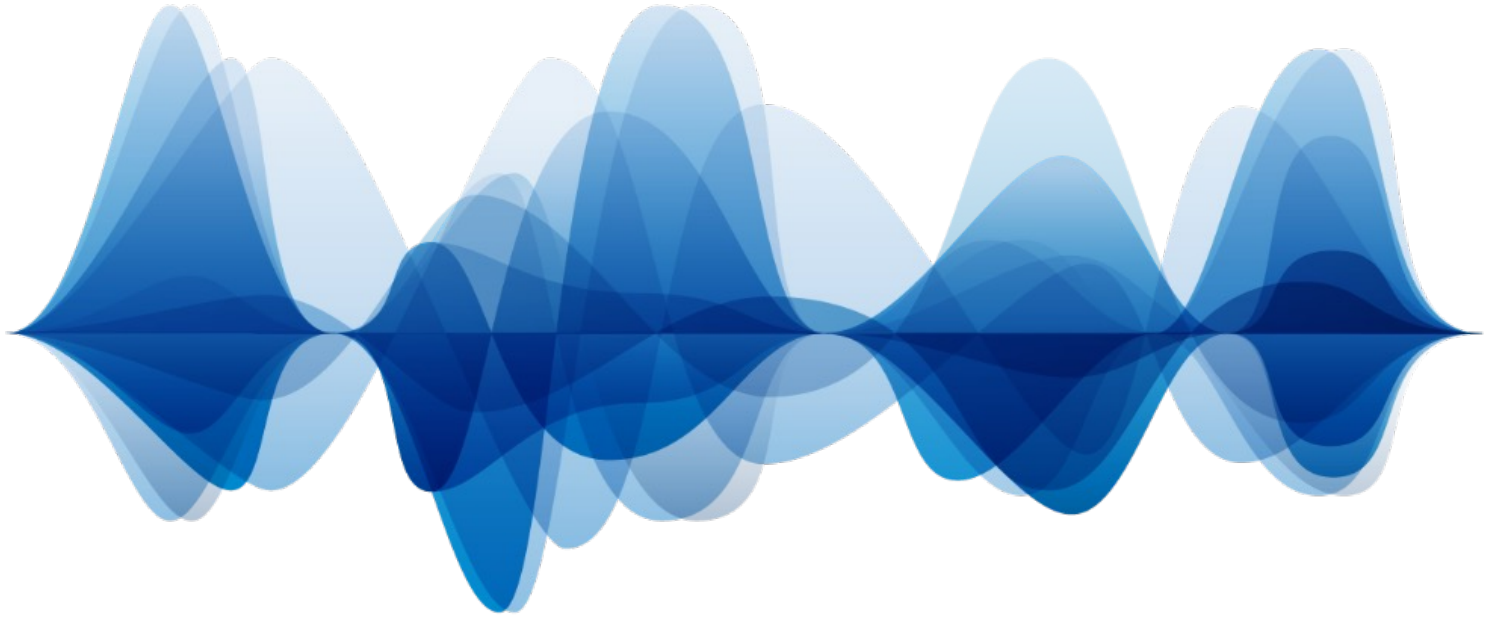
机器学习介绍

Xin Tao



推断

猜、预测、学习、智能、语言与统计：从过往数据和经验中学习规则，根据新条件作出判断 *Inference*

- **0，2，4，？，8，10，？** — *推断、预测*
- **秋收** — *日历？经验？统计？*
- **12中午点半了，在张三哪？** — *常识？经验推断？统计推断？*
- **这几天天天天气不好。** — *语言学习？条件概率推断？*
-  — *语言？数据？*
-  — *市场？数据？市场的语言？*

“每当我开除一个语言学家，语音识别系统就更准确了”

