned}$$

序号	地区	年销售额 (百万美元)	广告投放额 (百万美元)	促销投放额 (百万美元)	竞争者年销售额 (百万美元)
1	Selkirk	101.8	1.3	0.2	20.4
2	Susquehanna	44.4	0.7	0.2	30.5
3	Kittery	108.3	1.4	0.3	24.6
4	Acton	85.1	0.5	0.4	19.6
5	Finger Lakes	77.1	0.5	0.6	25.5
6	Berkshire	158.7	1.9	0.4	21.7
7	Central	180.4	1.2	1.0	6.8
8	Providence	64.2	0.4	0.4	12.6
9	Nashua	74.6	0.6	0.5	31.3
10	Dunster	143.4	1.3	0.6	18.6
11	Endicott	120.6	1.6	0.8	19.9
12	Five-Towns	69.7	1.0	0.3	25.6
13	Waldeboro	67.8	0.8	0.2	27.4
14	Jackson	106.7	0.6	0.5	24.3
15	Stowe	119.6	1.1	0.3	13.7

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \quad \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_N \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{pmatrix}$$

- 第七，如何求解上述的最优化问题？
  - 梯度下降法（稍后讲）

机器学习应用举例-回归的例子

- 第八，确定预测模型的具体形式

$$Y \approx f_\beta(\mathbf{X}) = 65.70 + 48.98 \times \text{广告投放额度} + 59.65 \times \text{促销投放额} - 1.84 \times \text{竞争者年销售额}$$

通过梯度下降法估计参数

$$Y \approx f_\beta(\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

β_0的估计值：$\hat{\beta}_0 = 65.70$

β_1的估计值：$\hat{\beta}_1 = 48.98$

β_2的估计值：$\hat{\beta}_2 = 59.65$

β_3的估计值：$\hat{\beta}_3 = -1.84$

- 第九，如何对新的数据进行预测?

Columbus（广告投放额 =1.2, 促销投放额 =0.5, 竞争者年销售额 =25）：

$$Y_{Columbus} = 65.70 + 48.98 \times 1.2 + 59.65 \times 0.5 - 1.84 \times 25 = 108.36$$

Bloomington（广告投放额 =1, 促销投放额 =0.6, 竞争者年销售额 =10）

$$Y_{Bloomington} = 65.70 + 48.98 \times 1 + 59.65 \times 0.6 - 1.84 \times 10 = 132.09$$

Buffalo（广告投放额 =1.4, 促销投放额 =0.7, 竞争者年销售额 =8）：

$$Y_{Buffalo} = 65.70 + 48.98 \times 1.4 + 59.65 \times 0.7 - 1.84 \times 8 = 1631.33$$

地区	广告投放额（百万美元）	促销投放额（百万美元）	竞争者年销售额（百万美元）
Columbus	1.2	0.5	25
Bloomington	1	0.6	10
Buffalo	1.4	0.7	8

什么是机器学习：总结

- 机器学习研究的对象是数据，即从数据出发，提取数据的特征或者规律，抽象出数据模型，利用学习的模型对新数据进行预测或者分析。
- 实现机器学习的步骤如下：

① 确定研究目标；

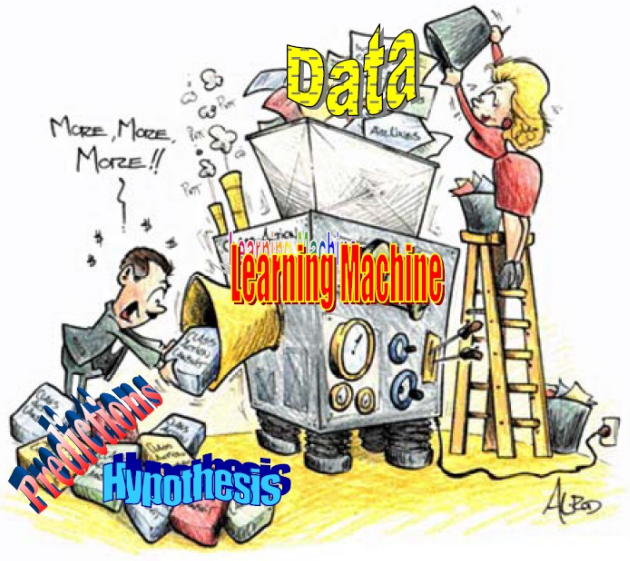
② 得到一个有限的训练数据集合；

③ 确定包含所有可能的模型的 假设空间，即学习模型的集合；

④ 确定模型选择的准则，即学习的策略；

⑤ 通过某种算法求解最优的模型；

⑥ 利用学习的最优模型对新数据进行预测或者分析。



机器学习=数据+模型+策略+算法

机器学习三要素：模型

- 机器学习首先要考虑的是学习什么样的模型。
- 在监督学习过程中，模型就是要学习的决策函数或者条件概率分布。模型的假设空间包含所有可能的条件概率分布或者决策函数的集合：

$$\mathscr{F} = \{f | Y = f_\theta(X), \theta \in R^n\}$$

其中，X 和 Y 是定义在输出空间 $\mathscr{X}$ 和输出空间 $\mathscr{Y}$ 上的变量，这时 $\mathscr{F}$ 通常是一个由参数向量决定的函数族。而参数向量 θ 取值于 n 维欧式空间 R^n，称为参数空间。

例如，线性模型中 $Y \approx f_\beta(\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

机器学习三要素：策略

策略：有了模型以及对应的假设空间，机器学习接着需要考虑按照什么样的准则（又称策略）选择最优的模型。因此，统计机器学习的目标在于从假设空间中选取最优模型。

损失函数：

- 监督学习的目的是找到决策函数 f(x), 使得给定输入 X，能够有输出预测值 f(X)，该值可能与真实值 Y一致也可能不一致，损失函数（loss function）或代价函数 （cost function) 用来度量预测错误的程度。

- 损失函数是预测值 f(X)与真实值 Y的非负实值函数，记作 $L(Y, f(X))$

机器学习三要素：策略

机器学习常用的损失函数有以下几种：

- 0-1损失函数（0-1 loss）

$$L(Y, f(\boldsymbol{X})) = \begin{cases} 0 & \text{if } Y = f(\boldsymbol{X}) \\ 1 & \text{if } Y \neq f(\boldsymbol{X}) \end{cases}$$

二分类问题：
- 真实值 Y={0（负类），1（正类）}
- 预测值 f(X)={0,1}

- 平方损失 (squared loss)

$$L(Y, f(X)) = (Y - f(X))^2$$

回归问题：
- 真实值 Y取值连续
- 预测值 f(X)取值连续

- 绝对值损失 (absolute loss)

$$L(Y, f(X)) = |Y - f(X)|$$

- 其它：Huber 损失，分位数损失、合页损失，逻辑损失、指数损失等

机器学习三要素：策略

- 给定训练数据集合：

$$T = \{(x_1, y_1), \ldots, (x_N, y_N)\}$$

- 经验损失定义为：

$$R_{\text{emp}}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) \quad \xrightarrow{\text{回归问题}} \quad R_{\text{emp}}(\boldsymbol{\beta}) = \frac{1}{2N}\sum_{i=1}^{N}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2$$

- 给定假设空间，损失函数和数据集，经验风险最小化策略为：

$$\min_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) \quad \xrightarrow{\text{回归问题}} \quad \begin{aligned} \hat{\boldsymbol{\beta}} &= (\hat{\beta}_1, \ldots, \hat{\beta}_p) \\ &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \tfrac{1}{2N}\textstyle\sum_{i=1}^{N}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 \end{aligned}$$

机器学习三要素：算法

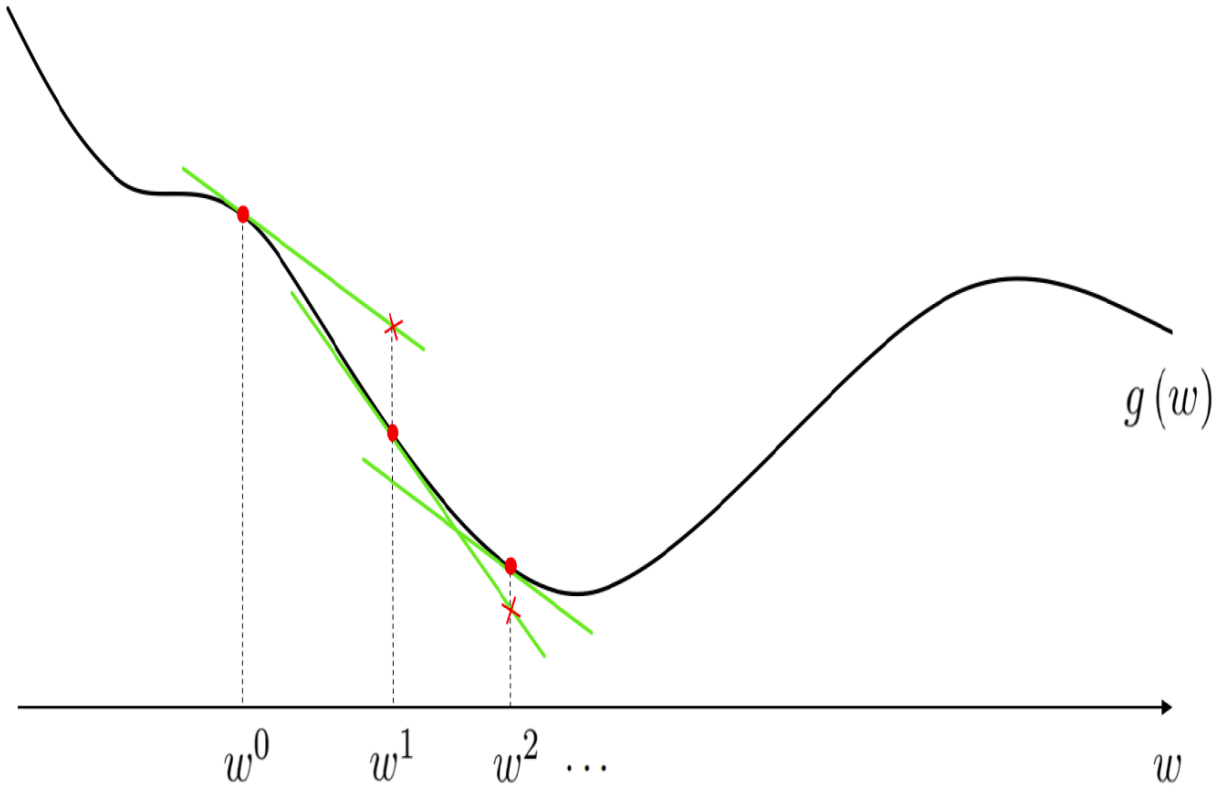
- 算法是指学习模型的具体计算方法。统计学习基于训练数据，根据学习策略，从假设空间中选择最优模型，最后需要考虑需要用什么样的计算方法求解最优模型。
- 机器学习最终都会归结为最优化问题。
- 机器学习方法的不同主要来自于模型、策略以及算法的不同（以回归问题为例）。

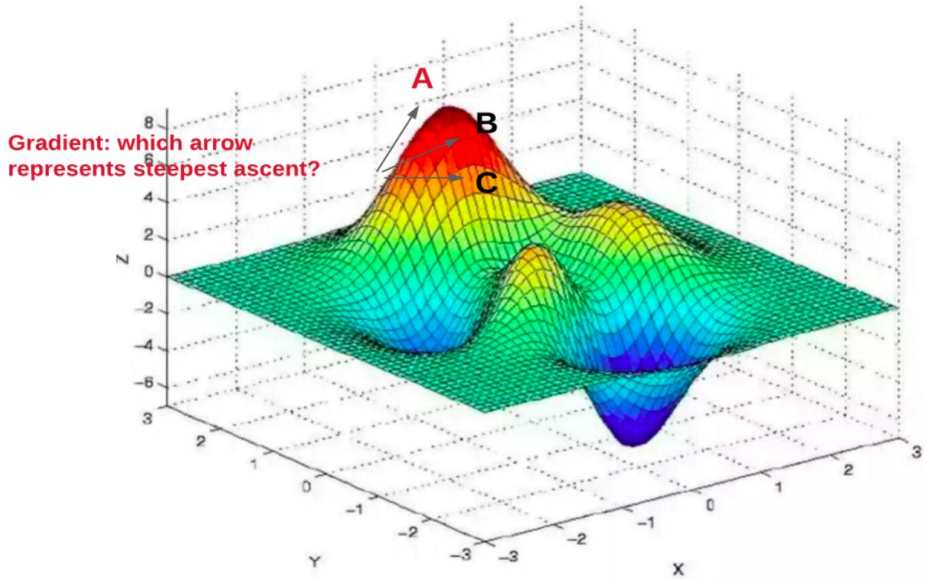
- ***梯度下降法***，*牛顿法，拟牛顿法，共轭梯度下降法，**随机梯度下降法**，坐标下降法等*
- *梯度投影算法，罚函数方法，增广的拉格朗日算法，ADMM算法（凸集与凸优化）*
- *SVD分解算法*
- *EM算法*
- *MH 算法，Gibbs 抽样算法*

《统计计算》课程

算法补充：梯度下降法

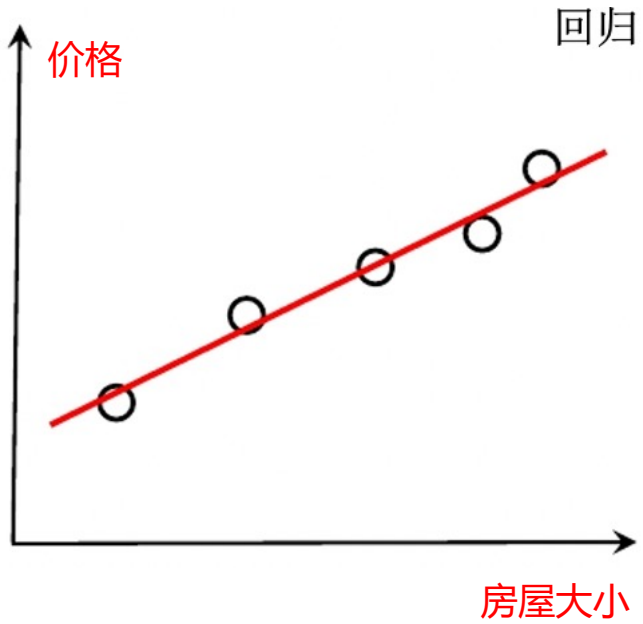
命题：沿函数的梯度负方向是函数值下降最快的方向。





机器学习应用举例：逻辑回归模型

- 线性模型主要解决的是机器学习中当输出值为连续取值的预测问题，例如
  - 根据房屋的位置、面积、房龄来预测房屋价格
  - 根据上市公司历史财务数据、历史股票价格、成交量来预测其未来几天的股票价格
  - 根据出租车出发地和目的地距离、时间段、天气状况等来预测打车费用
  - 根据产品的历史销售数据、广告支出、促销折扣、产品类别等预测未来某段时间的销售额



- 可以看到，房屋价格、股票价格、打车费用、销售额等均是连续型变量
- 很自然，我们可以构造线性决策函数来对输出值进行预测：

$$Y \approx f(\boldsymbol{x}) = w_1 x_1 + \cdots + w_d x_d + b$$

输出值　决策函数　变量1　变量d

思考：输出结果为连续值的情况下为什么可以直接构造线性决策函数?

机器学习应用举例：逻辑回归模型

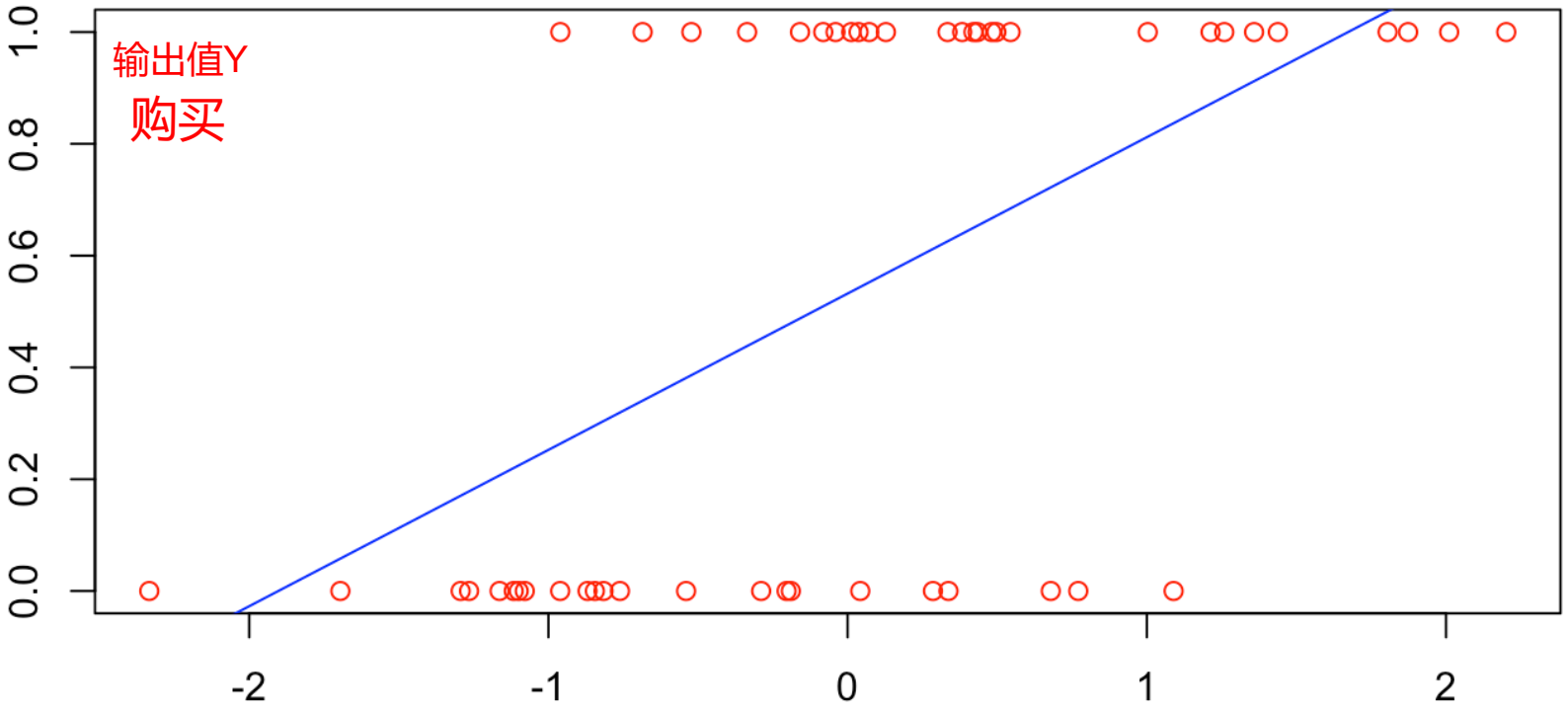
- 除预测连续值以外，实际应用中，我们还会碰到这样的问题，例如：
  - 假设你是一个广告商，你希望预测某个用户是否会点击你展示的广告，你可能搜集到用户的年龄、浏览历史、兴趣爱好等数据。
  - 假如你是银行信贷部门的经理，你想通过客户的历史信用评分、收入水平、是否有负债、家庭状况、教育程度、工作行业等数据判断客户是否会违约。
  - 假如你想通过电子邮件中的关键词、发件人地址、邮件长度等内容判断所接收到的邮件是否为垃圾邮件。
- 可以看到，上述问题的预测变量并非连续型数值，而是具有{0,1}取值的离散变量。
- **思考，我们是否可以像回归模型一样，构造决策函数以预测0-1输出值，即**

$$Y \approx f(\boldsymbol{x}) = w_1 x_1 + \cdots + w_d x_d + b$$

输出值　　决策函数　　变量1　　变量d

机器学习应用举例：逻辑回归模型

- 根据右侧图片，显然，采用自变量的线性组合无法很好的拟合我们的0-1输出值
- 通常情况下，我们又希望找到一个自变量的线性组合来解释我们的输出值
- 一个自然的问题是，当输出结果为离散情况下时又该如何相应的构造决策函数呢?



逻辑回归模型（Logistic regression model）

机器学习应用举例：逻辑回归模型

回顾统计学中的两点分布：

- 每次投掷一枚硬币，正面记为“成功”，反面记为“失败”。
- 假设用数字1表示：成功“，0表示“失败”，假设成功的概率为 $p \in (0,1)$
- 我们可以构造离散型随机变量Y

表：两点分布随机变量分布列

Y 的取值	0	1
概率	$1-p$	p

不包含任何特征信息

- 通常记作 $Y \sim b(1,p)$
- 两点分布可以刻画很多二分类的例子：顾客是否会点击广告；客户是否会违约；邮件是否为垃圾邮件等等。

机器学习应用举例：逻辑回归模型

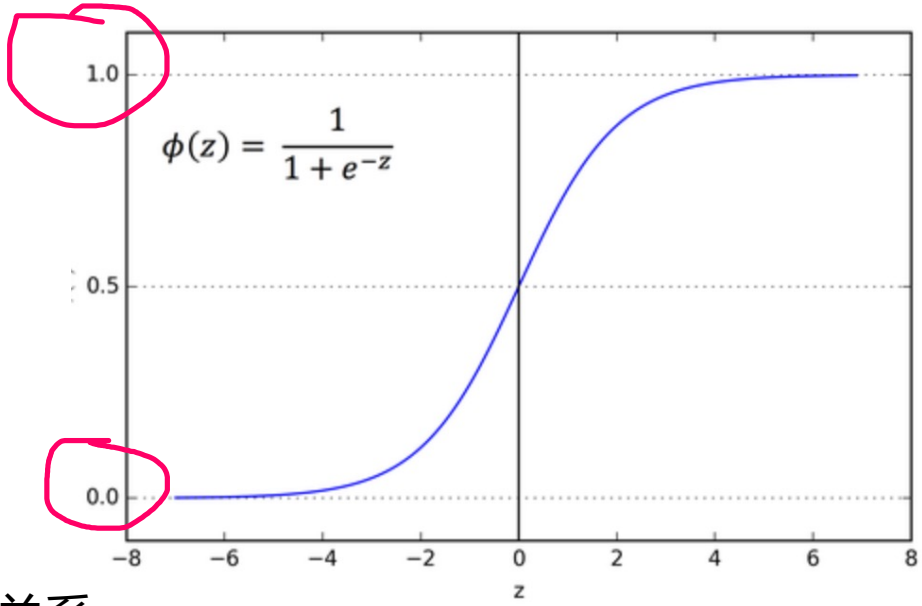
- 一个自然的想法是考虑如下的变换：

$$\boldsymbol{w}^\top \boldsymbol{x} + b = w_1 x_1 + \cdots + w_d x_d + b := f(\boldsymbol{x})$$

$$p(\boldsymbol{x}) = \frac{\exp(\boldsymbol{w}^\top \boldsymbol{x} + b)}{1 + \exp(\boldsymbol{w}^\top \boldsymbol{x} + b)}, \quad 1 - p(\boldsymbol{x}) = \frac{1}{1 + \exp(\boldsymbol{w}^\top \boldsymbol{x} + b)}$$

- Sigmoid 函数：

$$\sigma(z) = \frac{\exp(z)}{1 + \exp(z)} = \frac{1}{1 + \exp(-z)}, \quad -\infty < z < \infty$$



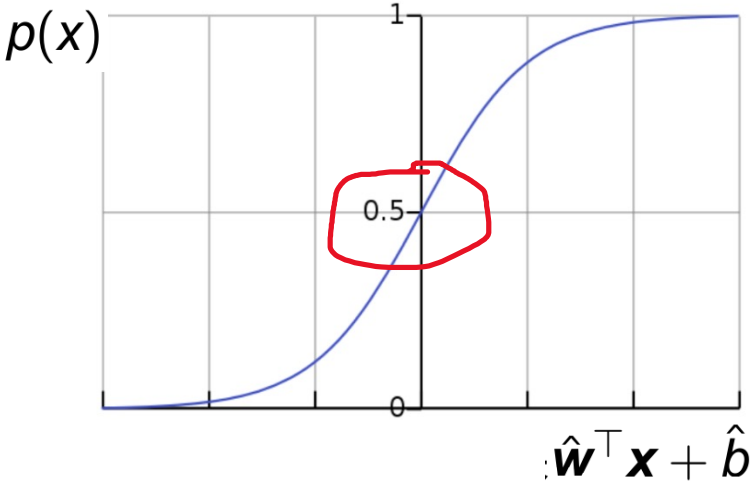
- 因此，我们可以通过sigmoid变换将决策函数f(x)与概率p(x)建立关系：

$$p(\boldsymbol{x}) := \sigma(\boldsymbol{w}^\top \boldsymbol{x} + b) = \sigma(f(\boldsymbol{x}))$$