-自然语言（统计）处理先驱 Frederick Jelinek

归纳与演绎智能

归纳: 观察-总结 vs 演绎: 演绎-假设（证实或证伪）

Inductive Reasoning and Deductive Reasoning

人类很早就会造桥（前1523 年），但桥梁力学出现很晚（1749年）

经验归纳

- 借助数据，总结规律
- 记忆
- 描述-或然
- 会计、记事、文字
- 技术
- 心理学-行为学
- 统计、大数据、机器学习

李约瑟难题*

理论推演

- 借助原理，推理规律、数据
- 想象
- 解释-必然
- 逻辑演绎
- 科学
- 科学、数学理论

**为什么古代中国技术这么发达、却没有产生科学革命（从而错过了工业革命）？*

两种智能的数学形式

机器学习：借助计算机算力的现代统计推断 – 注重预测、不注重解释

$$f(x) = ax + b = y$$

参数、自变量\数据、模型形式

条件：f(x)参数未知，形式待确定，已知部分数据X，Y

目的：使用已知数据，求 f(x)的“最合适”的形式和参数，具体评价标准为：

1. f（x）是否适配现有数据X、Y，或者f（x）能否用新X*成功预测Y*？（预测性）
2. f(x)的参数和形式本身是否可以描述特定规律（解释性）

机器学习所谓“训练”或“学习”即是寻找这个参数的过程。

条件：f(X)形式、参数已知，已知X

目的：

1. 使用X，通过f(X)，求 Y值（计算）
2. 变换f(X)的数学形式（推导或证明）

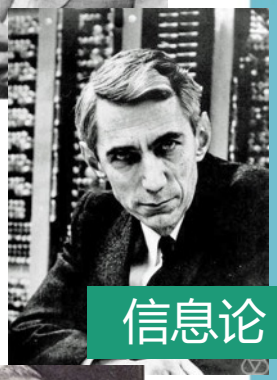
例：欧姆定律

$$V = IR \quad I = \frac{V}{R}$$

机器学习: 概念与历史

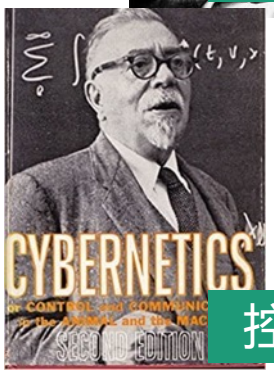
数理统计

1940's



信息论

火控研究



控制论

2014 AlphaGo

Artificial Intelligence
人工智慧

专家系统

Machine Learning
機器學習

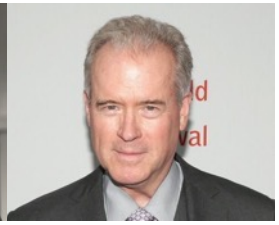
Deep Learning
深度學習

神经网络

1950's 1980's 2010's

Renaissance

1982



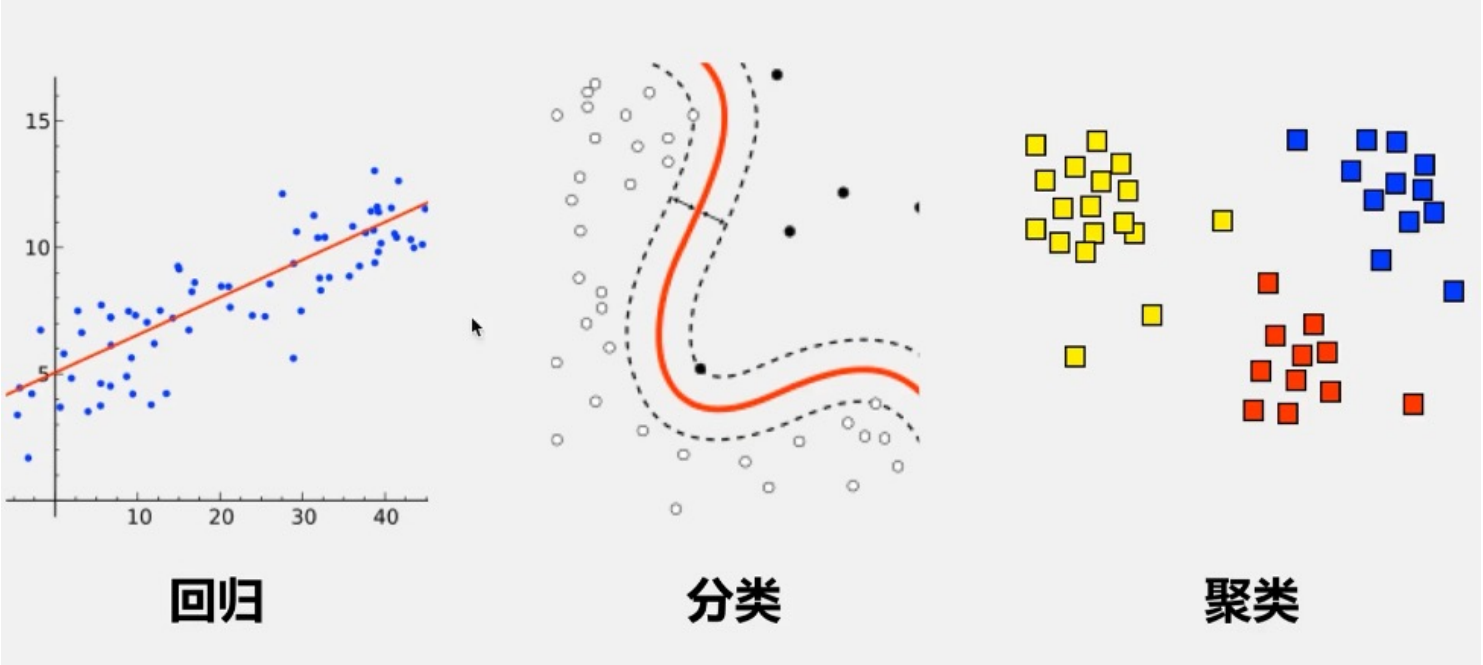
1993

2014 Cambridge Analytica

机器学习三板斧:　回归、分类、聚类

大部分模型都是为了描述关系、找出区分的条件、和总结大类