机器学习应用举例：逻辑回归模型

- 根据概率变换公式，我们可以得到当输入值为x情况下的概率值

$$p(x) = \sigma(f(\boldsymbol{x})) = \frac{\exp(\hat{\boldsymbol{w}}^\top \boldsymbol{x} + \hat{b})}{1 + \exp(\hat{\boldsymbol{w}}^\top \boldsymbol{x} + \hat{b})}$$

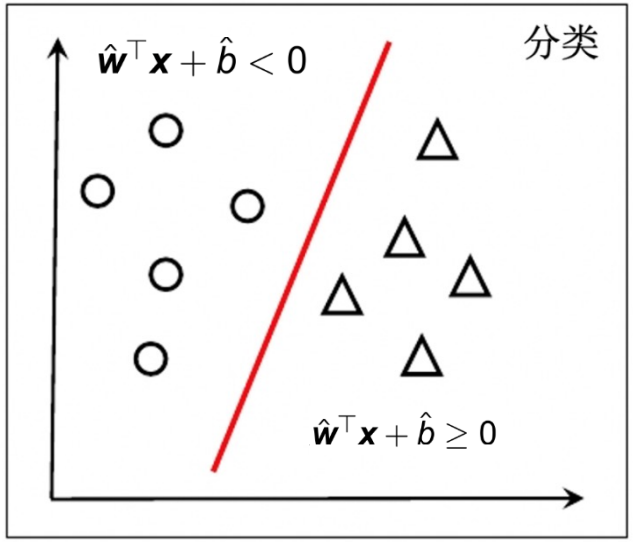


- 我们根据计算出的概率 p(x)，选择阈值(通常为0.5)进行分类：

$$\hat{y} = \begin{cases} 1, & \text{如果} p(x) \geq 0.5 \\ 0, & \text{如果} p(x) < 0.5 \end{cases}$$

$$\Longleftrightarrow$$

$$\hat{y} = \begin{cases} 1, & \text{如果} \hat{\boldsymbol{w}}^\top \boldsymbol{x} + \hat{b} \geq 0 \\ 0, & \text{如果} \hat{\boldsymbol{w}}^\top \boldsymbol{x} + \hat{b} < 0 \end{cases}$$



实例分析

- 对一个企业而言，员工的流失会给企业带来巨大成本，包括业务中断，雇用新员工和培训新员工的成本。因此，了解员工流失背后的各种因素，并最大程度地减少人员流失是企业人力资源部门非常感兴趣的问题。
- 该数据集提供了IBM的员工调查，目的是调查哪些因素是人员流失的主要因素。数据集包含大约1500个条目。
- 目标变量：是否离职。
- 解释变量：年龄，性别，月收入，工作满意度，加班频率等，总计15个影响因素。
- 使用Logistic回归模型可以比较准确的预测员工是否有可能离职，从而可以大大提高人力资源部门按时干预并对错误予以纠正，而防止了人员的流失。

因素	变量名	中文描述	变量类型	可能取值
工作经历	ATTRITION	是否离职（0=否，1=是）	离散	0, 1
个人基本情况	AGE	年龄	连续	数值型
个人基本情况	GENDER	性别	离散	1=女性, 2=男性
收入情况	MONTHLY INCOME	月收入	连续	数值型
收入情况	PERCENT SALARY HIKE	薪资涨幅百分比	连续	数值型
收入情况	STOCK OPTIONS LEVEL	股票期权等级	连续	数值型
工作满意度	JOB SATISFACTION	对工作的满意度	连续	数值型
工作强度	BUSINESS TRAVEL	出差频率	离散	1=不出差, 2=频繁出差, 3=偶尔出差
工作强度	DISTANCE FROM HOME	家庭与工作地点的距离	连续	数值型
工作强度	JOB INVOLVEMENT	工作投入度	连续	数值型
工作强度	WORK LIFE BALANCE	工作与生活平衡程度	连续	数值型
工作经历	NUMCOMPANIES WORKED	曾工作公司数量	连续	数值型
工作经历	TOTAL WORKING YEARS	总工作年限	连续	数值型
工作经历	TRAINING TIMES LAST YEAR	去年培训次数	连续	数值型
工作经历	YEARS AT COMPANY	在公司年限	连续	数值型
工作经历	YEARS SINCE LAST PROMOTION	距离上次晋升年数	连续	数值型

逻辑回归应用

- p：离职的概率
- $x_1, \ldots, x_{15}$：性别，年龄，月收入，薪资涨幅比例，工作地点离家距离等等
- 模型：
- $\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \cdots + \beta_{16} x_{15}$

$$p = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_{16} x_{15})}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_{16} x_{15})}$$

逻辑回归应用

	项	系数	变换后的系数
效应	Age	-0.0283	0.972097
反向	JobInvolvement	-0.554	0.574647
反向	JobSatisfaction	-0.303	0.738599
反向	MonthlyIncome	-0.000063	0.999937
反向	PercentSalaryHike	-0.0158	0.984324
反向	StockOptionLevel	-0.539	0.583331
反向	TotalWorkingYears	-0.0499	0.951325
反向	TrainingTimesLastYear	-0.1635	0.849166
反向	WorkLifeBalance	-0.255	0.774916
反向	YearsAtCompany	-0.053	0.94838
正向	NumCompaniesWorked	0.1393	1.149469
正向	YearsSinceLastPromotion	0.1242	1.132242
正向	DistanceFromHome	0.03613	1.036791
正向	BusinessTravel		
正向	Travel_Frequently	1.722	5.595709
正向	Travel_Rarely	0.925	2.521868
正向	Gender		
正向	Male	0.231	1.259859

在其它因素不变的情况下

股权水平每增加一单位，离职的几率缩小0.58倍

上年度培训次数每增加一单位，离职的几率缩小0.84倍

距离上次晋升时间每增加一单位，离职的几率扩大1.13倍

相比于不出差，出差频繁，离职的几率扩大5.59倍

逻辑回归应用

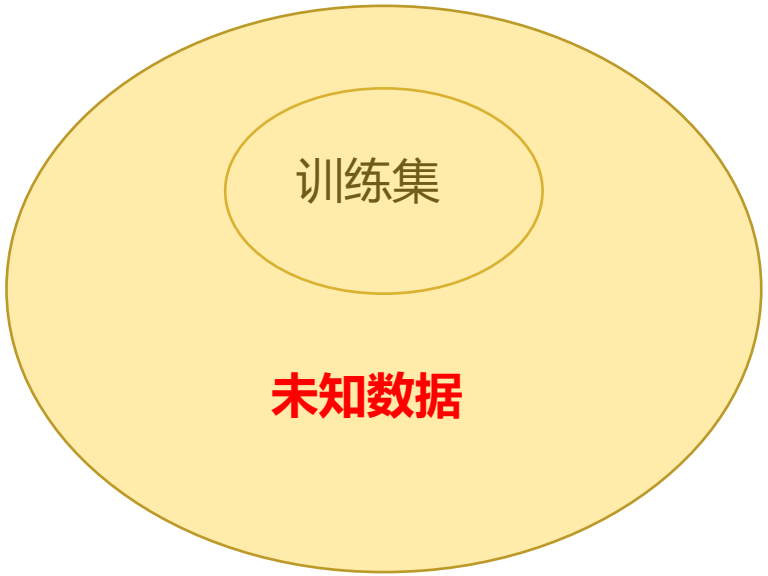
	个体1	个体2	个体3	个体4
Age	28	30	39	53
Attrition	Yes	Yes	No	No
BusinessTravel	Travel_Frequently	Travel_Frequently	Travel_Rarely	Travel_Rarely
DistanceFromHome	2	26	3	2
EnvironmentSatisfaction	3	3	3	3
Gender	Male	Female	Female	Male
JobInvolvement	2	2	4	3
JobSatisfaction	1	1	2	3
MonthlyIncome	2561	6696	6389	14852
NumCompaniesWorked	7	5	9	6
PercentSalaryHike	11	15	15	13
RelationshipSatisfaction	3	3	3	3
StockOptionLevel	0	0	1	1
TotalWorkingYears	8	9	12	22
TrainingTimesLastYear	2	5	3	3
WorkLifeBalance	2	2	1	4
YearsAtCompany	0	6	8	17
YearsSinceLastPromotion	0	0	3	15
预测离职概率	0.82761156	0.6650553	0.133366879	0.047670099

模型评估与选择

- 泛化能力：学习方法的泛化能力（generalization ability）是指该方法学习到的模型对未知数据的预测能力。
- 从本质上讲，机器学习的目的是学到一个模型能够**有好的泛化能力**。
- 泛化能力越好，该机器学习方法一定越好。

- 泛化误差：在“未来”样本上的误差
- 经验误差：在训练集上的误差，亦称“训练误差”

结论：泛化误差越小越好！

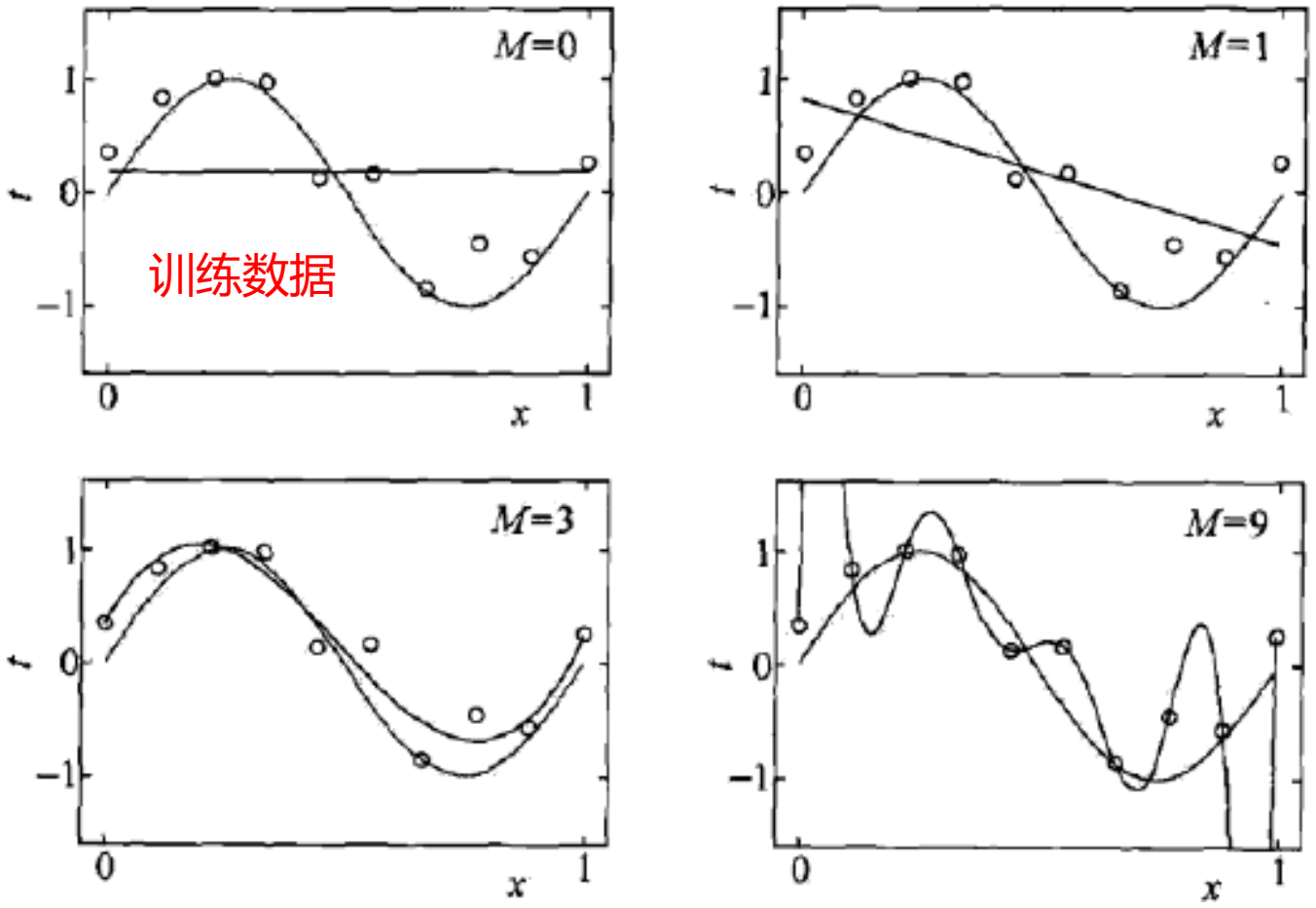


NO！因为会出现过拟合(Overfitting)

模型评估与选择

- M=0，欠拟合
- M=9，过拟合
- 不能一味的追求经验误差最小，否则会出现过度拟合现象。

不能简单的通过经验误差来衡量模型的泛化能力。



模型评估与选择

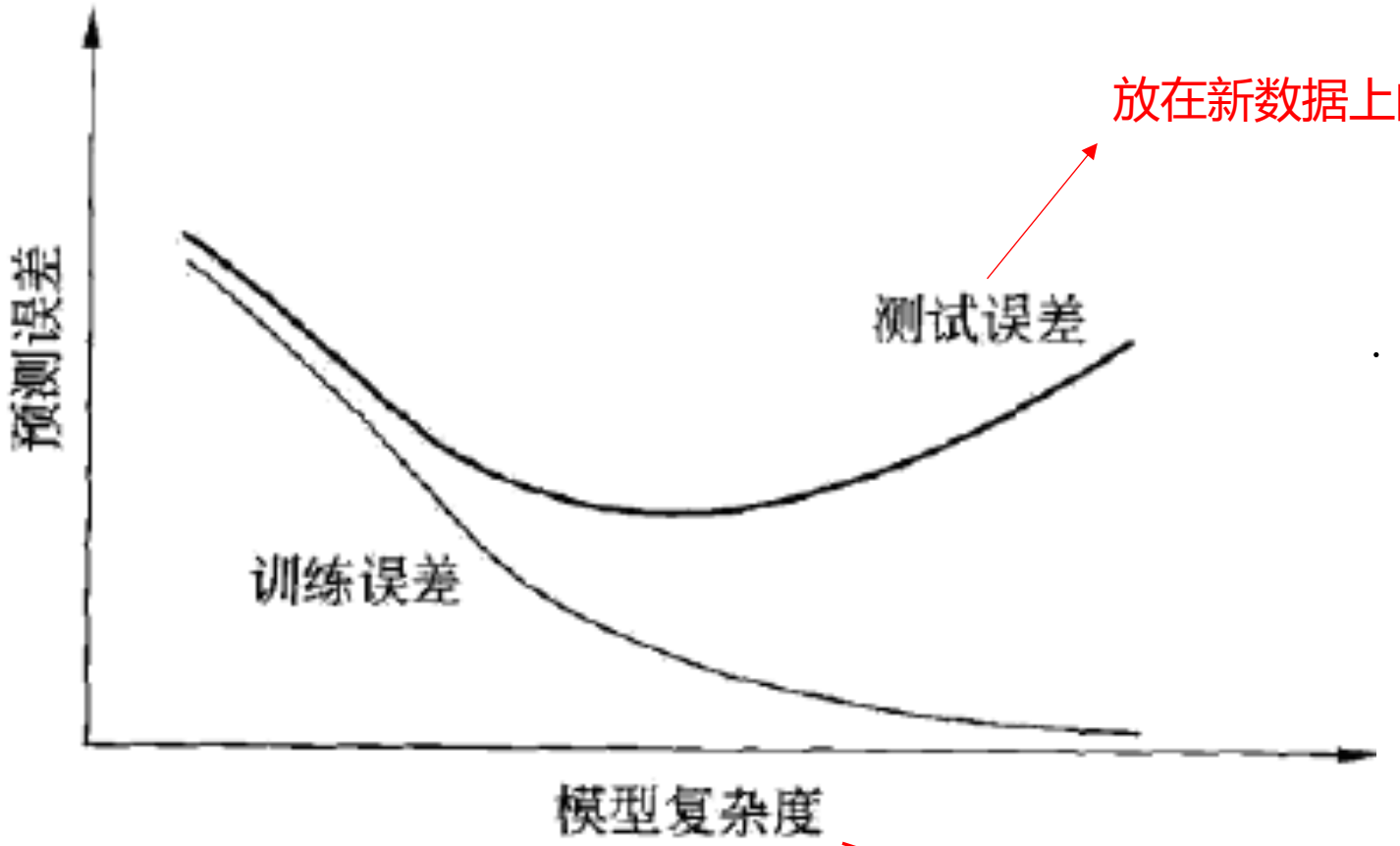
新样本

过拟合模型分类结果:
不是树叶
(误认为树叶必须有锯齿)

欠拟合模型分类结果:
是树叶
(误认为绿色的都是树叶)

过拟合和欠拟合与模型复杂度到底什么关系?

模型评估与选择



放在新数据上的预测误差

1，有没有理论解释?
2，如何获得测试误差?

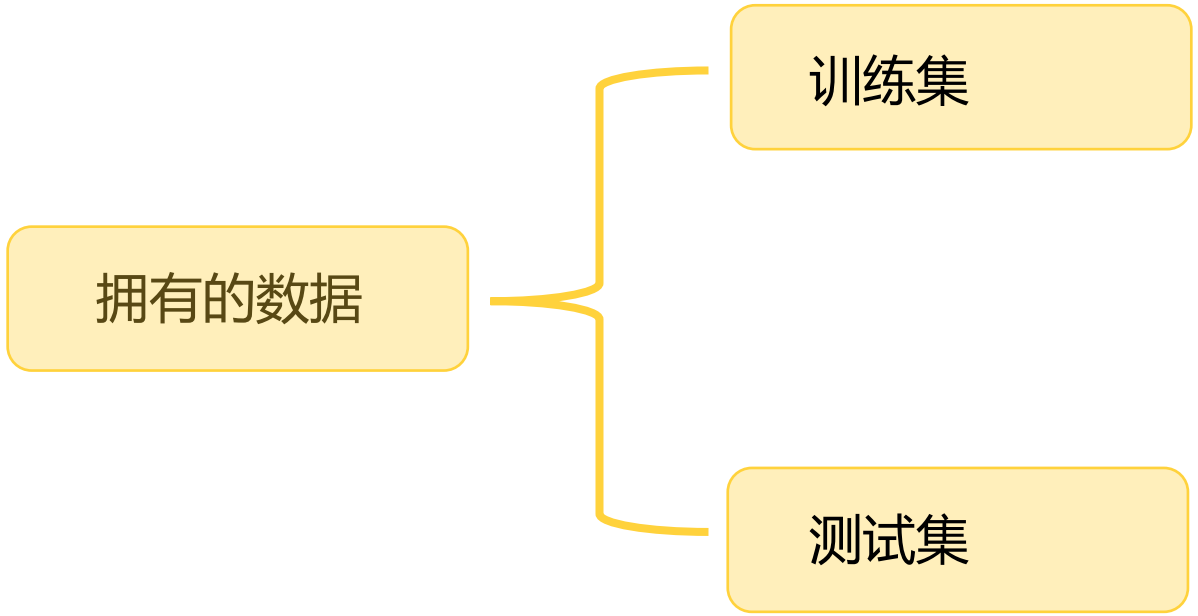
模型参数的个数

模型评估与选择

三个关键问题：

- **如何获得测试结果？** → 评估方法
- **如何评估性能优劣？** → 性能度量
- **如何判断实质差别？** → 比较检验

模型评估与选择



测试误差：

$$e_{\text{test}} = \frac{1}{N'} \sum_{i=1}^{N'} L(y_i, \hat{f}(x_i))$$

训练集（training set）：训练模型

测试集 （testing set）：通过测试集上的“测试误差”作为泛化误差的近似

要求：测试集应尽可能与训练集互斥

关键：怎么获得“测试集”(test set)?

模型评估和模型选择

关键：怎么获得“测试集”(test set)?

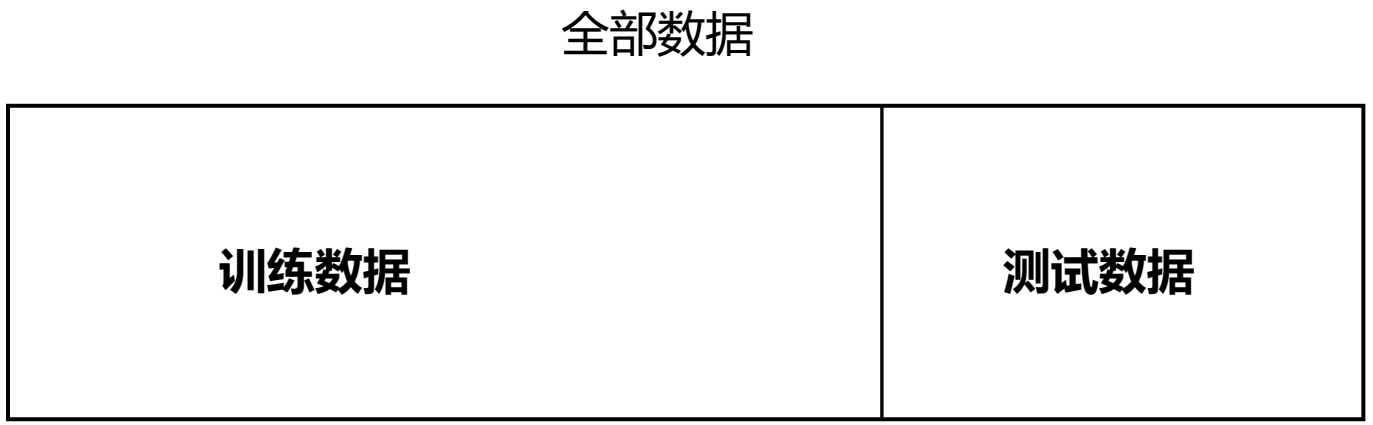
难点：测试集应该与训练集“互斥”

- 根据测试集选择方法的不同，模型选择方法可分为：
  - 留出法、
  - k折交叉验证法
  - 自助法等。

模型评估和模型选择

◆ 留出法：

- 直接将数据集划分为互斥的两个集合，一个作为训练集，一个作为测试集；
- 用训练集在各种条件下（如不同参数）训练模型，从而得到不同的模型；
- 将这些模型带入到测试集上计算测试误差，选择误差最小的模型。



1. 适用于样本多的情形，划分要保持数据分布的一致性
2. 可多次重复随机划分，取测试误差的平均值
3. 训练集的比例一般为 2/3-4/5

模型评估和模型选择

交叉验证法

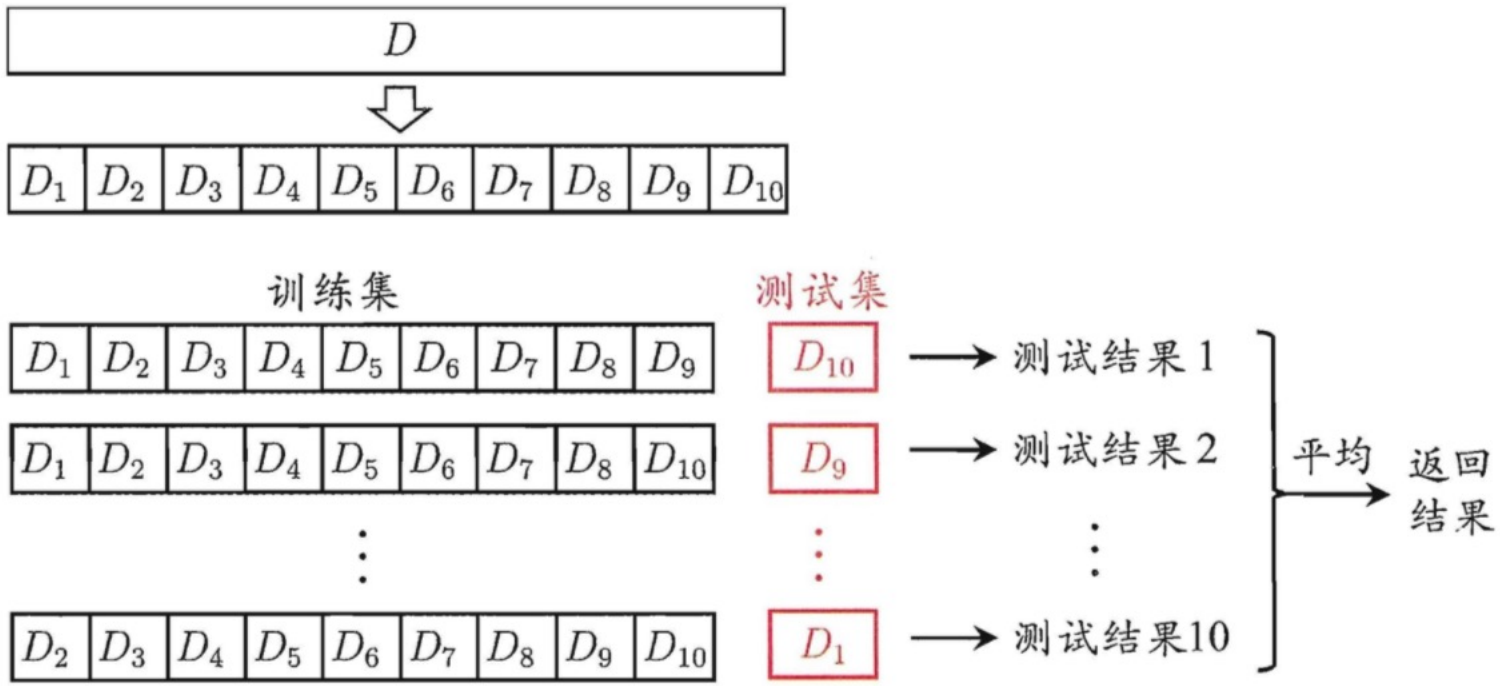
- 交叉验证法“ (cross validation)先将数据集D划分为k个大小相似的互斥子集：

$$D = D_1 \cup D_2 \cup \ldots \cup D_k,\ D_i \cap D_j = \varnothing\ (i \neq j).$$

- 然后，每次用 k-1 个子集的并集作为训练集，用其余的那个子集作为测试集;
- 这样就可获得 k 组训练/测试集，从而可进行 k次训练和测试， 最终返回的是这 k个测试结果的均值
- 通常把交叉验证法称为 " k 折交叉验证" (k-fold cross validation)

模型评估和模型选择

- 交叉验证法



10-折交叉验证示意图

注：

1. 如果k=N,又称留一法 （Leave-one-out）。
2. 与留出法相似，将数据集D划分为k个子集同样存在多种划分方式。为减小因样本划分不同而引入的差别，k折交叉验证通常要随机使用不同的划分重复 p次,最终的评估结果是这 p次k折交叉验证结果的均值 。

模型评估与选择

三个关键问题：

- **如何获得测试结果？** → 评估方法
- **如何评估性能优劣？** → 性能度量
- **如何判断实质差别？** → 比较检验

性能度量

- 对学习器的泛化性能进行评估，不仅需要有效可行的实验估计方法；还需要有衡量模型泛化能力的评价标准，这就是性能度量 (performance measure)。
- 性能度量反映了任务需求，在对比不同模型的能力时，使用不同的性能度量往往会导致不同的评判结果

什么样的模型是“好”的，不仅取决于算法和数据，还取决于任务需求

分类任务常用

- 查准率、查全率、F1度量
- PR图
- ROC、AUC
- Lift提升

回归任务常用

- 均方误差
- 模型拟合优度
- R^2方

性能度量

- 假设测试集的样本量是 m，对应的测试集为

$$D = \{(x_1, y_1), \ldots, (x_m, y_m)\}$$

- 回归任务最常用的性能度量是
- “均方误差” (mean squared error)

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2}$$

- 平均绝对偏差(Mean Absolute Error, MAE)

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m}|f(x_i) - y_i|$$

- 平均相对偏差(Mean Absolute Percentage Error, MAPE)

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{f(x_i) - y_i}{y_i}\right| \times 100\%$$

- 以及R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(f(x_i) - y_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2}$$

混淆矩阵

例：如有200个样本数据，预测为正类，负类两类。分类结束后得到的混淆矩阵为：

		预测	
		正类	负类
实际	正类	120（TP）	25（FN）
	负类	20（FP）	35（TN）

1，准确率是多少？
2，准确率是唯一的指标吗？

性能度量（查准率/查全率）

错误率/正确率的局限性

假定瓜农拉来一车西瓜，我们用训练好的模型对这些西瓜进行判别，显然，错误率衡量了有多少比例的瓜被判别错误.但是若我们关心的是

- “挑出的西瓜中有多少比例是真正的好瓜“---**查准率 （Precision）**
- ” 所有好瓜中有多少比例真正被挑了出来 ----**查全率 （Recall）**

类似的需求在信息检索、web 搜索等应用中经常出现。例如在信息检索 中，我们经常会关心

- "检索出的信息中有多少比例是用户感兴趣的"
- “用户感兴趣的信息中有多少被检索出来了。

性能度量（查准率/查全率）

- 查准率：查准率是指在所有被预测为正例的样本中，实际为正例的比例。换句话说，它是模型预测为正例的准确性。

$$P = \frac{TP}{TP + FP} = \frac{30}{30 + 10} = 0.75$$

- 查全率：在所有实际为正例的样本中，模型正确预测为正例的比例。它衡量的是模型找到所有正例的能力。

$$R = \frac{TP}{TP + FN} = \frac{30}{30 + 5} = 0.857$$

好瓜：正例；坏瓜：反例

实际情况	预测正例	预测反例
正例	真正例（TP)=30	假反例（FN）=5
反例	假正例（FP）=10	真反例（TN）=50

查准率和查全率是一对矛盾的度量，一般来说，查准率高时，查全率往往偏低; 而查全率高时，查准率往往偏低。

性能度量案例

你是复旦大学MBA项目的招生主管，现在你打算进行一个学生满意度调查，目的是了解学生对MBA项目的满意程度。你计划向所有学生发送调查问卷，但你知道，通常调查问卷的回收率只有10%。也就是说，你发送100份问卷，只有10份得到回复。（有效问卷要求回收率达到25%）

请你帮助项目提出对策提高调查问卷的回收率以减少无效的问卷发放?

性能度量案例

1

第一步： 搜集数据

搜集历史发放数据，比如历史发放数据中那些人回复了问卷，哪些人未回复，同时获取客户相应的信息，比如年龄、性别、学习成绩、是否参与过其它项目、教育程度等。

2

第二步： 构建逻辑回归模型

利用获取的历史数据，以是否回复问卷作为响应变量，以用户特征作为预测变量，构建用户是否会回复问卷的逻辑回归模型。

你使用训练集来训练逻辑回归模型，模型通过学习学生的历史数据与是否填写调查问卷之间的关系，来预测新的学生是否会填写问卷。

3

第三步，模型选择

根据所构建的模型，使用测试集来评估模型的性能。你可以ROC或者AUC、Lift提升度来评估模型的区分能力，选择最优的逻辑回归模型和阈值，让指标最大化。

4

第四步: 应用模型预测新学生

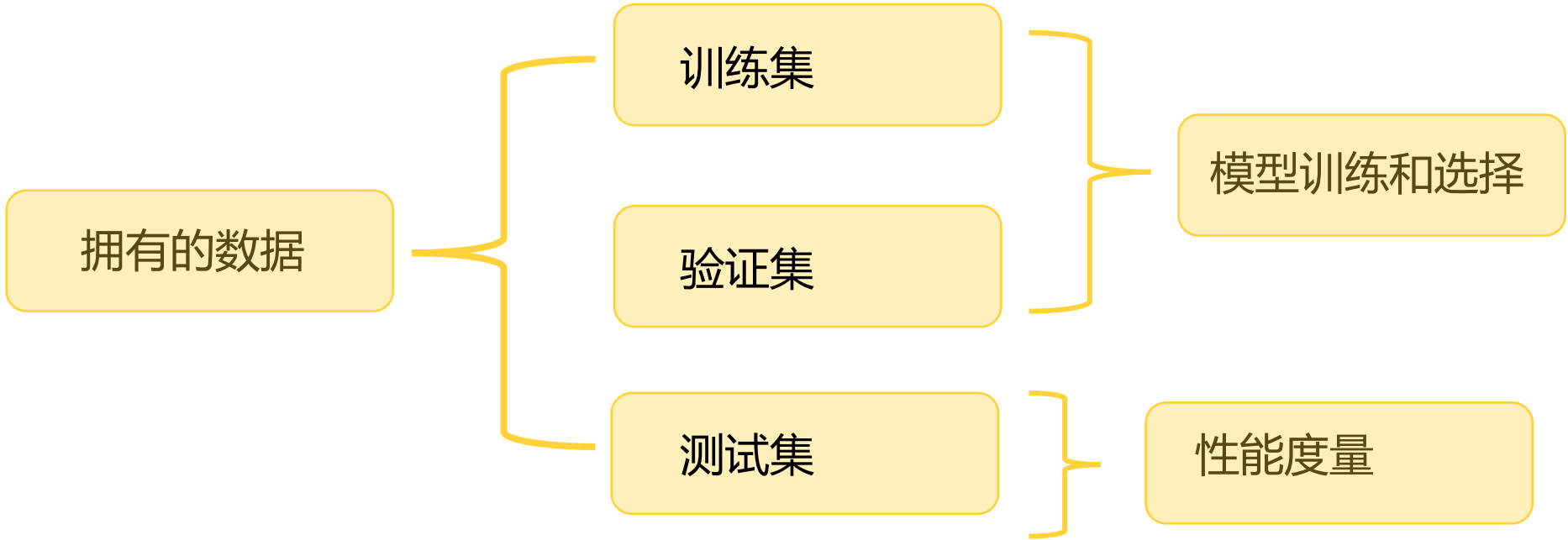
对200名新学生的数据，根据他们的个人信息（如年龄、性别、学习成绩等），利用所选择的模型预测他们是否会填写调查问卷，最终向模型预测为填写的学生发放问卷。

模型评估与选择

三个关键问题：

- **如何获得测试结果？** → 评估方法
- **如何评估性能优劣？** → 性能度量
- **如何判断实质差别？** → 比较检验

模型评估和模型选择



- 实际中，通常把用于模型选择部分的测试集称为验证集（validation set）。
- 在研究对比不同算法的泛化性能时，我们用测试集上的判别效果来估计模型在实际使用时的泛化能力。

章节目录

- 机器学习进阶2
  - **决策树算法**
  - Bagging与随机森林
  - Boosting 算法

决策树

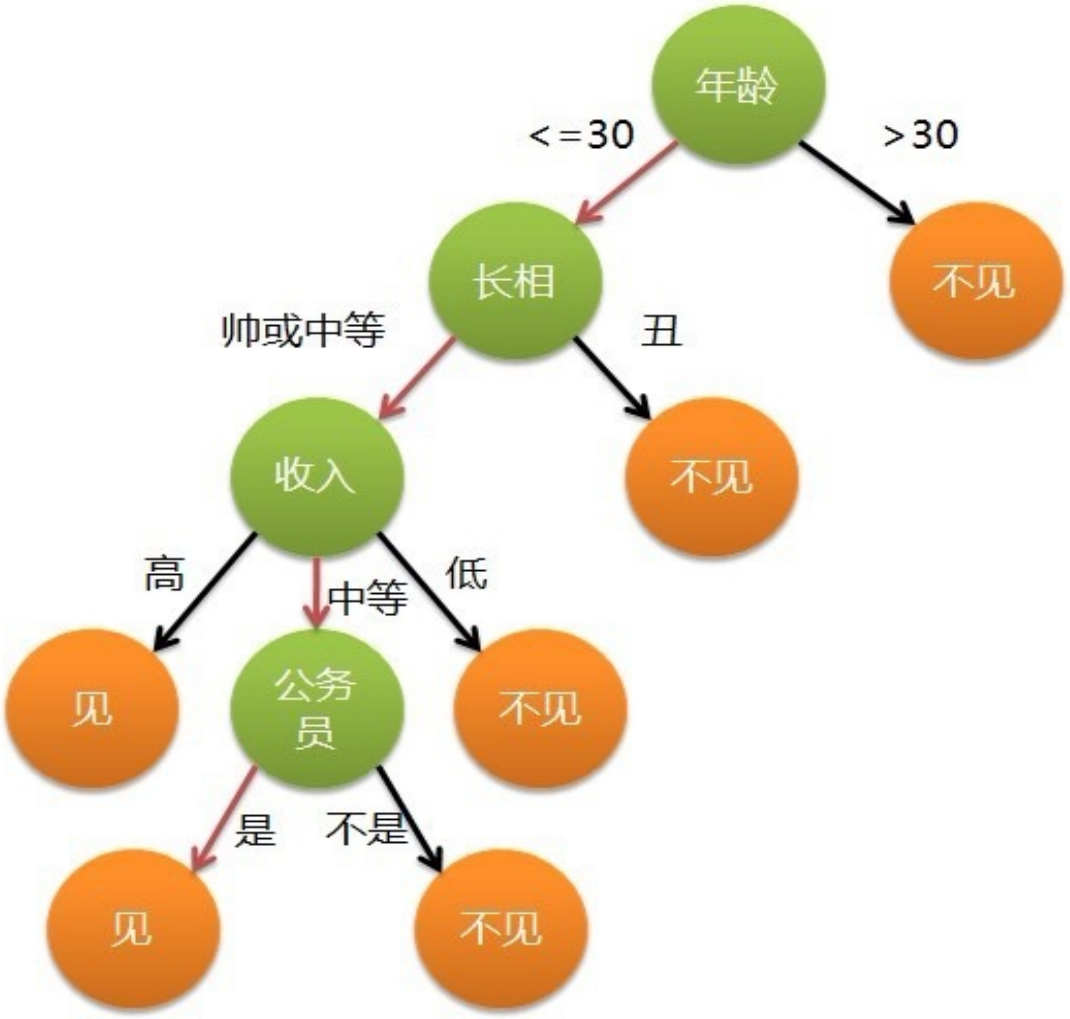
相亲决策树

■属性：年龄、长相、收入、是否公务员

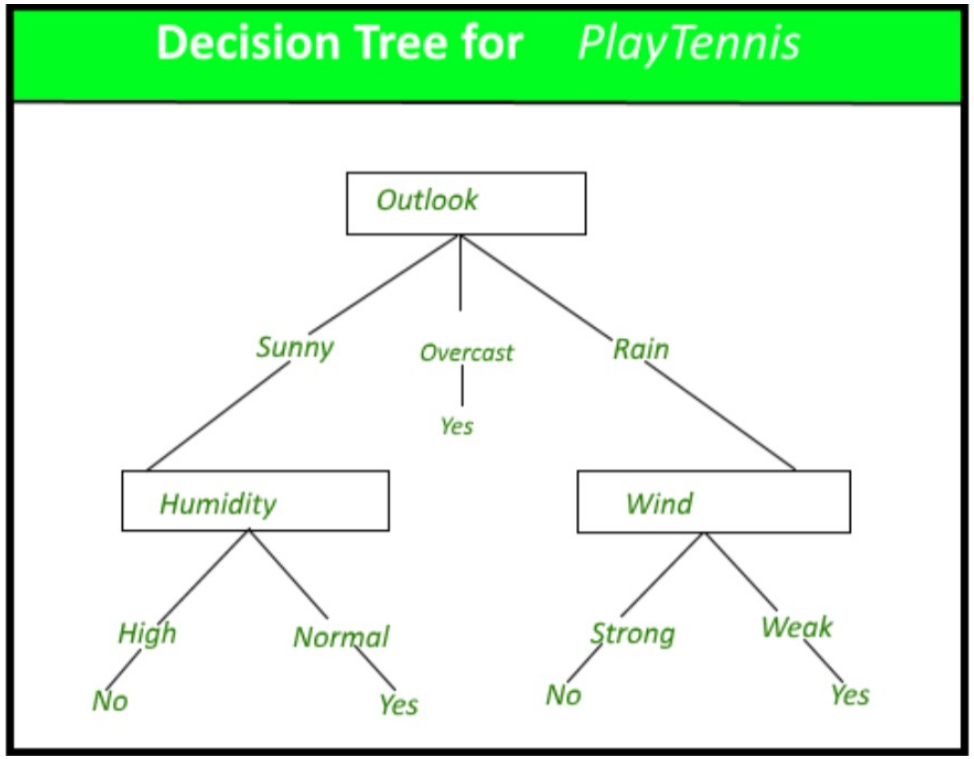
■分类结果：{见，不见}

■决策依据：

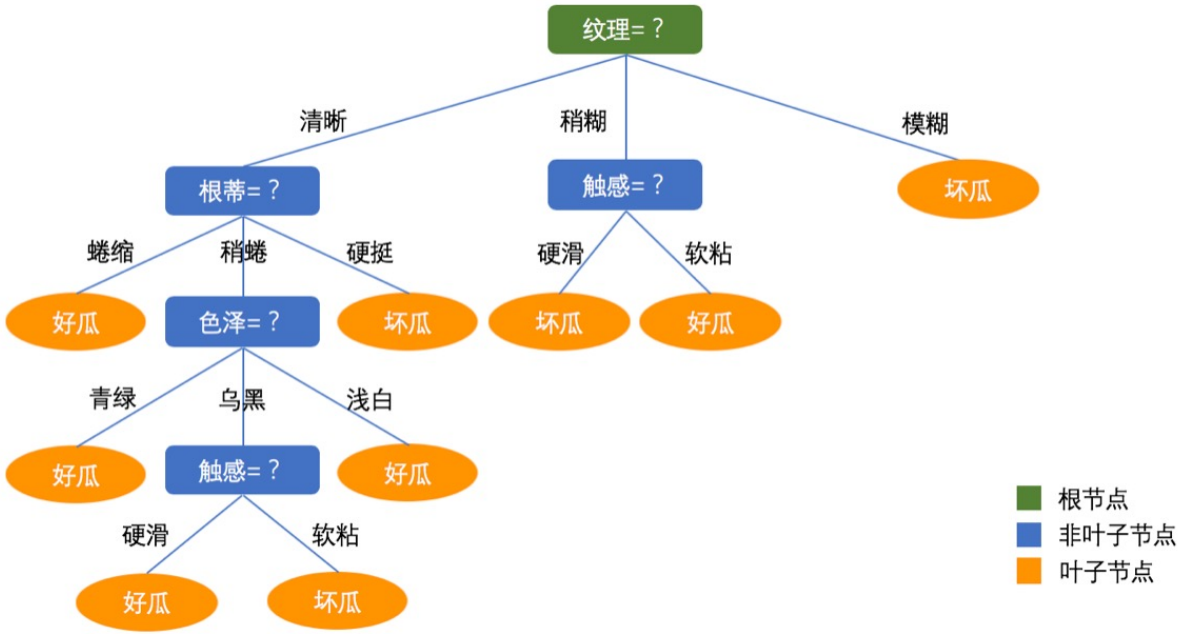
- 年龄>30=不见
- 年龄<30+长相丑=不见
- 年龄<=30+长相中等+收入高=见
- 等等



决策树



打网球决策树分类



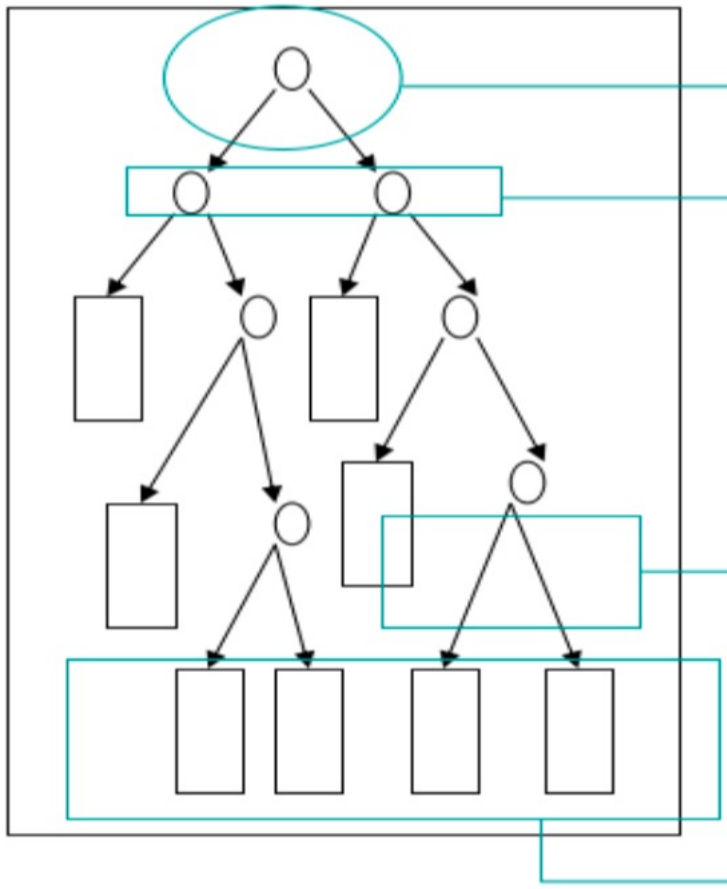
西瓜决策树分类

决策树

- 决策树(**Decision Tree**)，又称为判定树，是数据挖掘技术中的一种重要的分类方法，它是一种以树结构 (包括二叉树和多叉树) 形式来表达的预测分析模型
- 决策树模型常常用来解决分类和回归问题
- 决策树技术发现数据模式和规则的核心是归纳算法：
  - 归纳是从特殊到一般的过程
  - 归纳推理：从若干个事实表征出的特征、特性和属性中，通过比较、总结、概括而得出一个规律性的结论----即从特殊事实到普遍性规律的结论
  - 决策树学习是以样例为基础的归纳学习

决策树学习

- 决策树的构成：
  - 决策树基于“树”结构进行决策的一种分类模型
  - 决策树构成：结点和分支
  - 决策结点：根节点、内部节点、叶结点
- 决策树的学习过程：
  - 通过对训练样本的分析来确定“划分属性和分支”（即结点所对应的属性）
- 预测过程：
  - 对新示例沿着决策树从上到下进行遍历，在每个结点都有一个属性测试。对每个结点根据不同属性测试将会进入不同的分枝，最后会达到一个叶子结点。



决策树

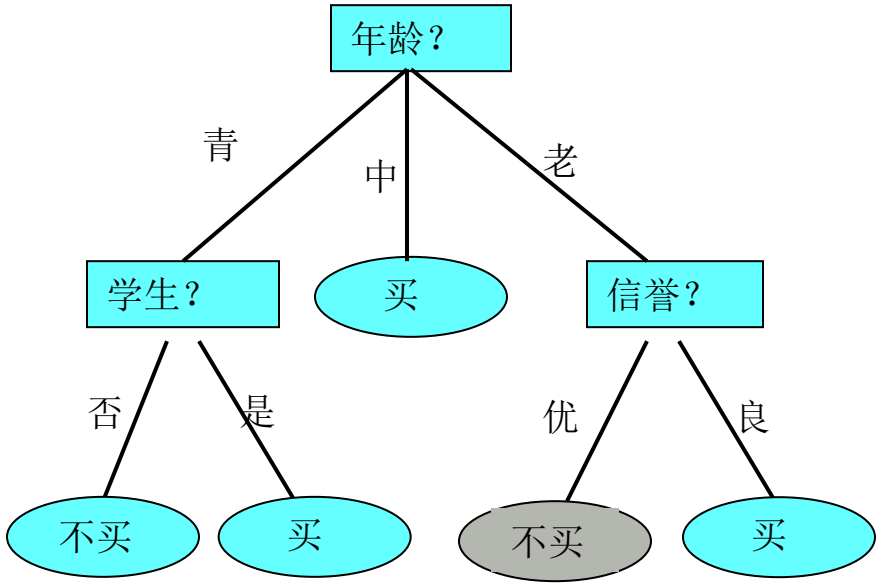
- 假定公司收集了右表数据，那么对于任意给定的客人（测试样例），你能帮助公司将这位客人归类吗？
- 即：你能预测这位客人是属于“买”计算机的那一类，还是属于“不买”计算机的那一类？

- 属性={年龄,收入，学生，信誉}
- 类别={买，不买}
- 年龄={青，中，老}
- 收入={低，中，高}
- 学生={是，否}
- 信誉={良、优}

计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	中	高	否	良	买
60	老	中	否	良	买
64	老	低	是	良	买
64	老	低	是	优	不买
64	中	低	是	优	买
128	青	中	否	良	不买
64	青	低	是	良	买
132	老	中	是	良	买
64	青	中	是	优	买
32	中	中	否	优	买
32	中	高	是	良	买
63	老	中	否	优	不买
1	老	中	否	优	买

决策树

谁在买计算机？



计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	中	高	否	良	买
60	老	中	否	良	买
64	老	低	是	良	买
64	老	低	是	优	不买
64	中	低	是	优	买
128	青	中	否	良	不买
64	青	低	是	良	买
132	老	中	是	良	买
64	青	中	是	优	买
32	中	中	否	优	买
32	中	高	是	良	买
63	老	中	否	优	不买
1	老	中	否	优	买

决策树

决策树构建的关键（难点）：

①如何对中间结点进行属性划分（特征选择）

②何时令某个结点为叶节点？（终止条件）

③如果落入叶节点的样本不都属于同一类，如何给该叶节点赋类别标记?

④如何具体构建一颗决策树（决策树的构建算法）

⑤如果树生长的过大，如何使其变小变简单，即如何剪枝?