Regression, Classification, Clustering



卿，四大夫禄。君，十卿禄。次国之卿，三大夫禄，君，十卿禄。

-《礼记》

君子泰而不骄，小人骄而不泰。

-《论语》

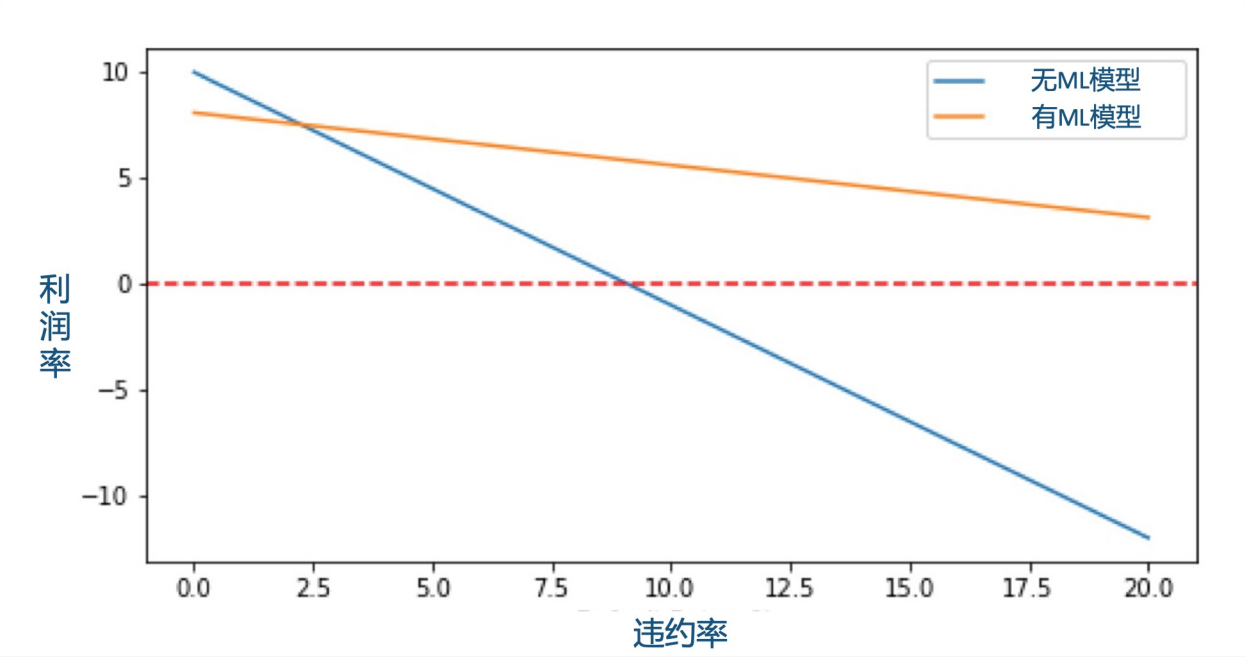
古者有四民：有士民，有商民，有农民，有工民。

-《春秋谷梁传》

机器学习与金融

机器学习与信用风险

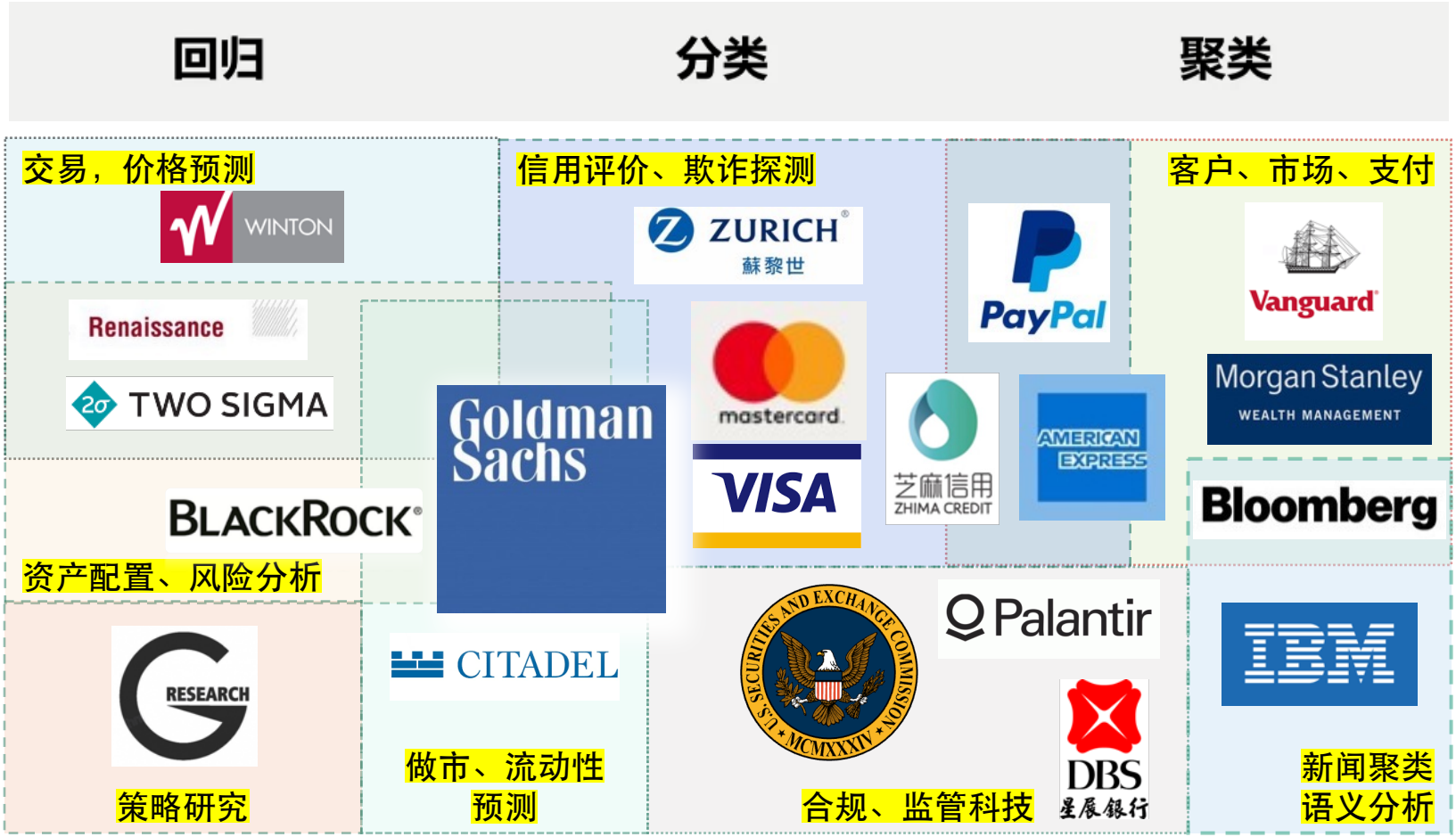
- 牛津大学Oxford-Man量化金融研究所关于使用机器学习对放贷者利润影响的研究：
  - 使用机器学习模型能显著减少违约率提高对利润的影响。
  - 显著提高放贷机构的危机生还率。



Dushimimana, Bernard; Wambui, Yvonne; Lubega, Timothy; McSharry, Patrick E. 2020. "Use of Machine Learning Techniques to Create a Credit Score Model for Airtime Loans" *J. Risk Financial Manag.* 13, no. 8: 180. https://doi.org/10.3390/jrfm13080180

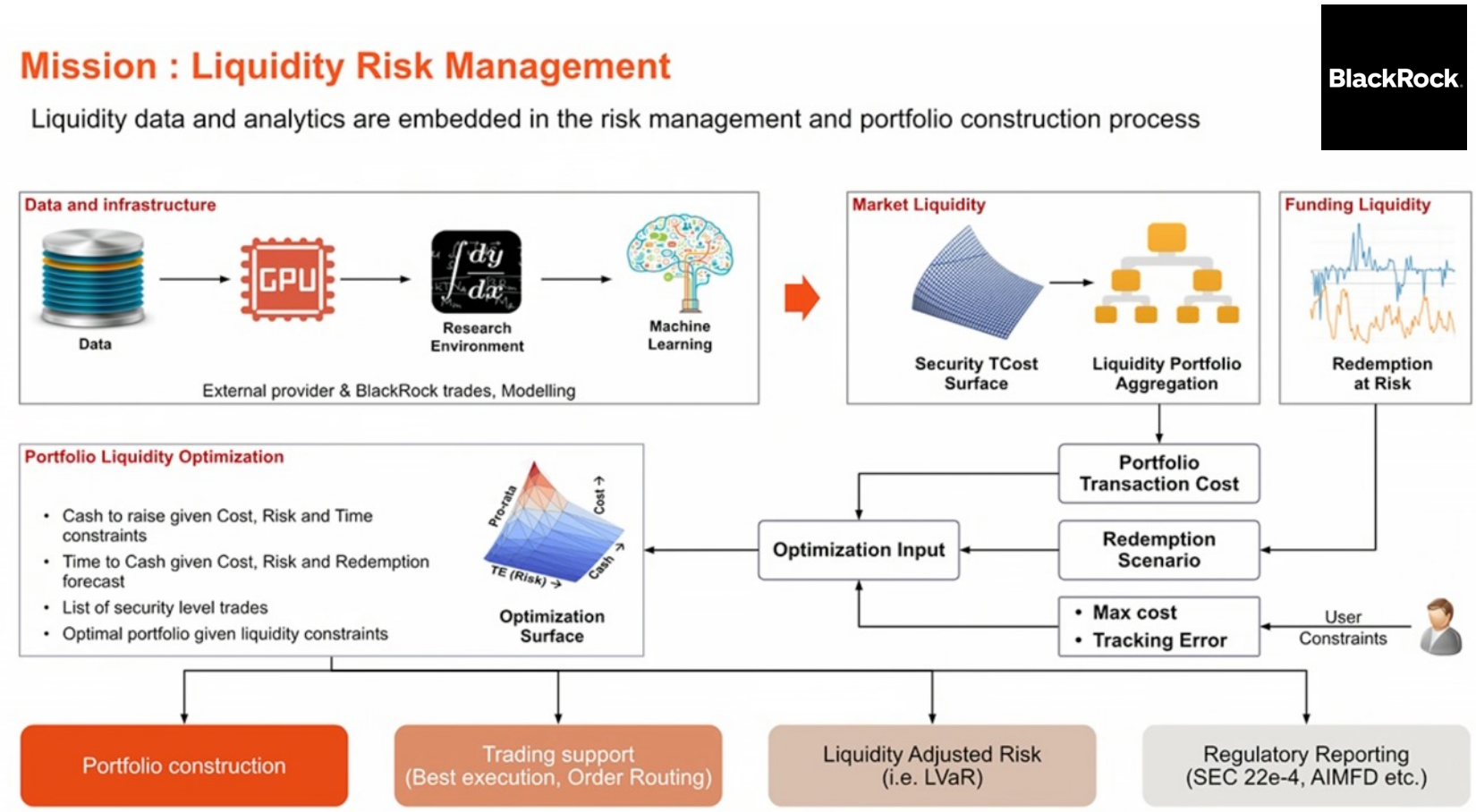
机器学习与金融

三大类功能对应的金融应用和典型产品、公司



机器学习与金融

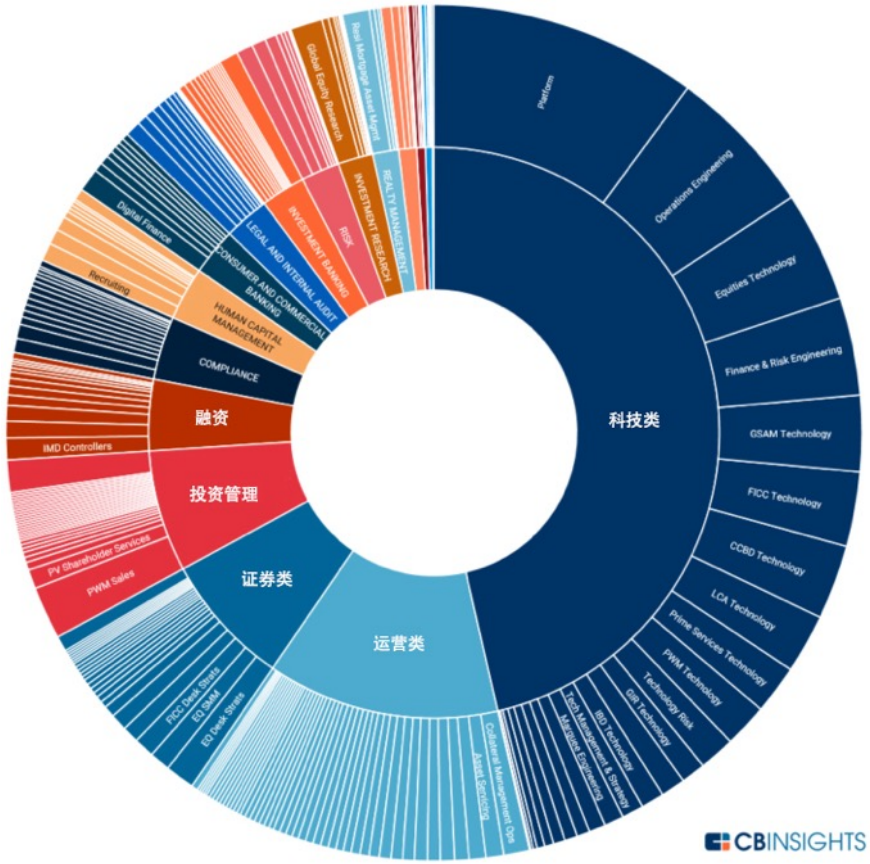
贝莱德（BlackRock）机器学习流动性管理框架



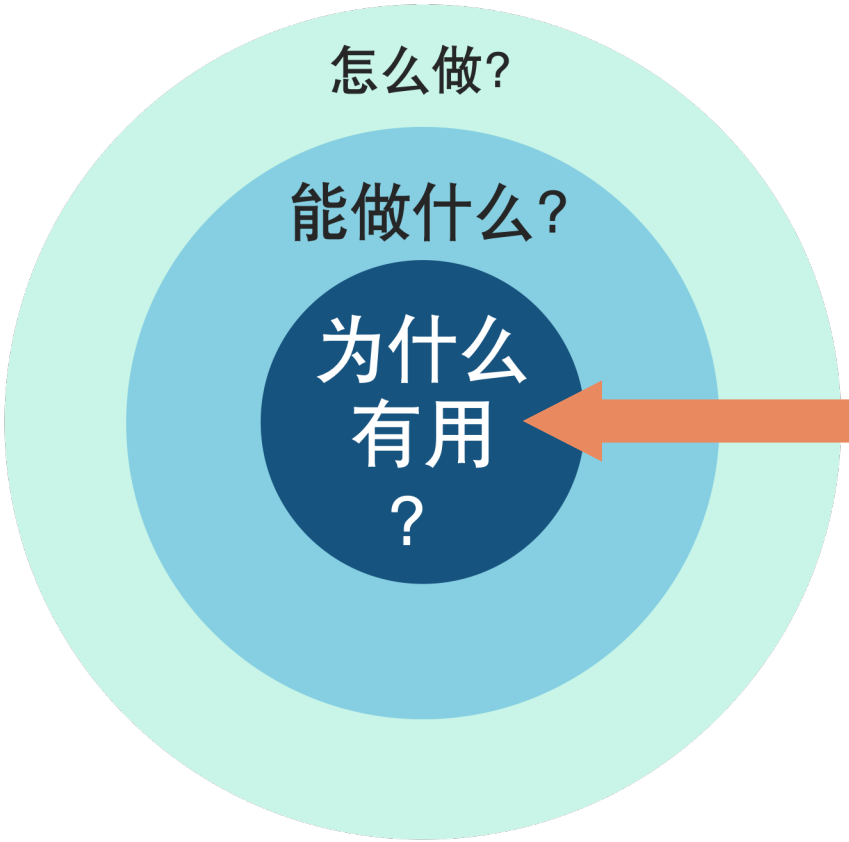
机器学习与金融

高盛2017年招聘最多的是信息科技类人才

- 高盛2017年招聘最多的是信息科技类人员。
- 并非由单独的科技部门招募，而是每个部门都有对应的信息科技招募。
    1. 很大程度还是优化流程和业务（为量化对冲基金服务需要很多自身技术改进，2020年开始建造统一平台）。
    2. 交易、合规、做市、内部数据分析平台、零售金融（Marcus）是主要的技术人员集中地。
    3. 仍有大量探索和创意性的工作。
    4. 资助机器学习领域研究和学术会议，紧跟技术前缘发展。
    5. 金融科技的主要风险投资者之一，以此取得前缘技术和人才。