特征选择

- 特征选择在于选择对训练数据具有分类能力的特征

思考：如何度量某个特征的分类能力？

- 换句话讲，我们希望给定某个特征以后，类别的 Y 不确定性 会很小

思考：如何度量类别 Y 的不确定性？

信息增益

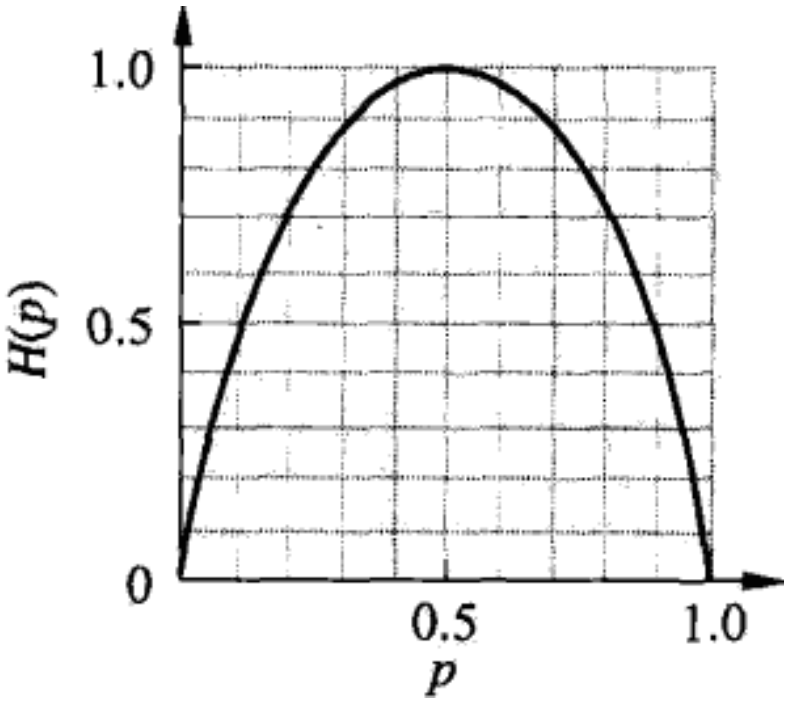
■ 熵：设类别变量 Y 为离散型随机变量，$P(Y=k)=p_k,\ k=1,\ldots,K,$ 其熵 (entropy)定义为：

$$H(Y)=-\sum_{k=1}^{K} p_k \log p_k$$

注：

- 熵是随机变量不确定性的一种度量
- 熵越大，随机变量的不确定性越大
- 可以验证，$0 \le H(Y) \le \log K$
- 如果 $H(Y)=0$，表示 Y 只有一种类别

■ 当 Y 为 0-1 分布时



注：最常用的是以2为底，单位为bit;在理论推导中常采用e为底，单位为Nat

信息增益

信息熵越大，意味着不确定性越大；信息熵越小，则不确定性越小。当熵为0时，表示不确定性为0.

- 例如，对于 一盏灯两种可能的状态：开和关。假设样本集合中一半是开的样本，一半是关的样本，那么信息熵为：

$$H(\boldsymbol{Y}) = -\sum_{k=1}^{K} \widehat{P}_k \cdot \log_2 \widehat{P}_k = -(\frac{1}{2} \cdot \log_2 \frac{1}{2} + \frac{1}{2} \cdot \log_2 \frac{1}{2}) = 1$$

类别	1（开）	2（关）
概率	1/2	1/2

在此情况下，信息熵很大，说明灯的两种状态不确定性很大。

- 假如这盏灯出现故障，无法正常打开，那么此时信息熵为：

$$H(\boldsymbol{Y}) = -\sum_{k=1}^{K} \widehat{P}_k \log_2 \widehat{P}_k = -(\frac{1}{1} \cdot \log_2 \frac{1}{1}) = 0$$

类别	1（开）	2（关）
概率	0	1

在此情况下，灯的状态是确定的，因此信息熵最小。

信息增益

类似的例子还有掷骰子，

- 假如骰子是公平的，即6个面出现的概率相同，均为1/6，那么信息熵为：

$$H(Y) = -\sum_{k=1}^{6} P_k \log_2 P_k = -6 \times \left(\frac{1}{6}\log_2 \frac{1}{6}\right) = \log_2 6 \approx 2.585$$

骰子点数Y	1	2	3	4	5	6
概率$P(Y)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

在此情况下，信息熵最大，说明骰子的点数的不确定性最大

- 假如骰子故障，总是掷出6，那么此时信息熵为：

$$H(Y) = -\sum_{k=1}^{6} P_k \log_2 P_k = 0 + 0 + 0 + 0 + 0 + 0 = 0$$

骰子点数Y	1	2	3	4	5	6
概率$P(Y)$	0	0	0	0	0	1

在此情况下，骰子的状态是确定的，因此信息熵最小。

信息增益

◆ 你每天要根据当天的天气情况决定是否外出。
◆ 而你是否出门可能有两种结果：**出门**或者**不出门**。
◆ 天气可能有三种状态：**晴天**、**雨天**和**阴天**。
◆ 我们希望计算在给定天气的情况下，你是否出门的不确定性，也就是条件熵。

特征：天气={晴天，阴天，雨天}

类别={出门，不出门}

◆ 从过去的经验中统计出以下信息：

- 三种天气出现的可能性相同，各为1/3
- **晴天**时，你出门的概率是70%，不出门的概率是30%。
- **雨天**时，你出门的概率是20%，不出门的概率是80%。
- **阴天**时，你出门的概率是50%，不出门的概率也是50%。

条件概率 ⇒

天气状态	外出概率	不外出概率
晴天	0.7	0.3
雨天	0.2	0.8
阴天	0.5	0.5

◆ 此外，我们还知道你出门和不出门的概率各位1/2，因此首先可以未给定天气时的不确定性为

$$H(Y) = -0.5\log_2 0.5 - 0.5\log_2 0.5 = 1$$ ⇒ **不考虑天气时出不出门的不确定性**

信息增益

- 晴天时的熵：

$$H(Y|\text{晴天}) = -0.7\log_2 0.7 - 0.3\log_2 0.3 = 0.8813$$

- 雨天时的熵

$$H(Y|\text{雨天}) = -0.2\log_2 0.2 - 0.8\log_2 0.8 = 0.7219$$

- 阴天时的熵

$$H(Y|\text{阴天}) = -0.5\log_2 0.5 - 0.5\log_2 0.5 = 1$$

- 天气总的条件熵为

$$\begin{aligned} &H(Y|X) \\ &= P(\text{晴天}) * H(Y|\text{晴天}) + P(\text{雨天}) * H(Y|\text{雨天}) + P(\text{阴天}) * H(Y|\text{阴天}) \\ &= \tfrac{1}{3} \times 0.8813 + \tfrac{1}{3} \times 0.7219 + \tfrac{1}{3} \times 1 = 0.8677 \end{aligned}$$

天气状态	外出概率	不外出概率
晴天	0.7	0.3
雨天	0.2	0.8
阴天	0.5	0.5

天气的信息增益

$$\begin{aligned} IG(Y,X) &= H(Y) - H(Y|X) \\ &= 1 - 0.8677 = 0.1323 \end{aligned}$$

特征：天气={晴天，阴天，雨天}

类别={出门，不出门}

信息增益

- 信息增益：表示得知随机变量 X 的信息使得随机变量 Y 信息的不确定性减少的程度:

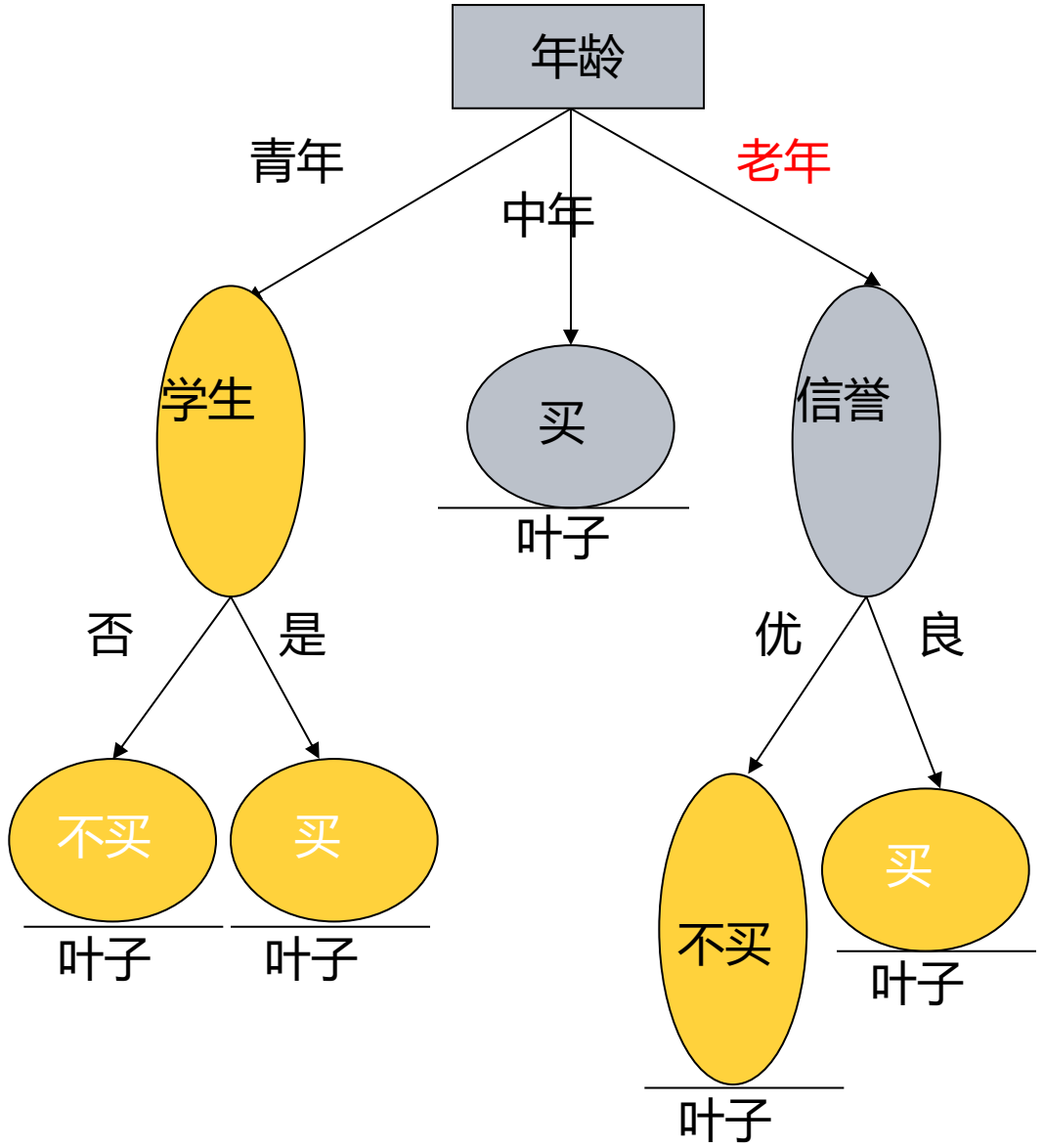
$$I(Y;X) = H(Y) - H(Y|X)$$

- 又称 $I(Y;X)$ 为互信息 (mutual information)

- 互信息描述了随机变量 X 提供的关于Y的信息量的大小
- 信息增益越大表明随机变量X所能提供的关于Y的信息越多！

ID 3算法

计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	中	高	否	良	买
60	老	中	否	良	买
64	老	低	是	良	买
64	老	低	是	优	不买
64	中	低	是	优	买
128	青	中	否	良	不买
64	青	低	是	良	买
132	老	中	是	良	买
64	青	中	是	优	买
32	中	中	否	优	买
32	中	高	是	良	买
63	老	中	否	优	不买
1	老	中	否	优	买



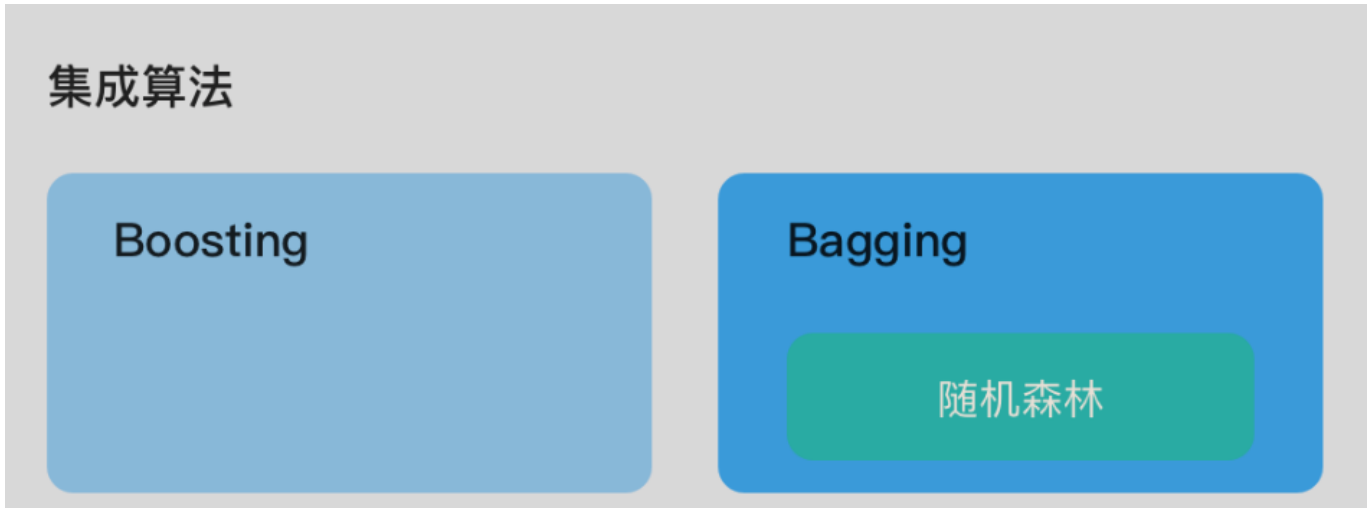
章节目录

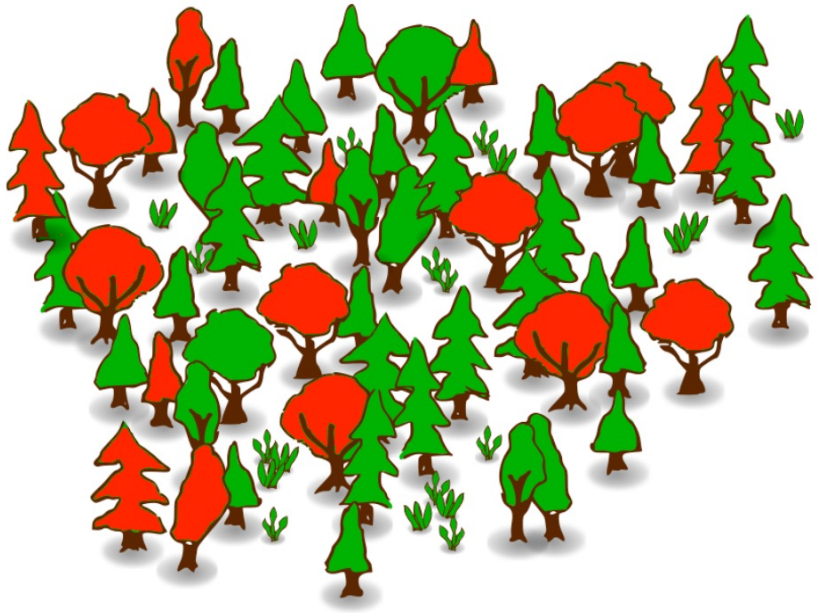
- 机器学习进阶2
  - 决策树算法
  - Bagging与随机森林
  - Boosting 算法

Random Forest

随机森林 (随机决策森林) :

- 是用于分类，回归和其他任务的集成学习方法，其通过在训练时构建多个决策树并输出类的众数（分类）或平均预测（回归）来进行操作。
- 随机森林拥有广泛的应用前景，从市场营销到医疗保健保险甚至金融量化投资交易。
- Kaggle数据科学竞赛，参赛者对随机森林的使用占有相当高的比例。

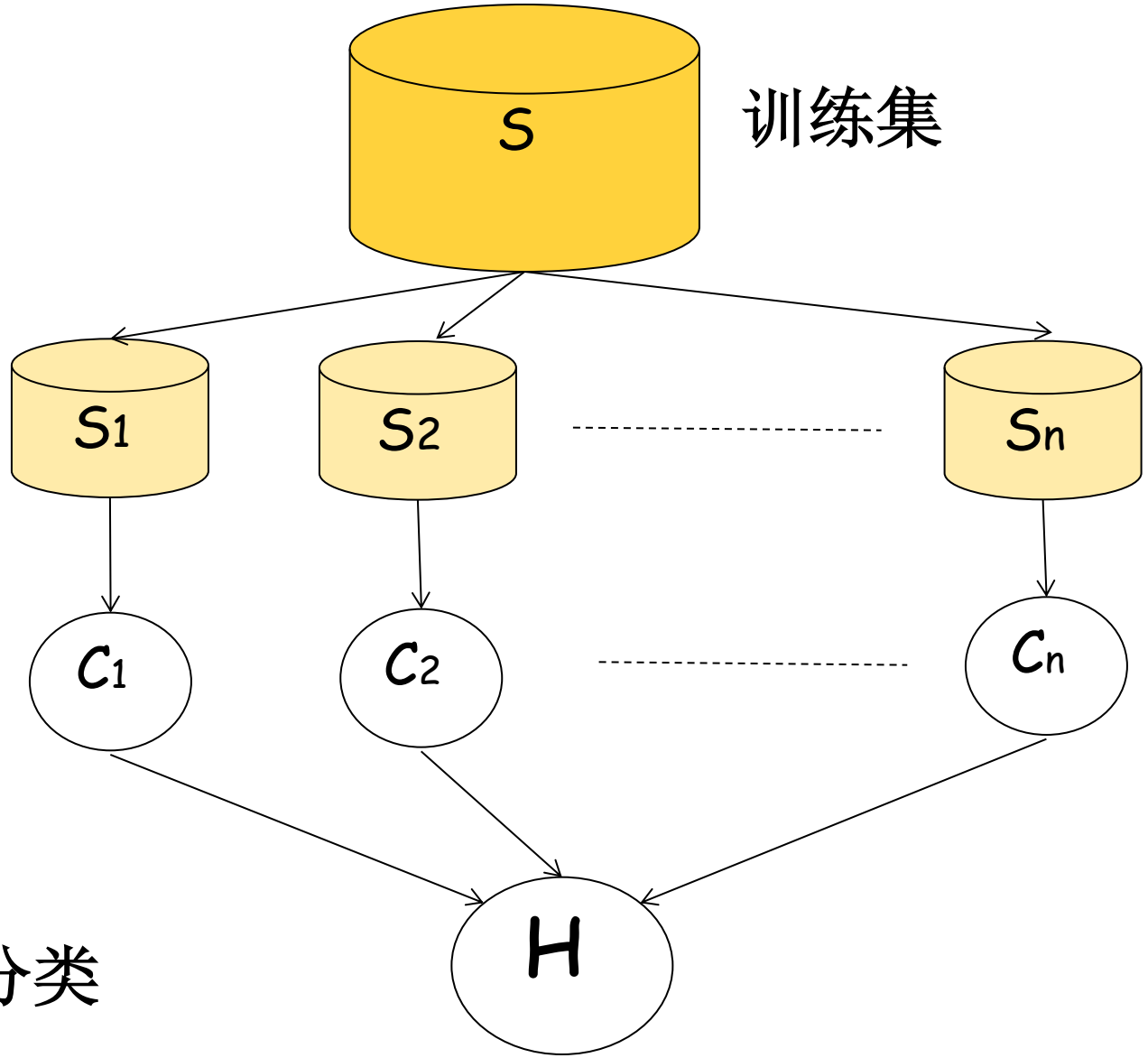




随机森林=随机+森林

集成学习思想

机器学习中的**集成算法（Ensemble Learning）**，是指通过**将多个模型组合起来**，以提升整体预测性能的一种方法。就像“多个臭皮匠顶个诸葛亮”，单个模型可能有偏差或容易过拟合，但多个模型合起来往往可以相互补充、降低误差、提升稳定性。



原理

- 假设有 25 个分类器，
- 每个分类器的犯错误的概率为 $\varepsilon = 0.35$
- 所有分类器是独立的，
- 则组合分类器犯错误的概率为:

$$\sum_{i=13}^{25} \binom{25}{i} \varepsilon^{i} (1-\varepsilon)^{25-i} = 0.06$$

假设你在看一场比赛，裁判团有 **25个评委（模型）**，他们要判断一个模糊的视频回放是否“进球”。

每个评委判断“进球”还是“没进”，有点靠感觉：

- 每个评委**判断错误的概率是 35%（ε = 0.35）**
- 每个评委**互相不交流、完全独立做决定**
- 规则是：**只要13个以上说“进了”，就算“进了”**

集成学习

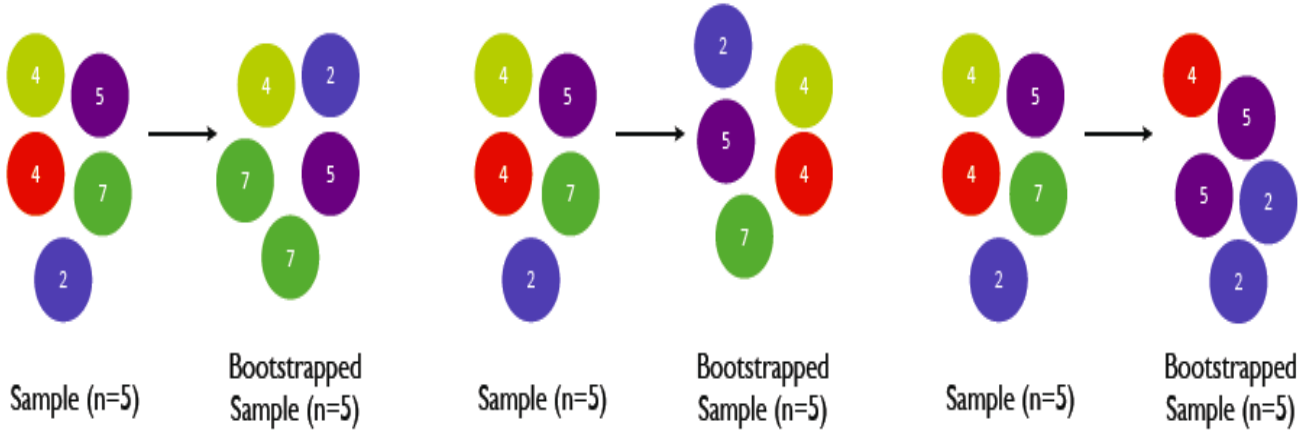
- 集成学习大致可分为两大类：
  - Boosting：Adaboost+ GBDT+XGBoost +LightGBM
  - Bagging （Bootstrap Aggregation）与随机森林

- Bootstrap
  - Bootstrap 方法是在1979年由斯坦福大学统计系的Bradley Efron提出的。
  - Bootstraping的名称来自成语“pull up by your own bootstraps”, 意思是依靠你自己的资源，称为自助法。
  - 它是一种有放回的抽样方法
  - 统计学中非参数统计方法

- Bootstrap本义是指高靴子口后面的悬挂物，小环，带子， 是穿靴子时用手向上拉的工具。
- “pull up by your own bootstraps”即“通过拉靴子让自己上升”，意思是“不可能发生的事情”。
- 后来意思发生了转化，隐喻“不需要外界帮助， 仅依靠自身力量让自己变得更好”。

集成学习

自助法：



- 假设我们有一个小数据集，表示5个人的收入（单位：万元）：

人员编号	收入 (万元)
1	40
2	50
3	45
4	60
5	55

第一次自助抽样：[50, 40, 40, 60, 55]

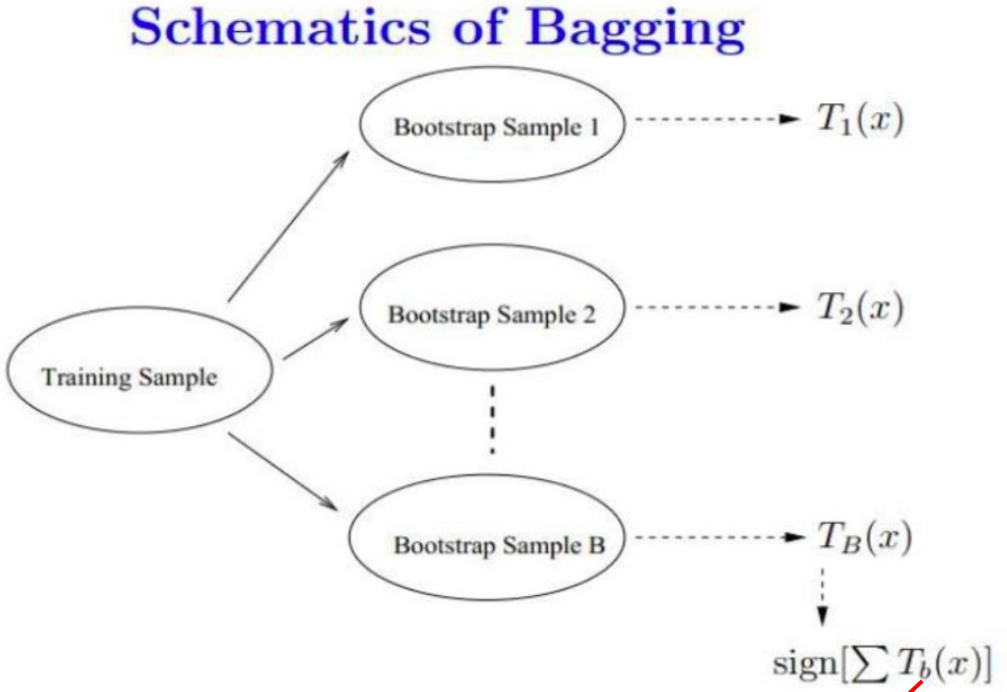
第二次自助抽样：[45, 45, 50, 60, 45]

第三次自助抽样：[40, 40, 60, 55, 55]

Bagging

Bootstrap aggregation

- 从训练样本集中有重复的选出n个样本， 对这n个样本建立分类器或预测器 (例如决策树、 SVM、Logistic回归、线性回归等)
- 重复以上步骤 B 次，即获得了B个分类器或预测器
- 将新的数据分别放在 B 个学习器上：
  - 分类：根据这B个分类器的投票结果决定数据属于哪一类（少数服从多数）
  - 回归：最终取这B个预测器的平均结果

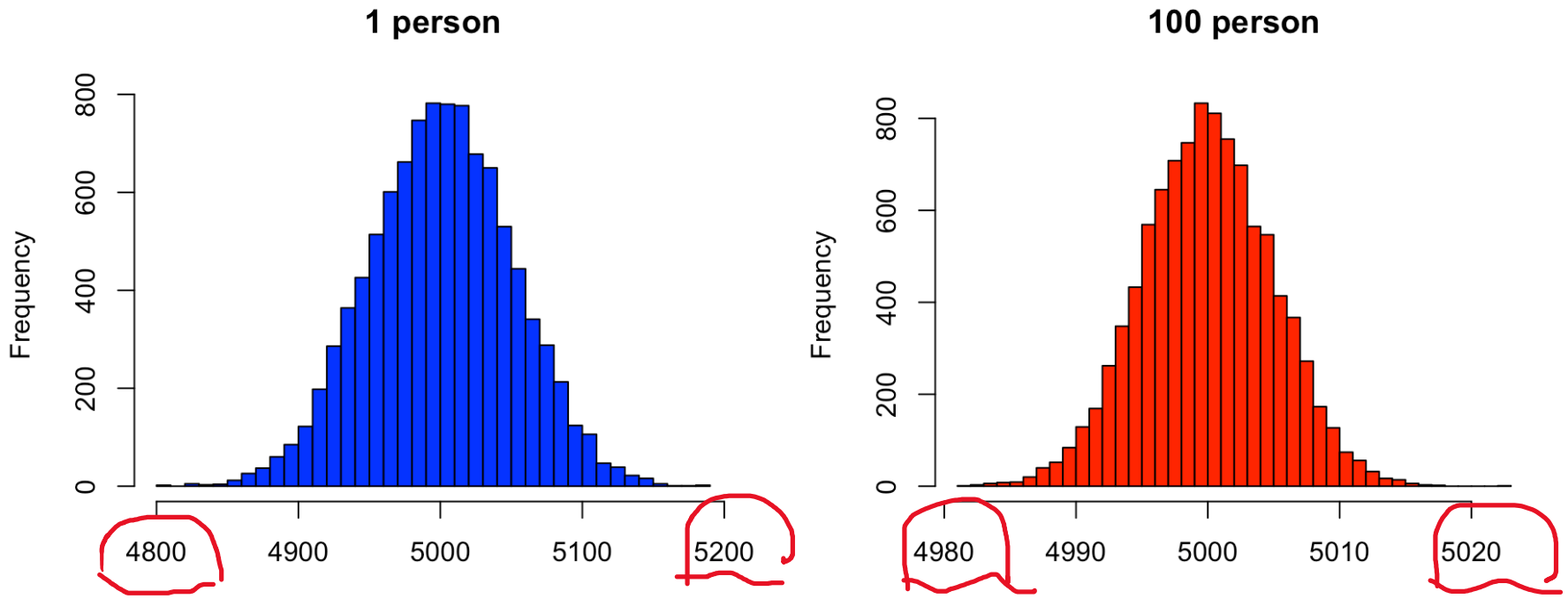


Bagging是并行式集成学习算法最著名的代表

二分类
{-1,1}

Bagging统计解释

- 假设我们要估计某个城市的平均月收入。我们知道这个城市的居民收入大致遵循正态分布，且其总体平均月收入为 **5000 元**，收入的总体方差为 **2500 元**。
- 两种方法：随机抽取1人问他的收入；随机抽取100人算收入的平均

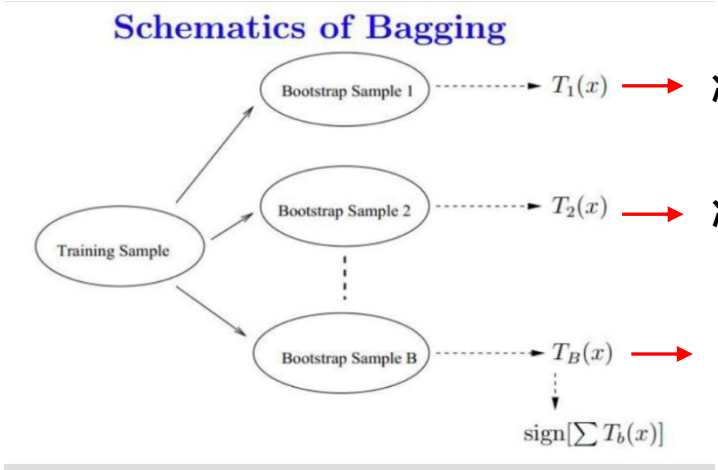


随机森林与Bagging

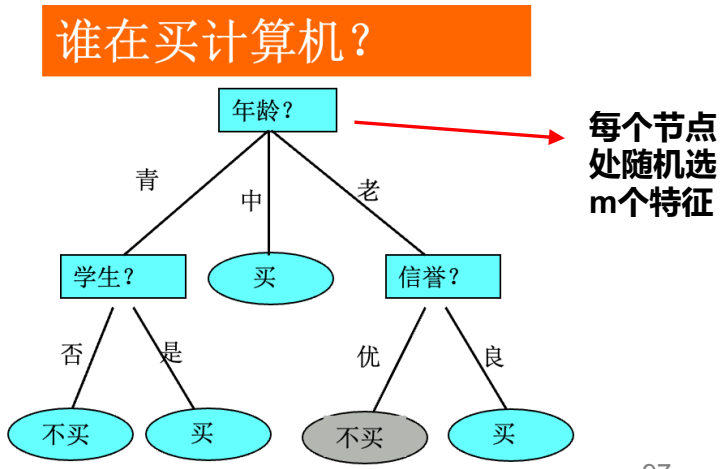
- 随机森林(Random Forest，简称RF)是bagging的一个扩展变种
  - 核心1: Bagging 中令基学习器为决策树
  - 核心2: 考虑了属性选择的随机性

从所有属性中随机选择m个属性，

从中选择最佳分割属性作为节点建立CART决策树；



- 上世纪八十年代Breiman等人发明CART算法 (Breiman et al. 1984)
- 2001年Breiman把分类树组合成随机森林 (Breiman 2001a)
- 被誉为当前最好的算法之一 (Iverson et al. 2008)

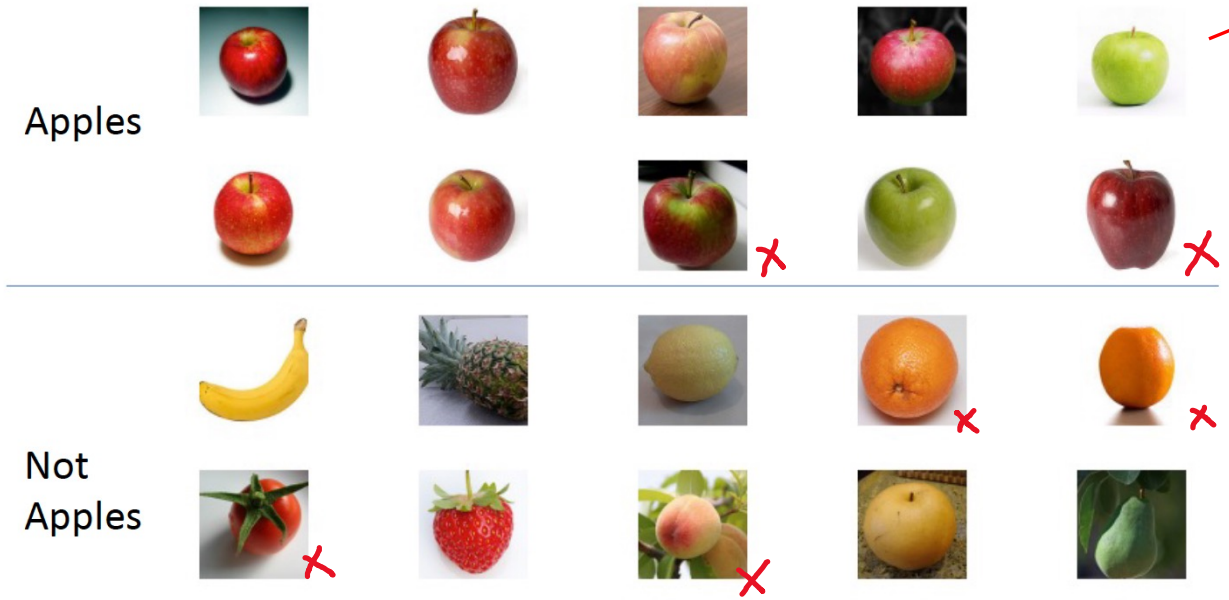


章节目录

- 机器学习进阶2
  - 逻辑回归与二元分类问题
  - 决策树算法
  - Bagging与随机森林
  - Boosting 算法

Boosting算法思想

训练识别苹果的“专家”

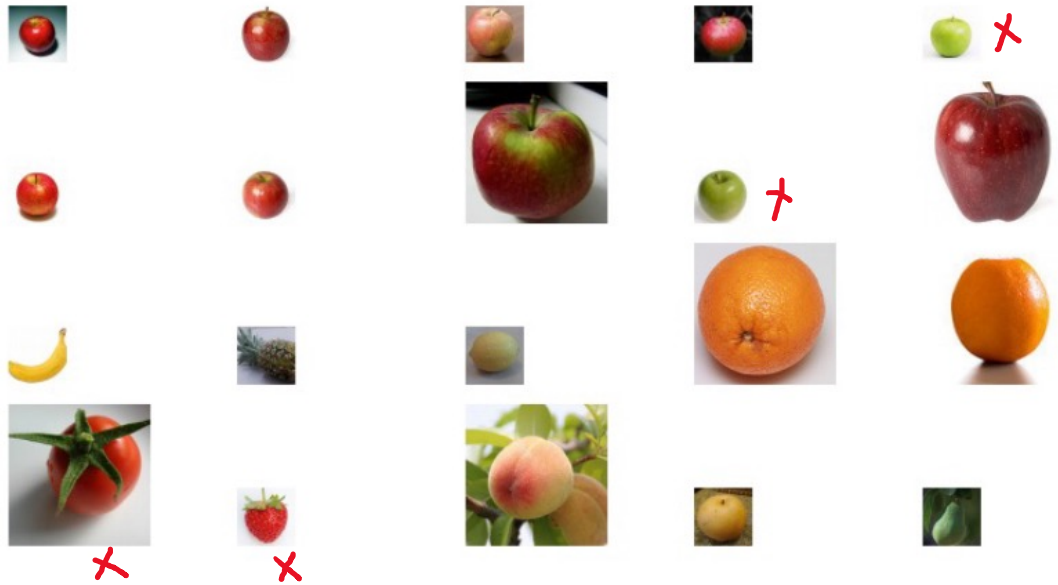


- 张三同学：苹果是圆的

- 老师：虽然这个特征有用，但是这个规则可能会导致一些错误，比如一些非苹果的圆形水果（如橙子、柠檬等）也会被错认为苹果。同时也会识别不出一些非圆的苹果
- 教师开始**标记错误**，
  - 把识别正确的图片变小
  - 把识别错误的图片变大

Boosting算法思想

训练识别苹果的“专家”

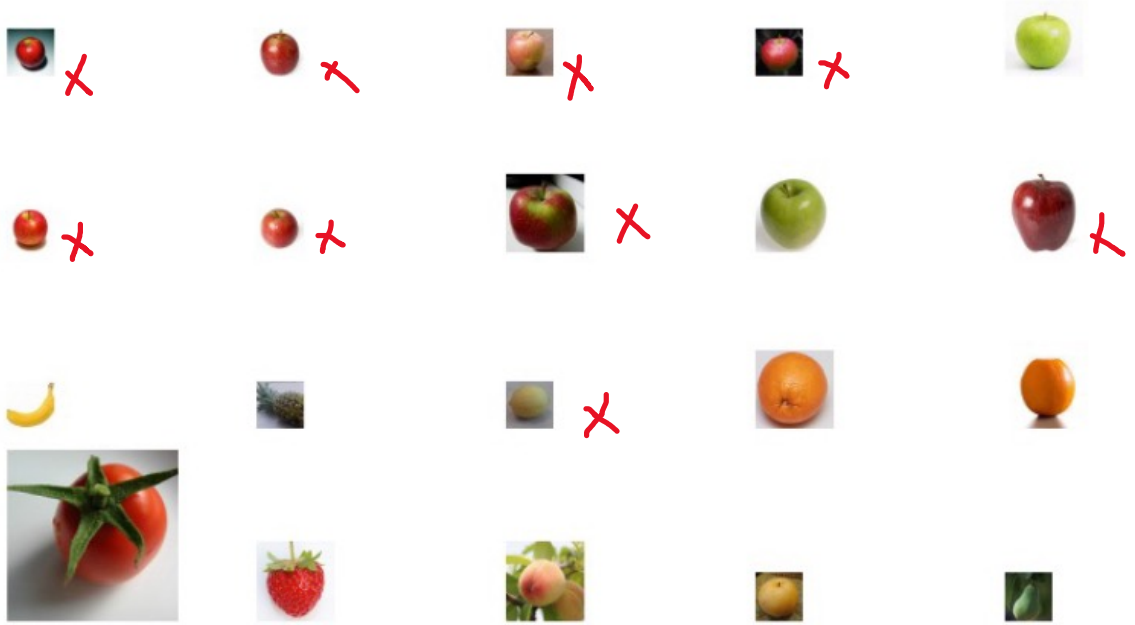


- 张三同学：苹果是圆的
- 李四：苹果是红的

- 老师：虽然这个特征也有用，但是这个规则可能会导致一些错误，比如一些非苹果的红色水果（如西红柿等）也会被错认为苹果。同时也会识别不出一些绿色的苹果
- 教师再次**标记错误**，
  - 把识别正确的图片变小
  - 把识别错误的图片变大

Boosting算法思想

训练识别苹果的“专家”

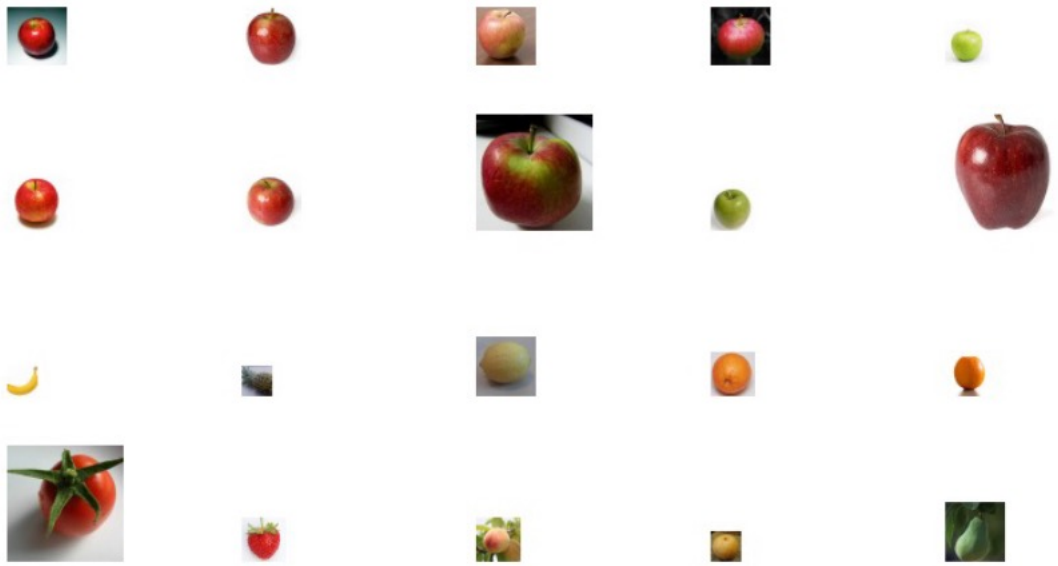


- 张三同学：苹果是圆的
- 李四：苹果是红的
- 王五：苹果是绿色的

- 老师：虽然这个特征也有用，但是这个规则可能会导致一些错误，比如一些非苹果的绿色水果（如桃子等）也会被错认为苹果。同时也漏掉了红色的苹果
- 教师再次**标记错误**，
  - 把识别正确的图片变小
  - 把识别错误的图片变大

Boosting算法思想

训练识别苹果的“专家”



- 张三同学：苹果是圆的
- 李四：苹果是红的
- 王五：苹果是绿色的
- 陈六：苹果上面有梗

经过多次迭代，学生们能够总结出：苹果是“有点圆的、有点红的、可能是绿色的、并且可能有茎”。

AdaBoost基本概念

Boosting的核心思想就是：对于一个复杂的任务而言，有效的综合多个专家的预测进而所得出的新的预测，要优于其中任何一个专家的单独预测。即通常所说的“**三个臭皮匠顶个诸葛亮**”的道理

$$h_1(x) \in \{-1, +1\}$$
$$h_2(x) \in \{-1, +1\}$$
$$\vdots$$
$$h_T(x) \in \{-1, +1\}$$

Weak classifiers

$$H_T(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

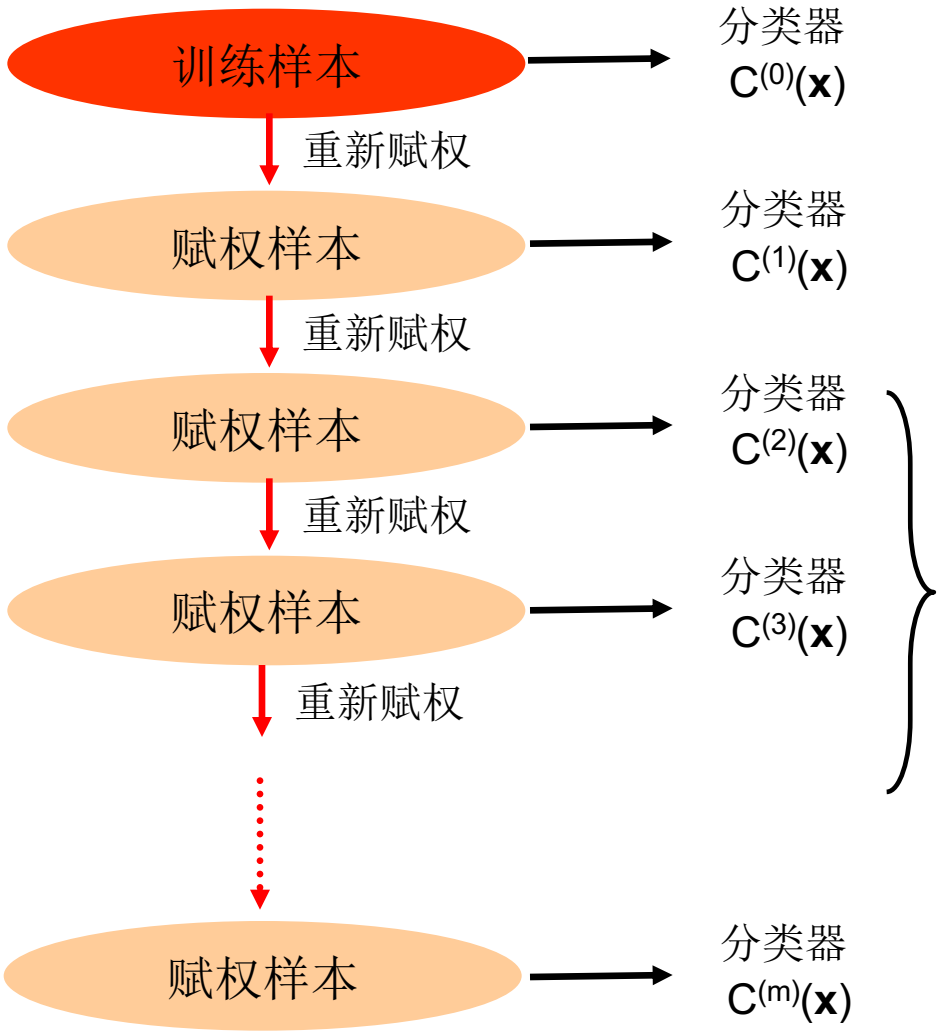
strong classifier

slightly better than random

Boosting算法

- 对每个训练样本赋予一个权值，每次用训练完的新分类器标注各个样本。
- 若某个样本点已被分类正确，则将其权值降低，并以该权重进行下一次模型的训练；
- 若样本点未被正确分类，则提高其权值，并以该权重进行下一次模型的训练。
- 权值越高的样本在下一次训练中所占的比重越大，即越难区分的样本在训练过程中会变得越来越重要。 整个迭代过程直到错误率足够小或达到一定次数才停止。

提升法（Boosting）



$$y(x) \square \sum_{i}^{N_{Classifier}} w_i C^{(i)}(x)$$

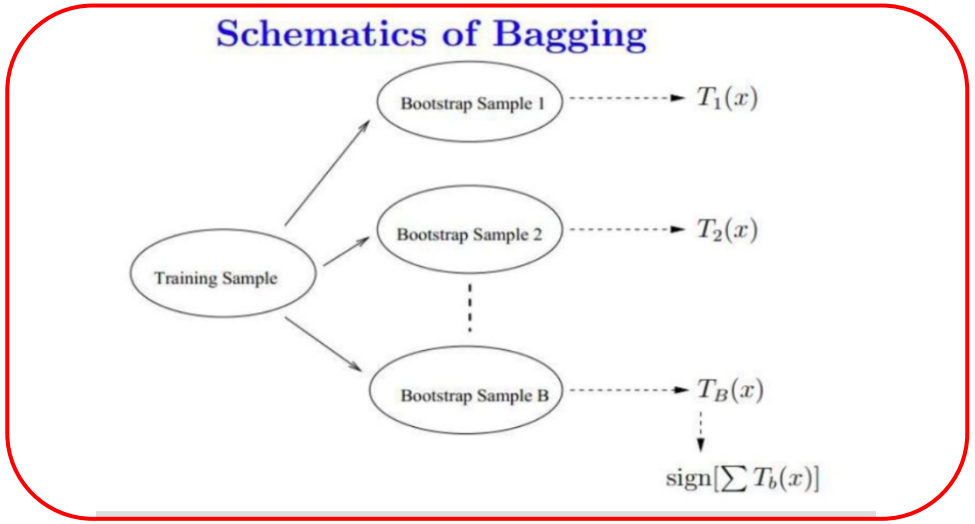
第i学习器的权重

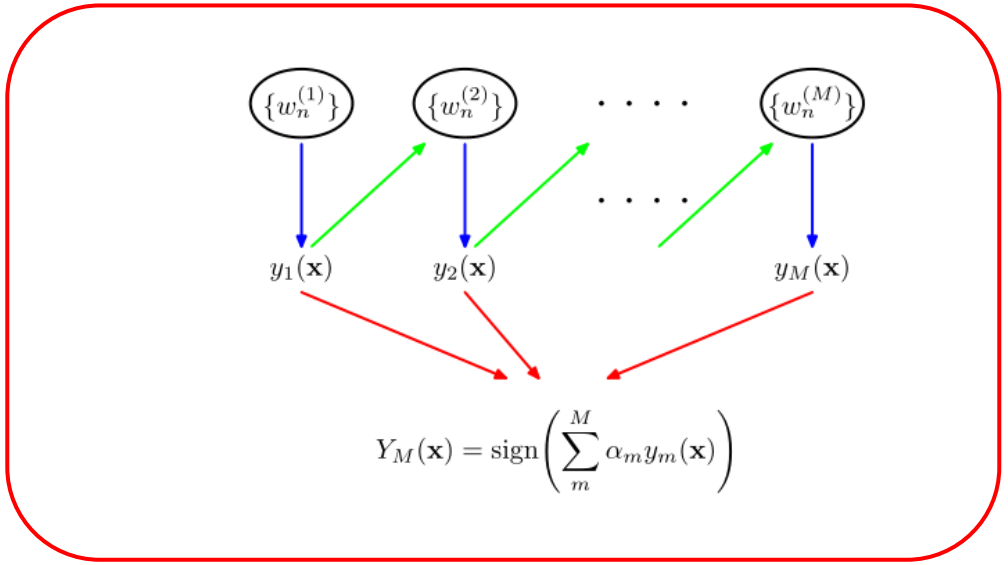
Adaboost、GBDT、XGBoost

Bagging，Random Forest，Boosting，和GBDT

（1）个体学习器间不存在强依赖关系，可同时生成的并行化方法，代表为Bagging/Random Forest。

（2）个体学习器间存在强依赖关系，必须串行生成的序列化方法，代表为Boosting，AdaBoost，GBDT，XGBoost。



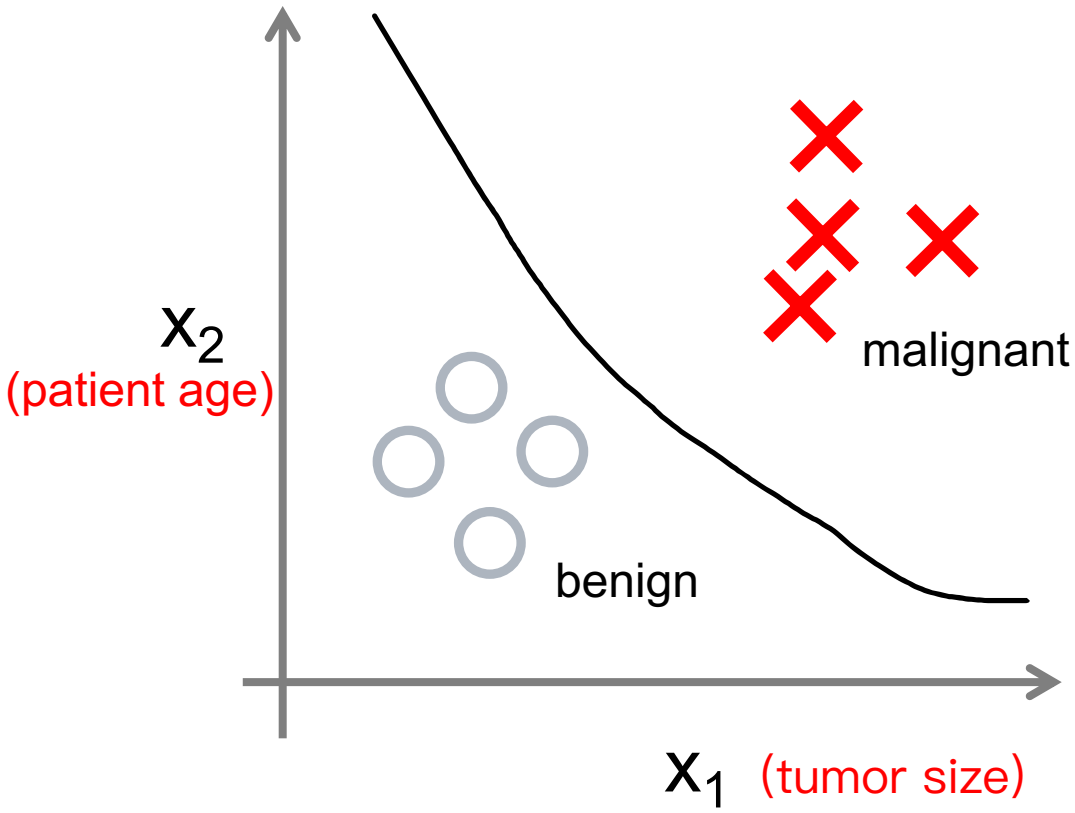


章节目录

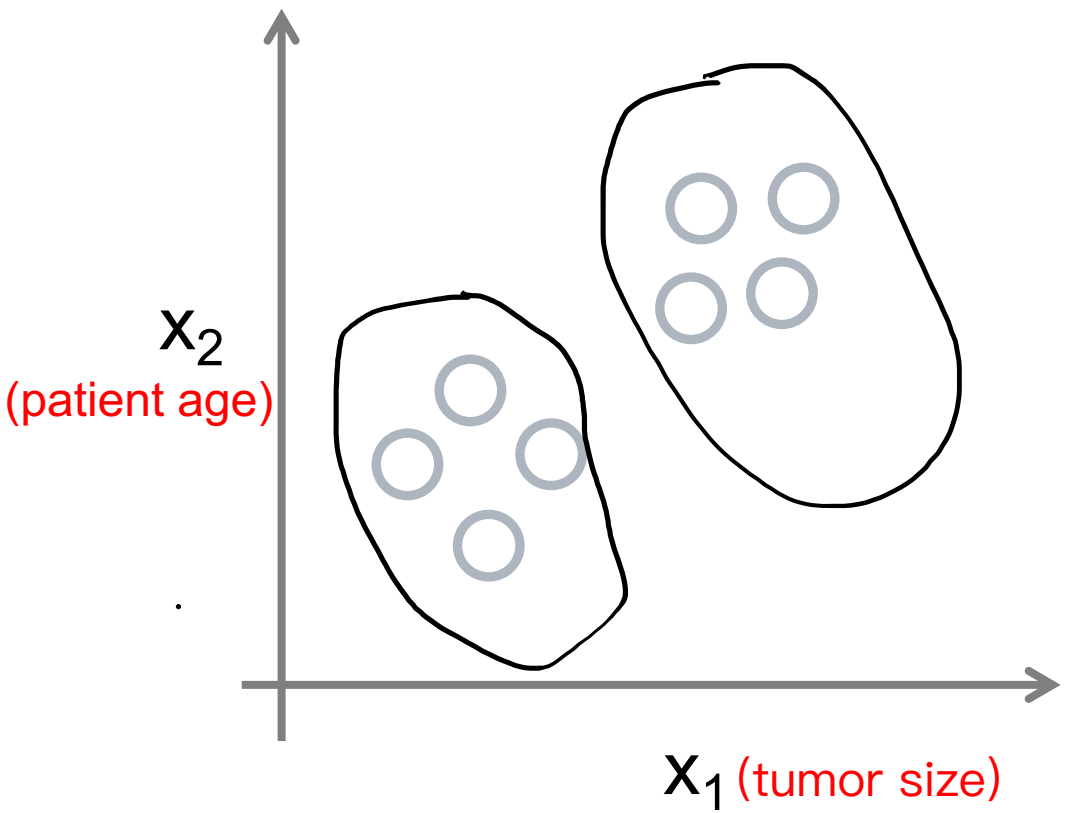
- 机器学习进阶3
  - 无监督学习
  - 什么是聚类
  - 距离的选择
  - K means聚类

无监督学习

Supervised Learning



Unsupervised Learning



无监督学习

- 在另一个场景中，我们只观察到特征向量，而没有测量输出。我们的任务是描述数据是如何组织或聚集的。这就是所谓的无监督学习问题。
- 无监督学习的基本想法是对给定数据（矩阵数据）进行某种“压缩”，从而找到数据的潜在结构。

注：假定损失最小的压缩得到的结果就是最本质的结构。

- 假设总共有 N 个实例，在无监督学习中，我们得到的训练集是 $U = \{x_1, \ldots, x_N\}$

其中，$x_i = (x_i^{(1)}, \ldots, x_i^{(n)})^\top$

样本	变量			
	x_1	x_2	$\cdots$	x_p
1	x_{11}	x_{12}	$\cdots$	x_{1p}
2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{np}

学生成绩表

	性别	语文	数学	英语	物理	化学	生物	历史	政治	地理
同学1	男	50	77	52	31	99	77	88	72	80
同学2	男	92	95	55	66	66	90	40	78	91
同学3	女	85	63	88	90	92	76	95	84	83
同学N	女	88	80	56	65	46	90	96	68	70

无监督学习

■ 无监督学习的基本想法是对给定数据（矩阵数据）进行某种“压缩”，从而找到数据的潜在结构。

注：假定损失最小的压缩得到的结果就是最本质的结构。

■ 考虑发掘数据的纵向结构：

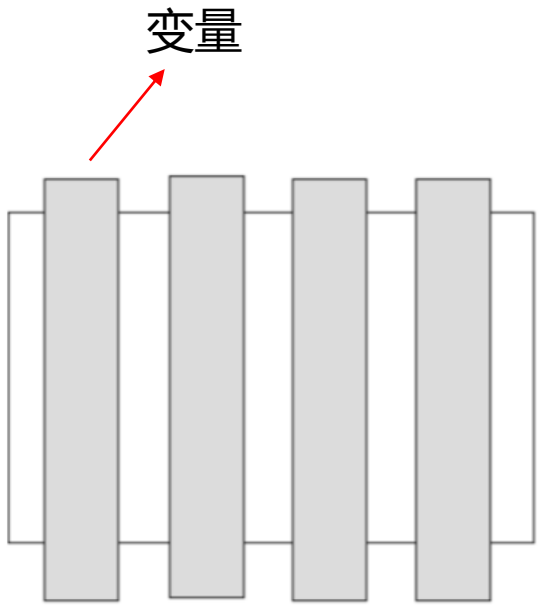
➢ 把高维空间的向量转换为低维空间的向量，

➢ 即对数据进行降维

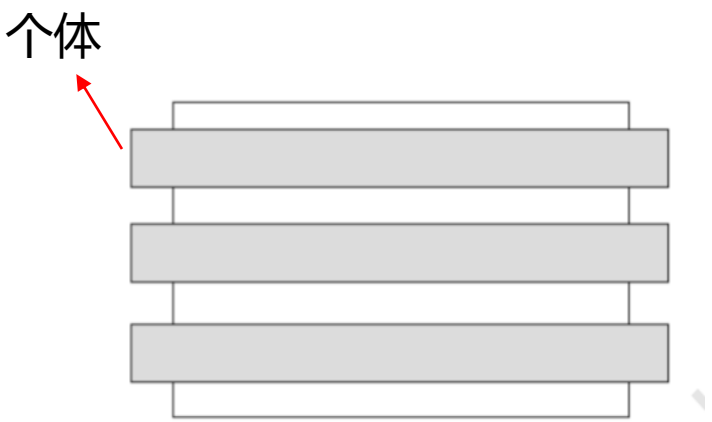
■ 考虑发掘数据的横向结构：

➢ 把相似的样本聚到同类，

➢ 即对数据进行聚类



(a) 数据纵向结构

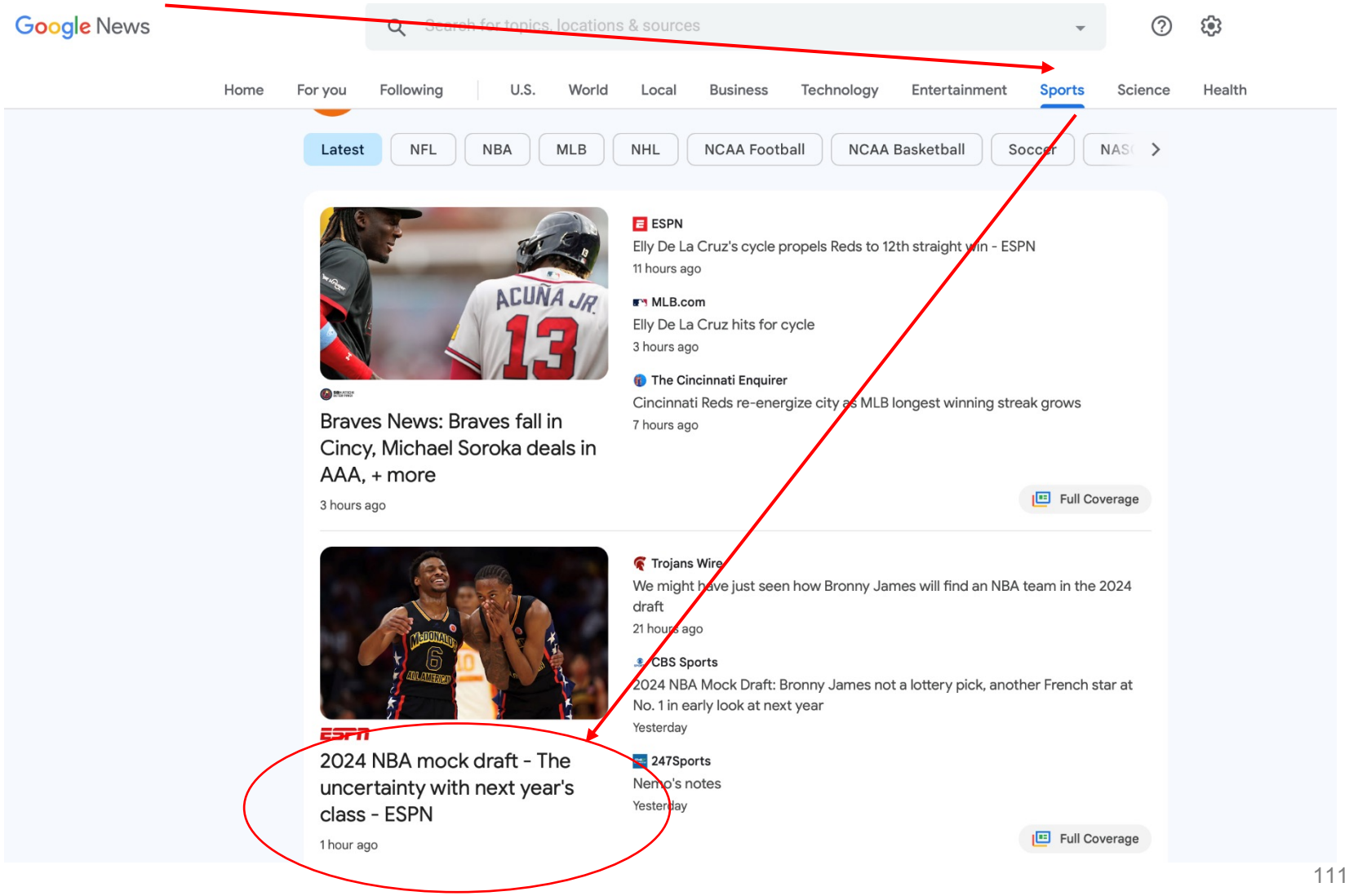


(b) 数据横向结构

无监督学习：聚类

谷歌新闻

- 谷歌新闻每天会报道大量的新闻
- 通过聚类算法，将这些新闻聚集在一起形成一个个专题
- 谷歌新闻要做的就是搜索成千上万的新闻，将这些新闻进行聚类

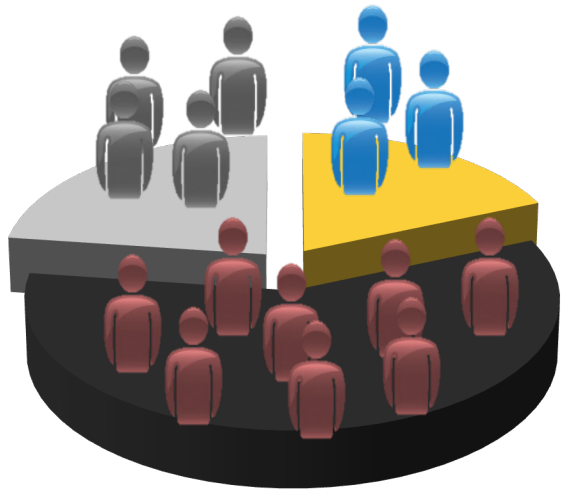


无监督学习：聚类

■ 市场分割

- 目标: 将市场细分为不同的客户子集，任何子集都可以被选择为市场目标。
- 搜集客户的基本信息（性别，年龄，受教育程度），购买信息（购买金额，购买数量，购买频率），产品信息等
- 寻找相似的客户群（潜力型客户、价值型客户、维持型客户等）
- 针对相应的客户群体给予不同的促销策略

- **潜力型客户**：这些客户可能暂时消费较少，但他们有很大的潜力成为高价值客户。企业可以通过优惠、会员奖励等手段吸引这些客户更多地进行消费。
- **价值型客户**：这些客户对企业贡献最大，通常是忠诚度高、消费频繁的用户群体。企业应注重维护这些客户，确保他们不会流失，并不断提升他们的用户体验。
- **维护型客户**：这些客户曾经有较高的消费，但最近消费减少或停滞。企业可以通过特别的促销活动或个性化服务，重新激发这些客户的消费兴趣。



例：市场分割

■ 文档聚类: 根据出现的重要术语查找彼此相似的文档组

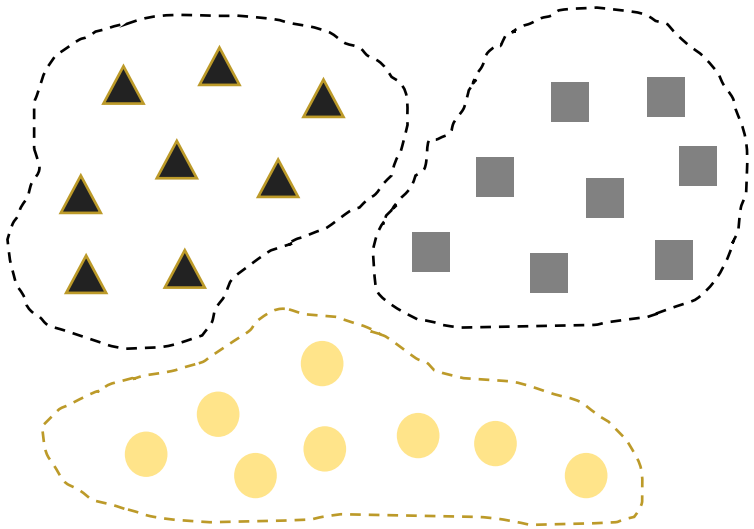
■ 根据每天的走势对股票进行分类

■ 天体聚类

例：天体分析

聚类

- 聚类目的
  - 希望“物以类聚”，即同一簇的样本尽可能彼此相似，不同簇的样本尽可能不同。
  - 换言之，聚类结果的“簇内相似度”（intra-cluster similarity）高，且“簇间相似度”（inter-cluster similarity）低，这样的聚类效果较好
- 聚类的作用
  - 聚类既可以作为一个单独过程（用于找寻数据内在的分布结构）
  - 也可作为分类等其他学习任务的前驱过程.



相似度的刻画-距离

- 第i个样品，第i行，$(x_{i1}, \dots, x_{ip})$
- 第j个样品，第j行，$(x_{j1}, \dots, x_{jp})$
- 定义第i, j个样品之间的距离：d_{ij}
  - $d_{ij} \geq 0$,任意i, j.
  - $d_{ij} = 0$,当且仅当两个样品的各变量值都相同
  - d_{ij}=d_{ji},任意i, j.
  - $d_{ij} \leq d_{ik}$+d_{kj},任意i, j, k.

样本	变量			
	x_1	x_2	$\cdots$	x_p
1	x_{11}	x_{12}	$\cdots$	x_{1p}
2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{np}

样品之间的距离

定量数据的常用距离

欧氏距离

$$d_{24} = \sqrt{\sum_{k=1}^{3} |x_{ik} - x_{jk}|^2}$$

$$= \sqrt{(18-4)^2 + (10-5)^2 + (36.3-11.5)^2} = 28.91$$

切比雪夫距离

$$d_{24} = \max_{1 \le k \le 3} |x_{ik} - x_{jk}| = \max\{(18-4), (10-5), 36.3 - 11.5)\} = 36.3 - 11.5 = 24.8$$

	V1	V2	V3
1	20	7	25.2
2	18	10	36.3
3	10	5	28.9
4	4	5	11.5
5	4	3	17

思考：Minkowski距离有什么缺点?

定量数据的常用距离

- Minkowski距离

$$d_{ij} = \left\{ \sum_{k=1}^{p} |x_{ik} - x_{jk}|^q \right\}^{1/q}$$

- Manhattan距离，绝对值距离

$$d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

- Euclide距离，欧式距离

$$d_{ij} = \sqrt{\sum_{k=1}^{p} |x_{ik} - x_{jk}|^2}$$

- Chebyshev距离

$$d_{ij} = \max_{1 \le k \le p} |x_{ik} - x_{jk}|$$

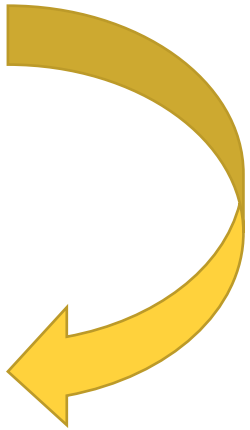
样本	变量			
	x_1	x_2	$\cdots$	x_p
1	x_{11}	x_{12}	$\cdots$	x_{1p}
2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{np}

✓ Minkowski距离中，$q = 1$即为绝对值距离； $q = 2$即为欧式距离；$q = \infty$为Chebyshev距离

离散数据的距离

	性别	专业	职业	学历	籍贯
甲	男	统计	咨询师	硕士	浙江
乙	女	法律	公务员	硕士	上海

变量	变量	甲	乙
X1	性别_男	1	0
X2	性别_女	0	1
X3	专业_统计	1	0
X4	专业_法律	0	1
X5	职业_咨询师	1	0
X6	职业_公务员	0	1
X7	学历_硕士	1	1
X8	籍贯_浙江	1	0
X9	籍贯_上海	0	1



1,对每个离散变量进行0-1 one-hot编码

2,对构造的变量按照欧氏距离计算数据点之间的距离

离散与连续数据的距离

个人	身高	体重	眼球颜色	头发颜色	优势手	性别
1	68	140	绿	金	右	女
2	73	185	棕	黑	右	男
3	67	165	蓝	金	右	男
4	64	120	棕	黑	右	女
5	76	210	棕	黑	右	男

方法1：将连续变量离散化，按照离散变量的方法计算样本距离
方法2：将离散变量转化成连续变量，按照连续变量方法计算样本距离
方法3：构造可以包含离散和连续变量的新的距离计算公式

离散与连续数据的距离（1）

个人	身高	体重	眼球颜色	头发颜色	优势手	性别
1	68	140	绿	金	右	女
2	73	185	棕	黑	右	男
3	67	165	蓝	金	右	男
4	64	120	棕	黑	右	女
5	76	210	棕	黑	右	男

低身高= [64, 67]； 中身高=[68, 73]； 高身高=74以上；　　轻体重： < 150：中：150–180：重：大于180

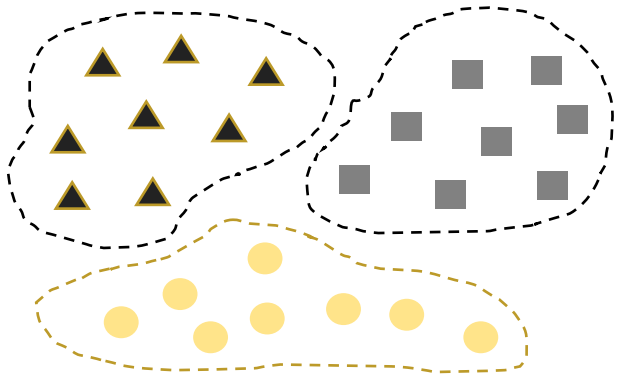
个人	身高段	体重段	眼球颜色	头发颜色	优势手	性别
1	中	轻	绿	金	右	女
2	中	重	棕	黑	右	男
3	低	中	蓝	金	右	男
4	低	轻	棕	黑	右	女
5	高	重	棕	黑	右	男

⇨ 离散变量

k均值法 kmeans

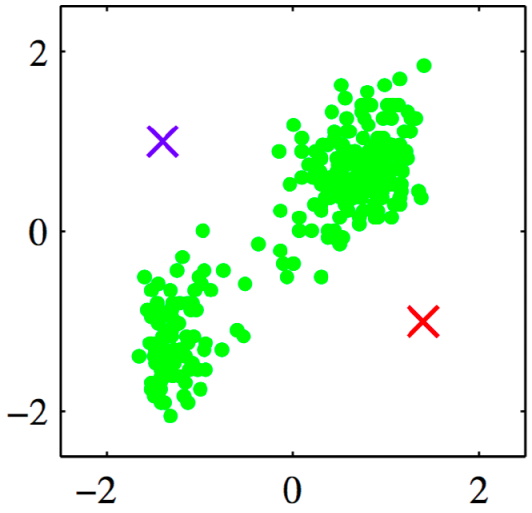
主要思想：

- 选择一批凝聚点（类中心）或给出一个初始的分类，
- 让样品按某种原则，让样本向凝聚点凝聚，
- 对凝聚点进行不断的修改或迭代，直至分类比较合理或迭代稳定为止，
- 选择初始凝聚点（或给出初始分类）的一种简单方法是采用随机抽选（或随机分割）样品的方法。

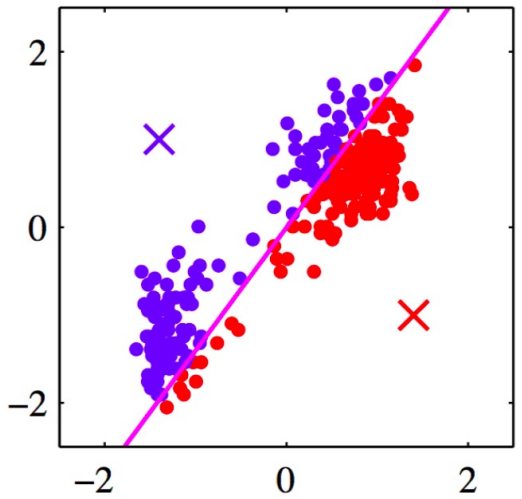


k均值法 kmeans

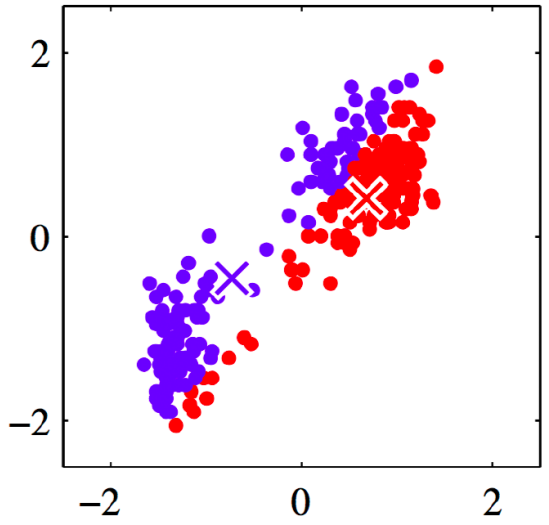
初始化类的中心



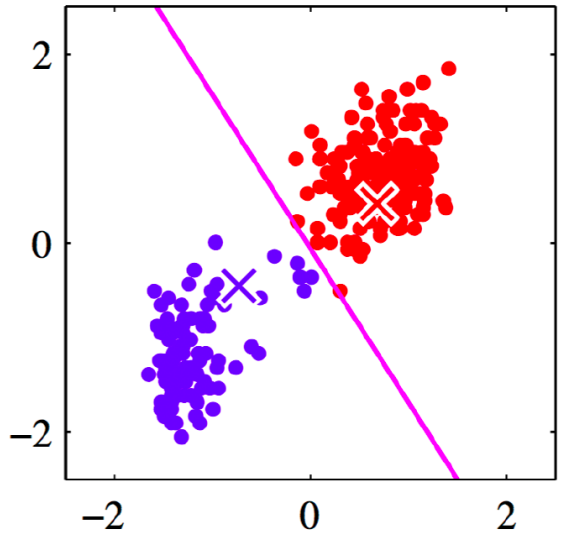
将每个样本点分配到离它最近的类中心



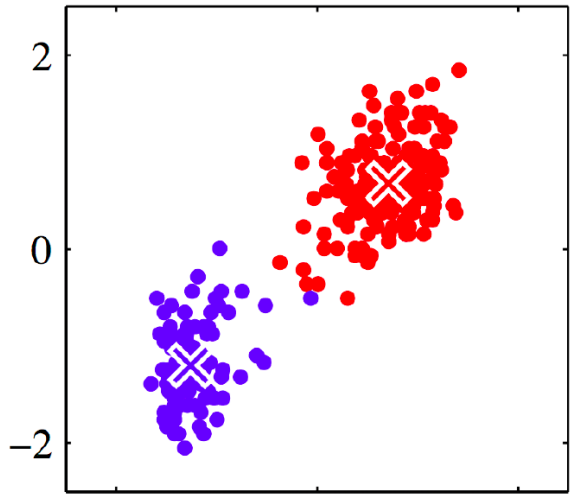
计算新的类中心



将每个样本点分配到离它最近的类中心



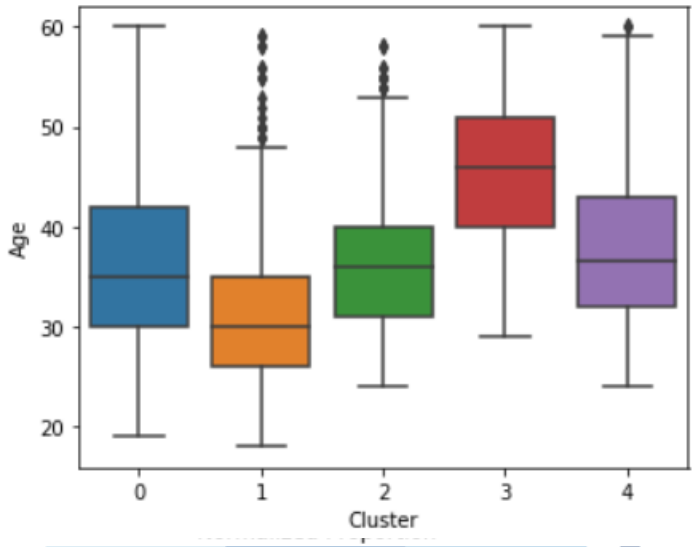
不断迭代直至收敛



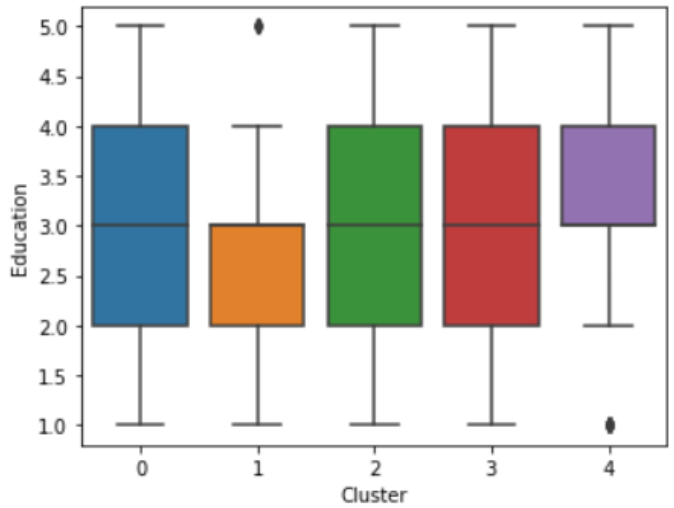
Kmeans聚类：实际案例

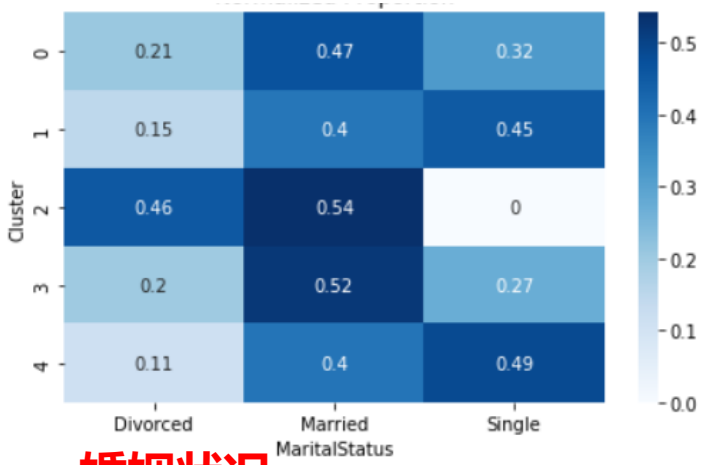
- 聚类后不同簇之间在特征上的差异——个人基本情况

年龄

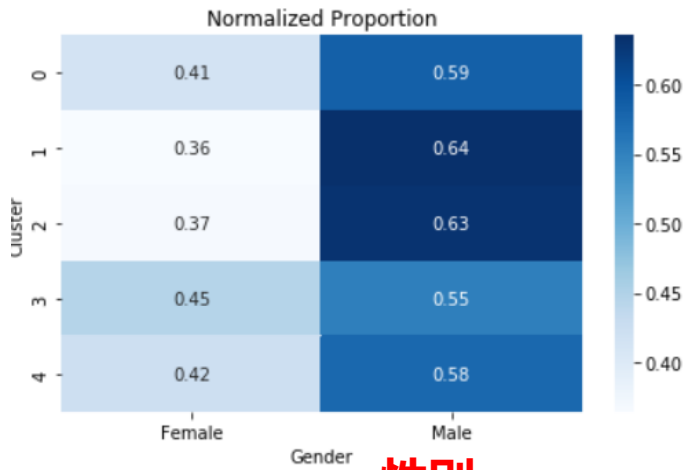


教育程度





婚姻状况



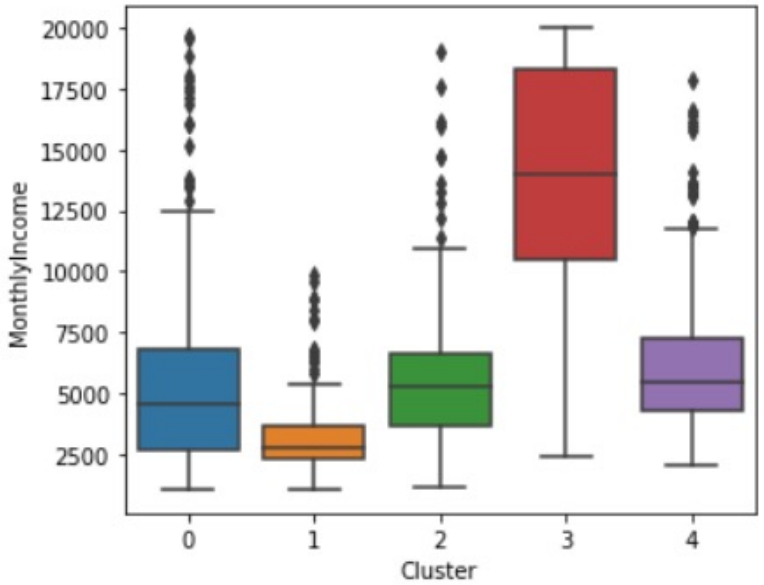
性别

- cluster3平均年龄最大、cluster1平均年龄最小；
- cluster1的受教育程度最低、cluster4受教育程度最高；
- cluster2均为已婚/离婚；
- cluster1、2男性更多，cluster3女性相对更对。

Kmeans聚类

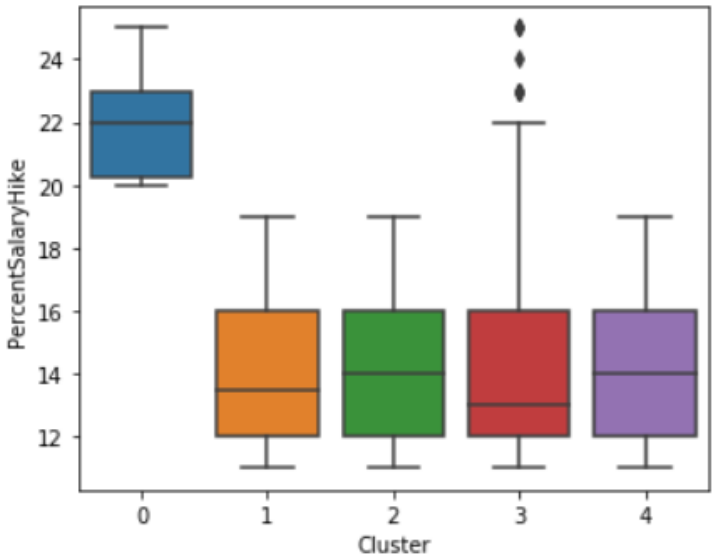
- 聚类后不同簇之间在特征上的差异——收入情况

月收入



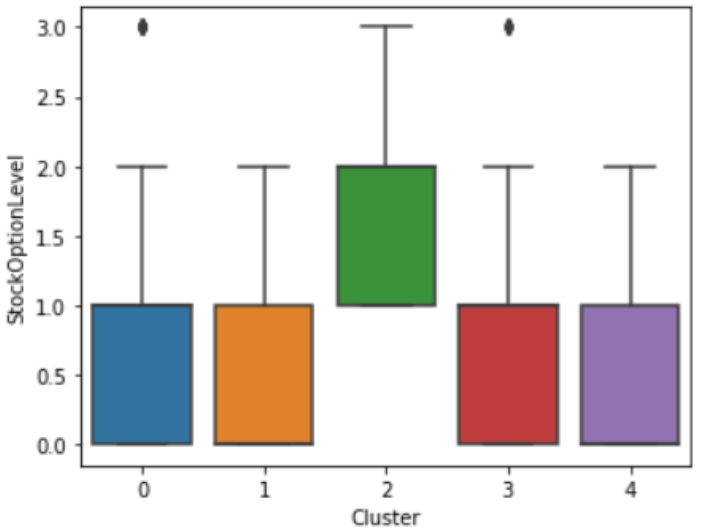
cluster3月收入最高，
cluster1月收入最低。

薪资涨幅



cluster0薪资涨幅比例最大，
cluster3涨幅比例最小。

期权等级

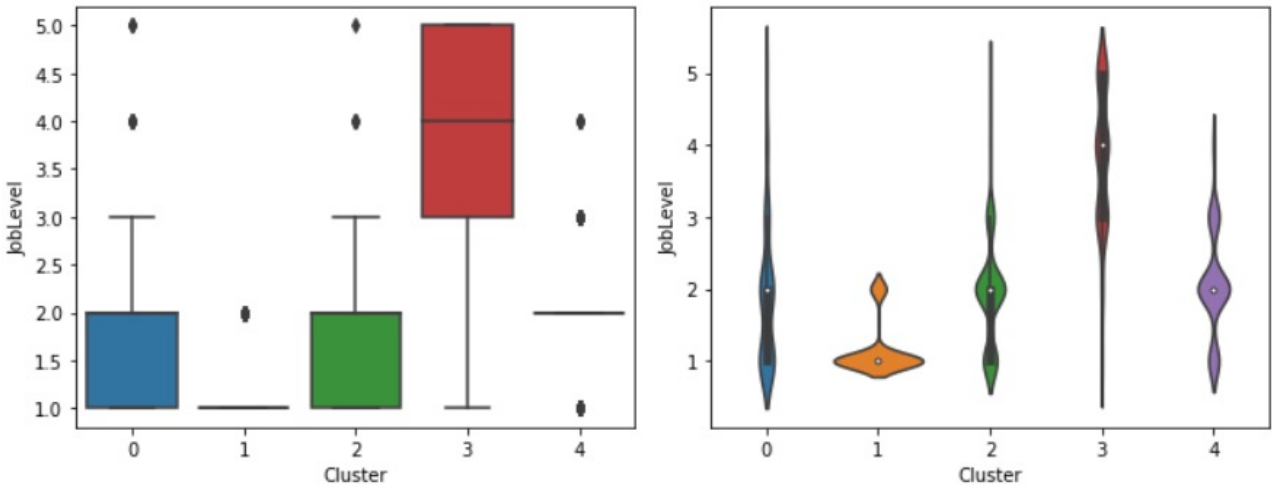


cluster2期权等级最高。

Kmeans聚类

- 聚类后不同簇之间在特征上的差异——工作部门

工作级别



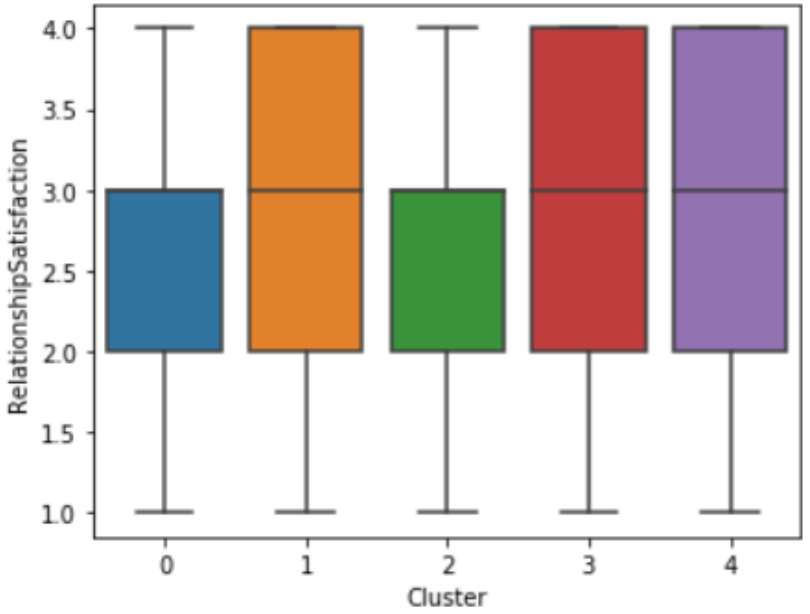
cluster1主要为研究科学家；
cluster3主要为管理经理；
cluster4主要为销售人员。

cluster3工作职级最高，cluster1工作职级较低。

Kmeans聚类

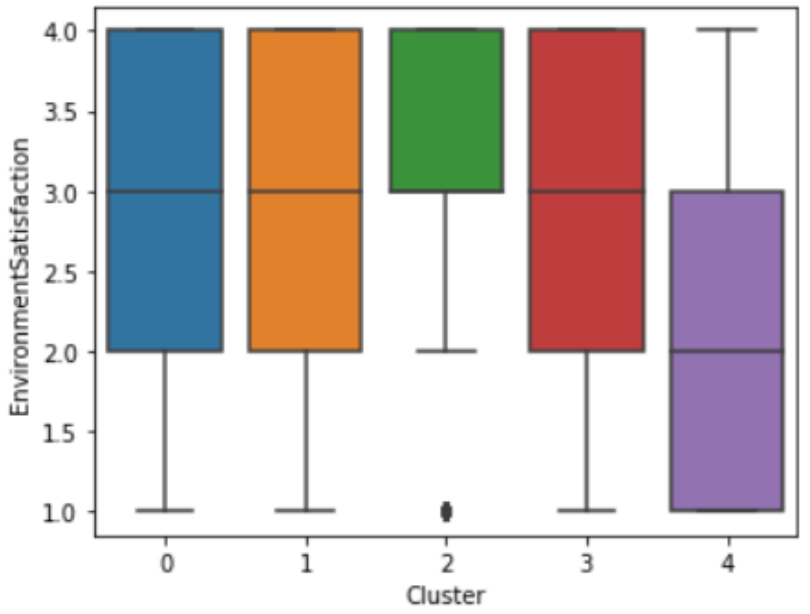
- 聚类后不同簇之间在特征上的差异——工作满意度

人际满意度



cluster1、3、4人际满意度相对较高；
cluster0、2人际满意度相对较低。

环境满意度

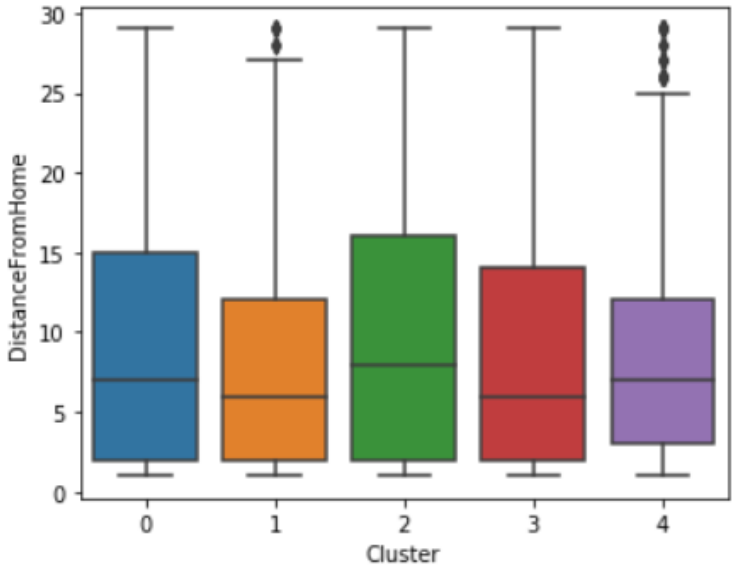


cluster2环境满意度较高；
cluster4环境满意度较低。

Kmeans聚类

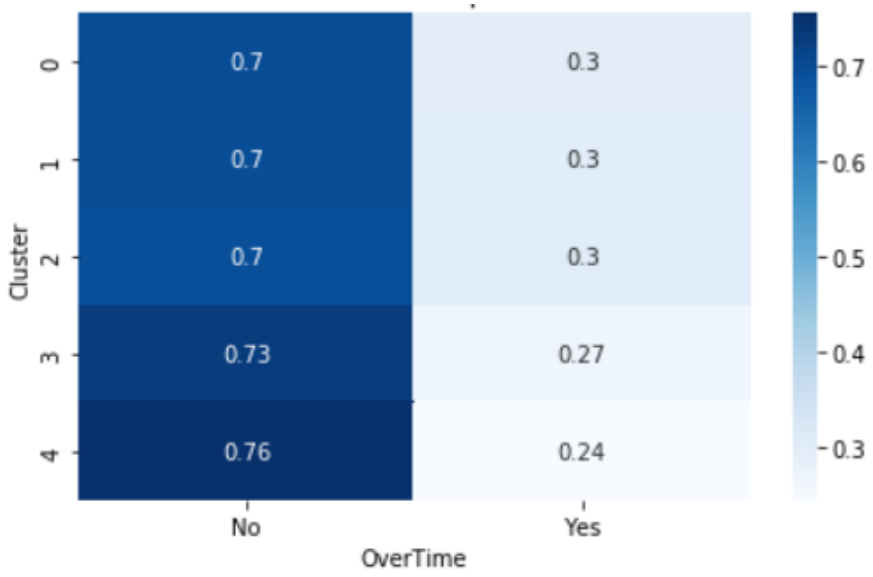
- 聚类后不同簇之间在特征上的差异——工作强度

家庭距离



cluster2家庭距离相对最远；
cluster1家庭距离相对最近。

加班频率

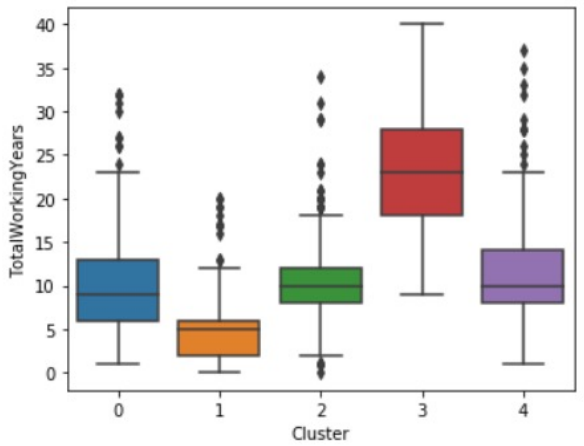


cluster3、4的加班率相对更低。

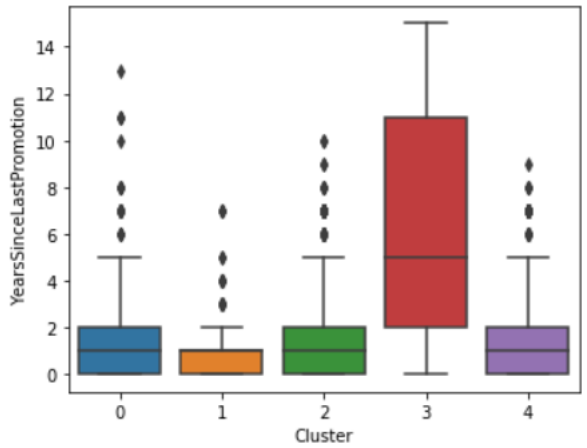
Kmeans聚类

- 聚类后不同簇之间在特征上的差异——工作经历

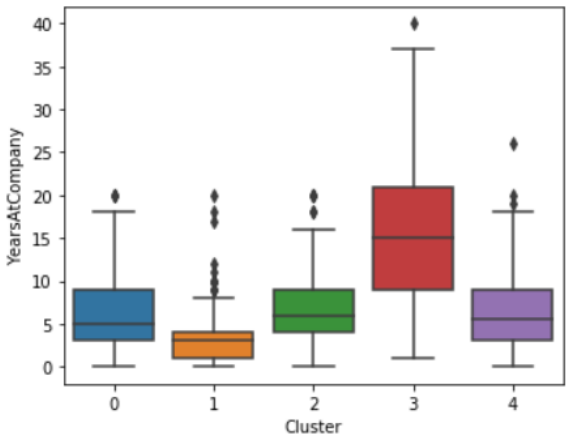
总工作年限



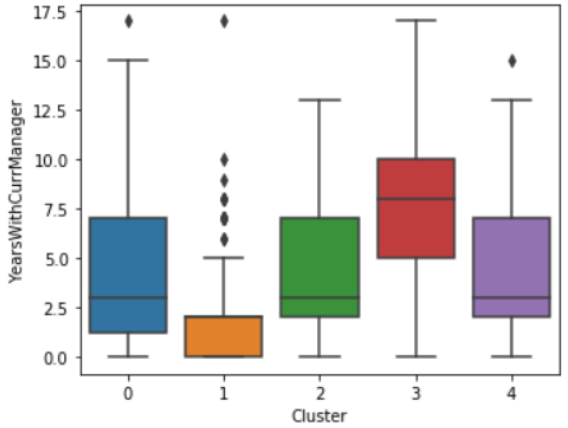
距离上次晋升时间



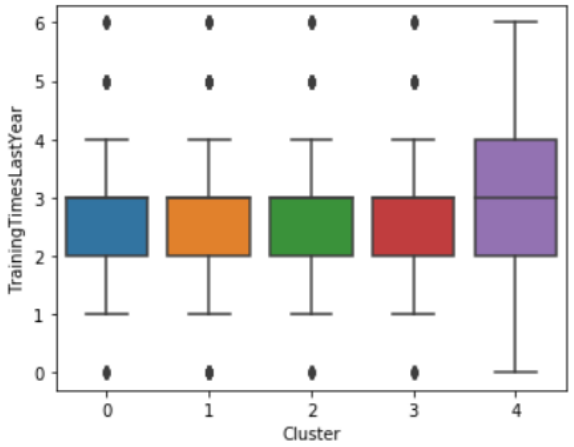
在公司的时间



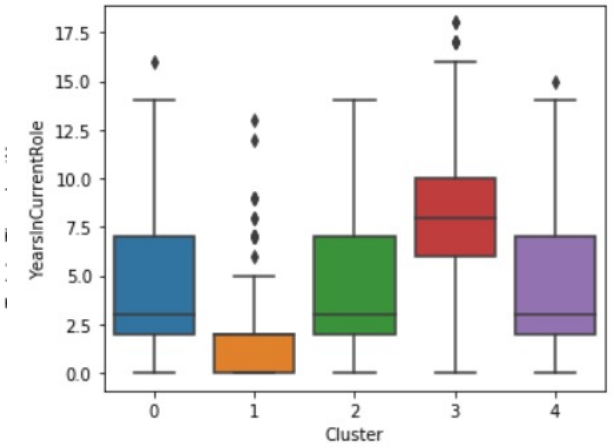
与当前经理的时间



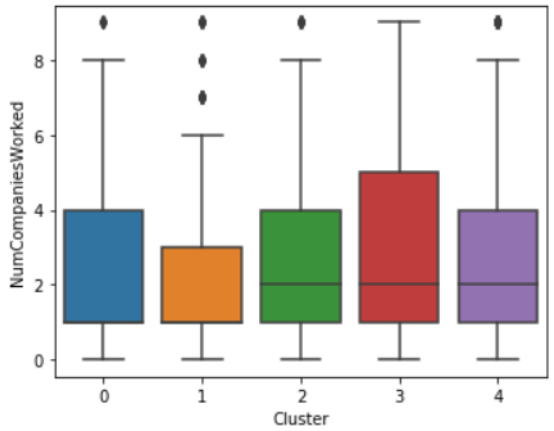
上一年度培训次数



从事当前职位时间



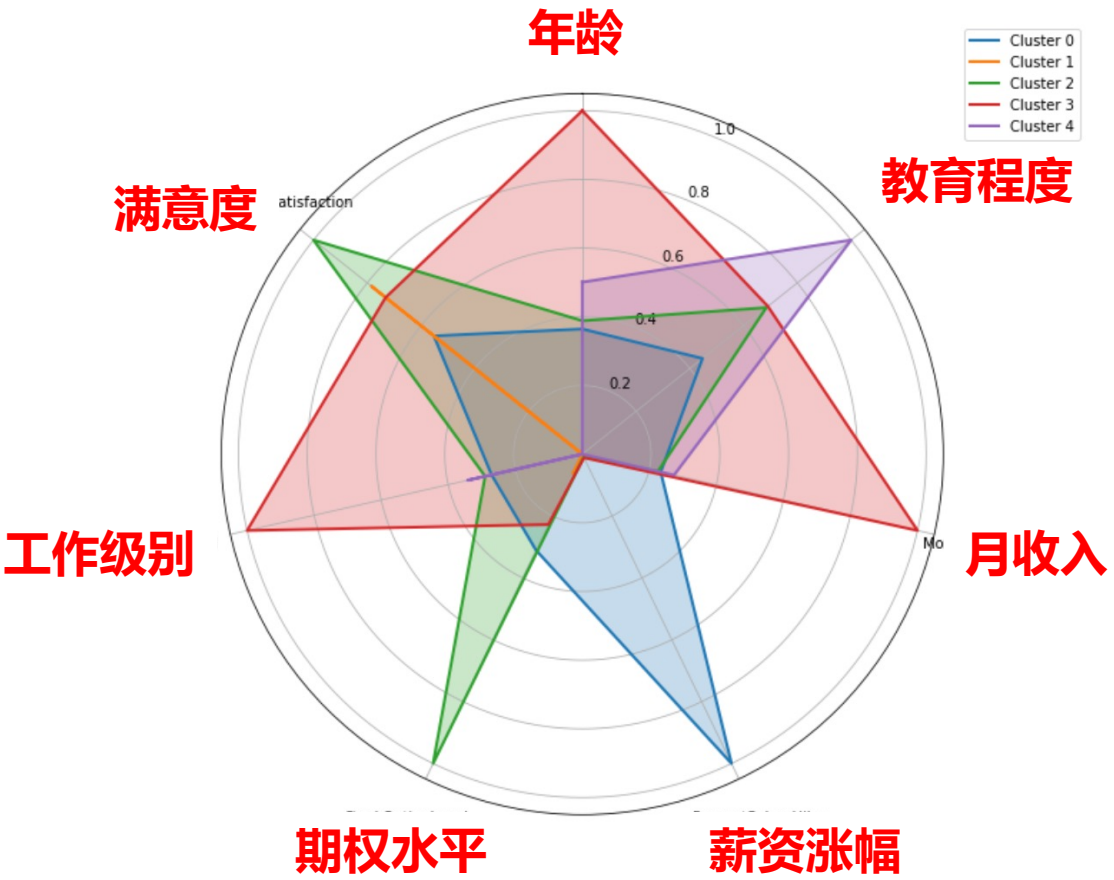
换工作的次数



cluster3总工作年限长、在公司时间长、工作环境与工作伙伴稳定；
cluster1总工作年限短、晋升速度快、在公司时间短；
cluster4培训次数多、工作经历相对较少。

Kmeans聚类：员工流失

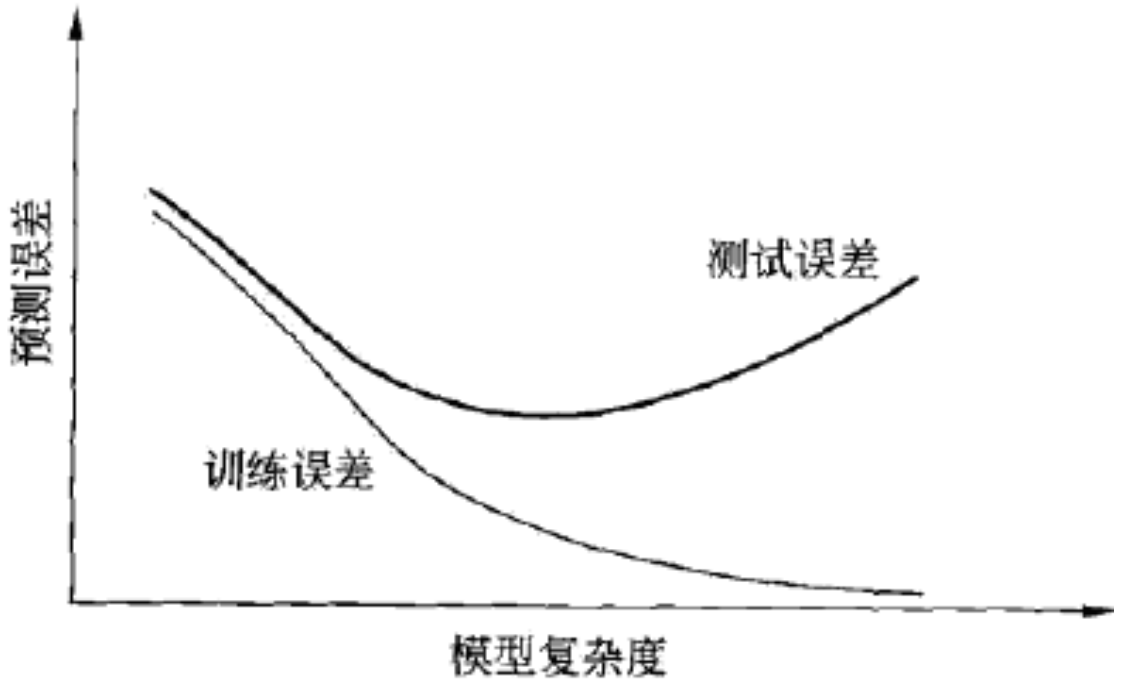
- 聚类后不同簇之间在特征上的差异



- cluster0——潜力员工：薪资涨幅比例最大；
- cluster1——初级员工：年龄小，受教育程度最低，月收入最低，工作职级最低，工作年限最短；
- cluster3——高级员工：年龄大，月收入最高，薪资涨幅比例最小，工作职级最高，工作年限最长。

机器学习指导思想：偏差与方差

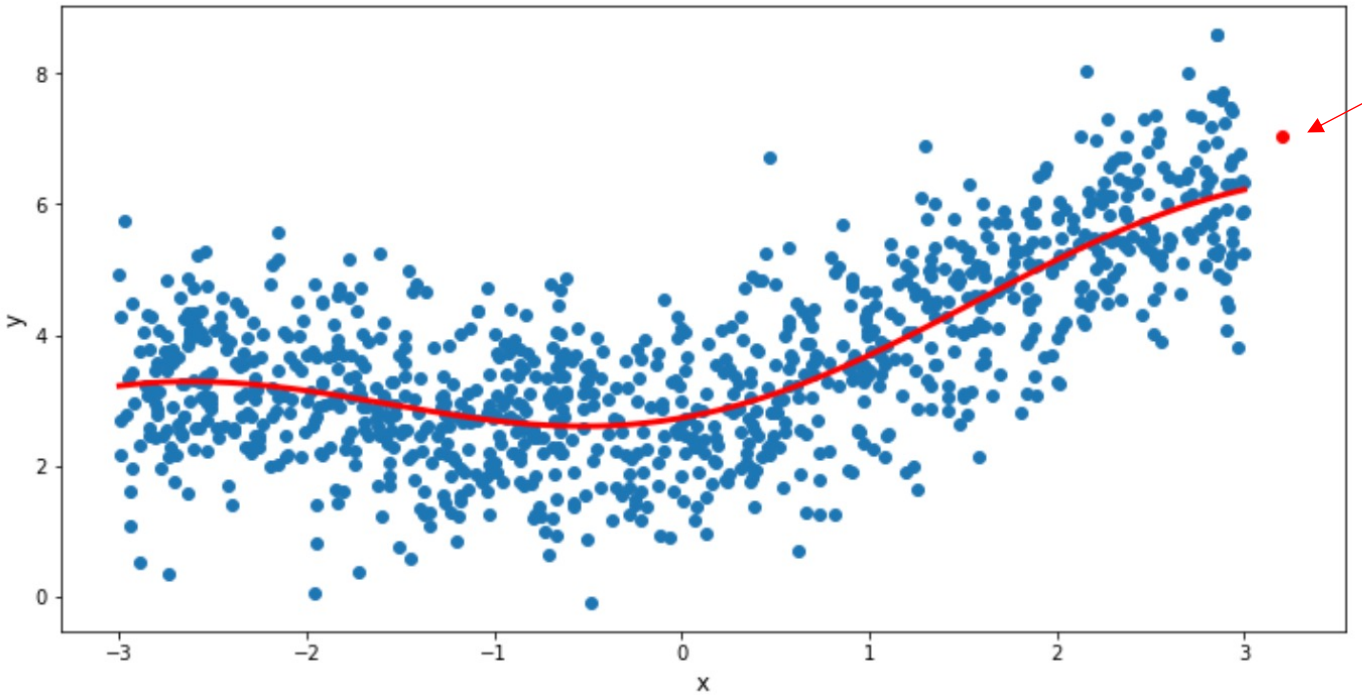
- 对学习算法除了通过实验估计其泛化性能以外，人们往往还需要了解它为什么具有这样的性能。
- **偏差与方差分解**（bias-variance decomposition）是解释学习算法泛化性能的重要工具。
- 通过对偏差与方差分解，可以指导如何进行模型调试以提高模型的性能。



机器学习指导思想：偏差与方差

- 假设 $y = f(x) + \epsilon$，其中 $\epsilon \sim \mathcal{N}(0, 1)$，以及

$$f(x) = \frac{1}{2}x + \sqrt{\max(x, 0)} - cosx + 2$$



待预测的点

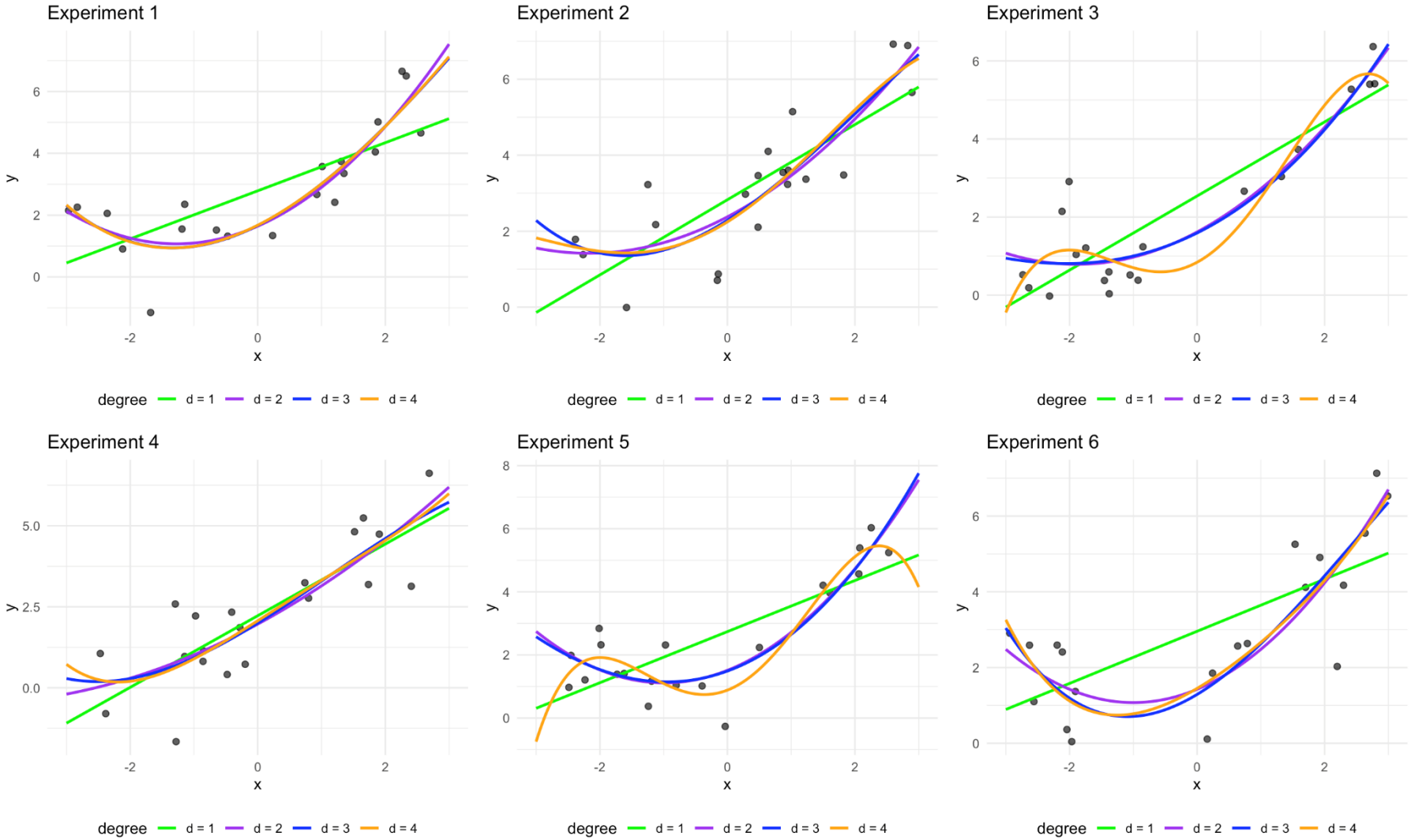
机器学习指导思想：偏差与方差

- 假如用d次多项式来拟合 x 与 y 之间的非线性关系：

$$\hat{f}(x) = w_0 + w_1 x + w_2 x^2 + \ldots + w_d x^d$$

- 假设我们每次只能使用20个样本点来训练多项式回归模型
- 考虑不同的模型：$d=1$, $d=2$, $d=3$ 以及 $d=5$.
- 重复6次试验（得到6次不同的训练样本（大小均为20），并进行模型学习）

机器学习指导思想：偏差与方差



机器学习指导思想：偏差与方差

- d=1对应的模型在不同的训练样本之下变化不大，而 d=4对应的模型则表现出非常大的波动性

——方差问题

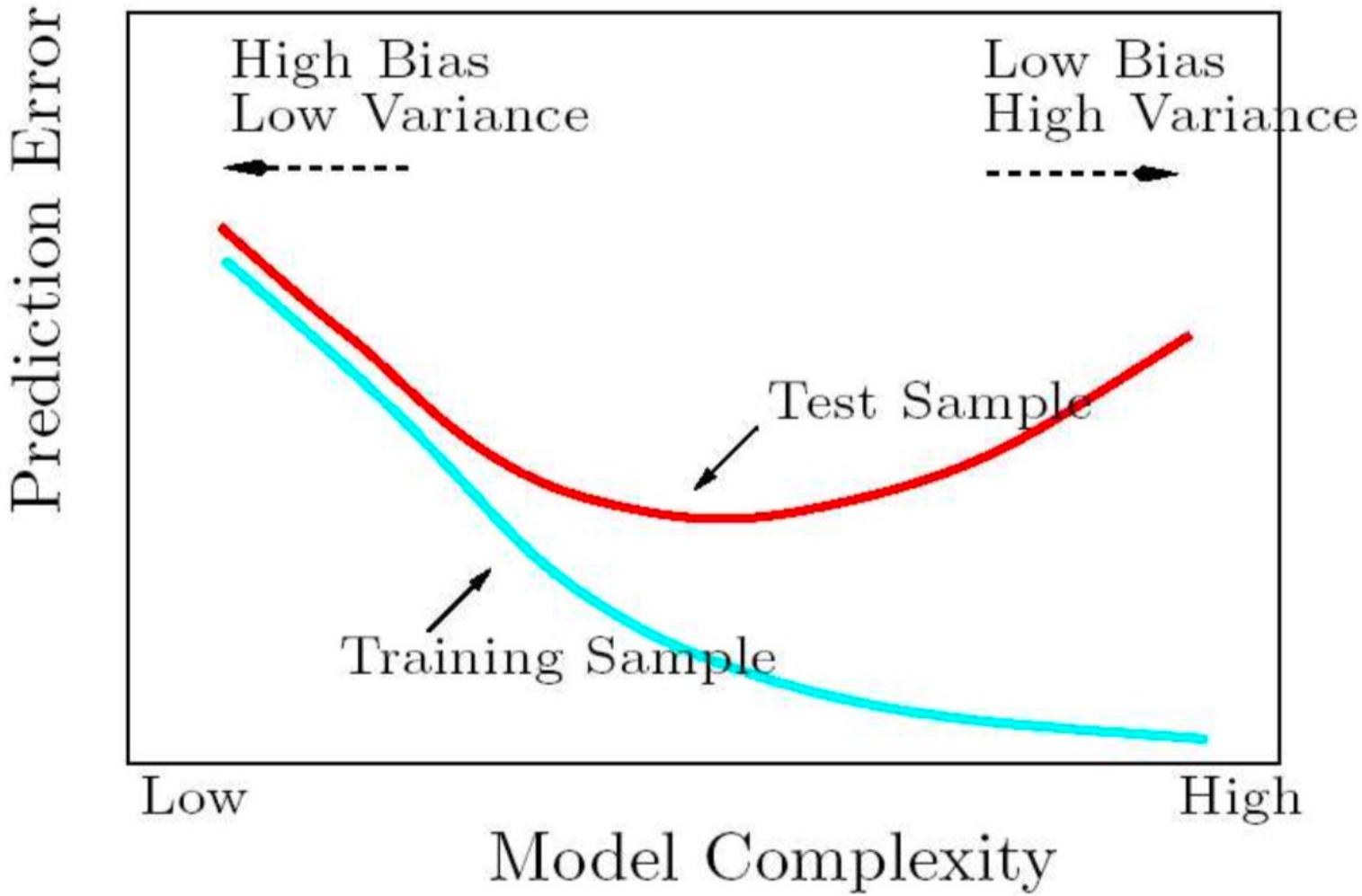
- d=1对应的模型 $\hat{f}(x)$ 距离真实 $f(x)$ 平均偏差更大，而 d=4对应的模型偏差相对较小

——偏差问题

- 为了进一步展示方差和偏差，假如有一个新的输入 $x_{\text{test}} = 3.2$,其真实标记为 $f(x_{\text{test}})$
- 重复10000次，得到10000个大小为20的训练样本，并分别计算 $\hat{f}(x_{\text{test}})$

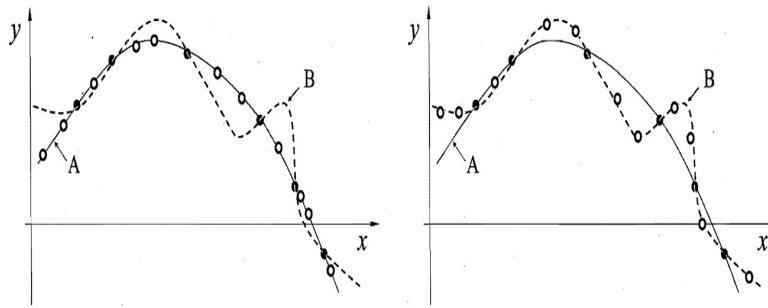
机器学习指导思想：偏差与方差分解

- 测试误差高于训练误差
- 对于High Bias情况，训练误差和测试误差都很大
- 对于High Variance 情况训练误差小，测试误差大



机器学习指导思想：没有免费的午餐

- 学习机器学习方法的目的是为了了解这些算法的原理和适用范围
- 在所有可能的问题中，没有任何一个算法能在所有任务上都优于其他算法。
  - 例如，如果算法A在情况下1表现优于算法B，那么必然存在情况2使得算法B优于算法A
  - 即理论上没有一致最优的算法！
- 脱离实际情况而空谈哪个算法好是没有意义的，即没有免费的午餐
  - No Free Lunch Theorem！
- 例如：考虑从A地至B地的算法
  - 情况1: A=A教学楼，B=B教学楼；最佳方案：走路
  - 情况2: A=A教学楼，B=浦东机场；最佳方案显然不是骑行车
- 例如图像分类问题：决策树or神经网络?



(a) A 优于 B　　(b) B 优于 A

思考：如何针对具体问题，分析比较不同的机器学习方法呢?

指导思想：奥卡姆剃刀原则

- 奥卡姆剃刀原则（Occam's Razor）原本是一种哲学原则，由14世纪英国哲学家和神学家威廉·奥卡姆（William of Ockham）提出。这个原则可以简单地概括为：

  “如无必要，勿增实体。”

换句话说，在解释事物或现象时，若有多种可能的解释，应该选择最简单的那个，即避免做出不必要的假设或增加不必要的复杂性。这一原则强调了简洁性和经济性，倡导在解释或理论中剔除那些不必要的、冗余的部分。

- 奥卡姆剃刀原则在机器学习中的核心思想是通过选择更简单的模型来提高模型的泛化能力，避免过拟合

- 模型选择
- 正则化
- 特征选择
- 模型复杂度和泛化能力