机器学习



机器学习

15个核心原理和思想

思想（越简单越强大）

-> 数学模型（看上去很难）

-> 智能程序（需要会编程）

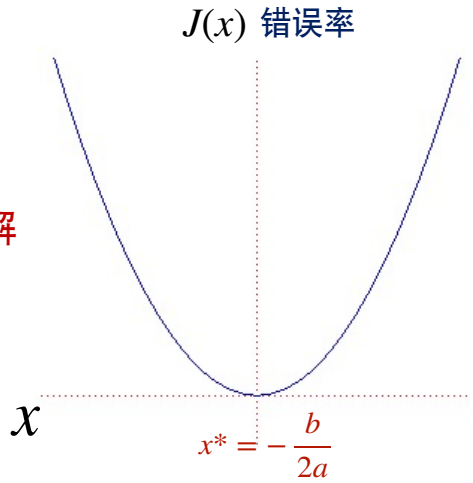
机器学习核心思想1: 迭代优化

先开始、再一步一步来。

Iterative Optimization

简单问题(一元二次方程极小值)

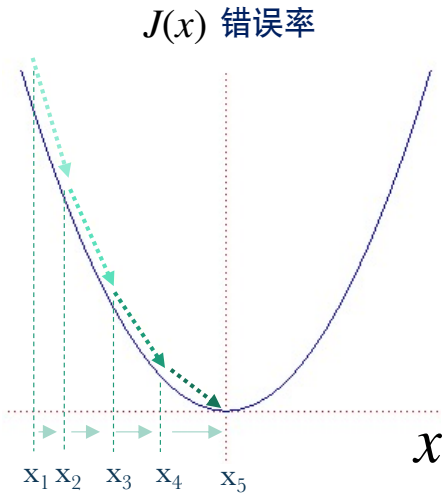
公式解



机器学习训练设计的核心问题:

1. 怎么设立一种机制让模型、参数一次比一次更准确，最快接近目标?

2. 怎么利用机器算力实现快速迭代?



机器学习:

梯度下降法

Gradient descent

利用斜率信息一步一步接近极值

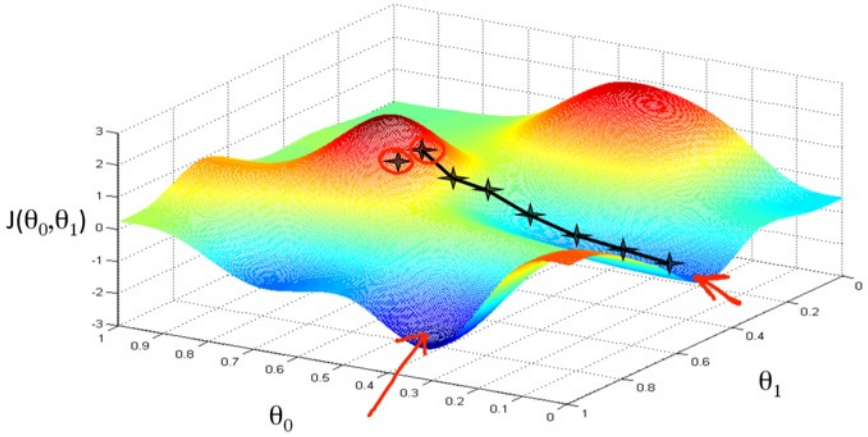
复杂问题

公式解:

可能不存在

机器学习:

梯度下降法逼近可能找到最优或次优解



机器学习核心思想2: 损失\奖励方程

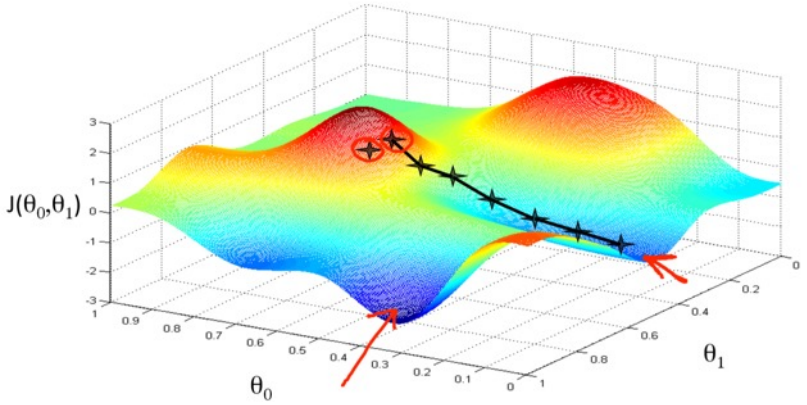
好坏的标准是模型训练最终成功的关键。

Loss \Reward Function

$J(\theta_1, \theta_2)$ 为损失\奖励方程

θ_1, θ_2 为模型参数

Parameters



- 每一步迭代训练的目标是：**找到让损失\奖励方程J更小\更大的参数组。**
- 但损失方程J是人为设定的，可以是错误率、概率、噪音比、波动率、含金量、胜率、回报、纯度等等，计算方式也可以自由选择。（*设计出好的损失方程可以让你成为公认机器学习大师*）
- **损失方程的好坏直接决定训练成败和模型好坏！**
- *思考：怎么设计一个股价预测程序的损失方程？（实际收益与预期收益的差？）*

“许愿需谨慎，万一实现了呢？”

--《伊索寓言》

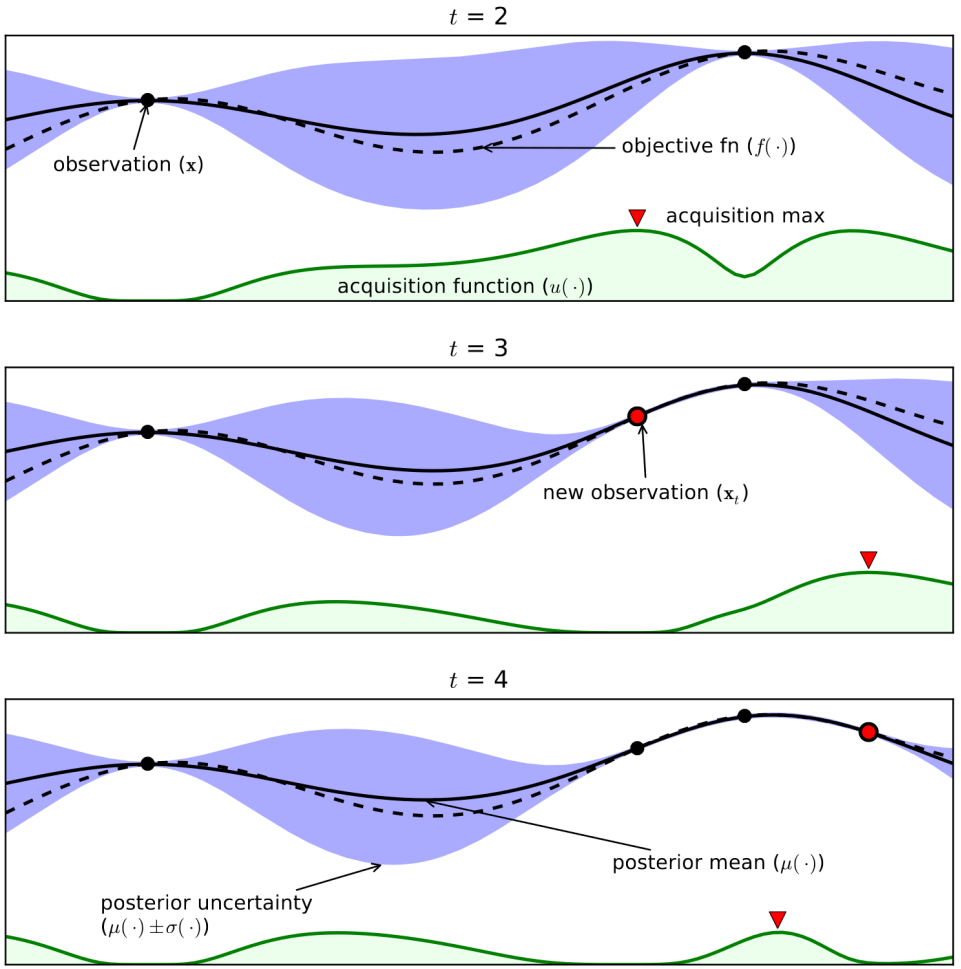
机器学习核心思想3: 贝叶斯优化

继续探索还是停止，这是一个问题；往哪个方向探索，又是一个问题。

Bayesian Optimization

探索vs利用难题：

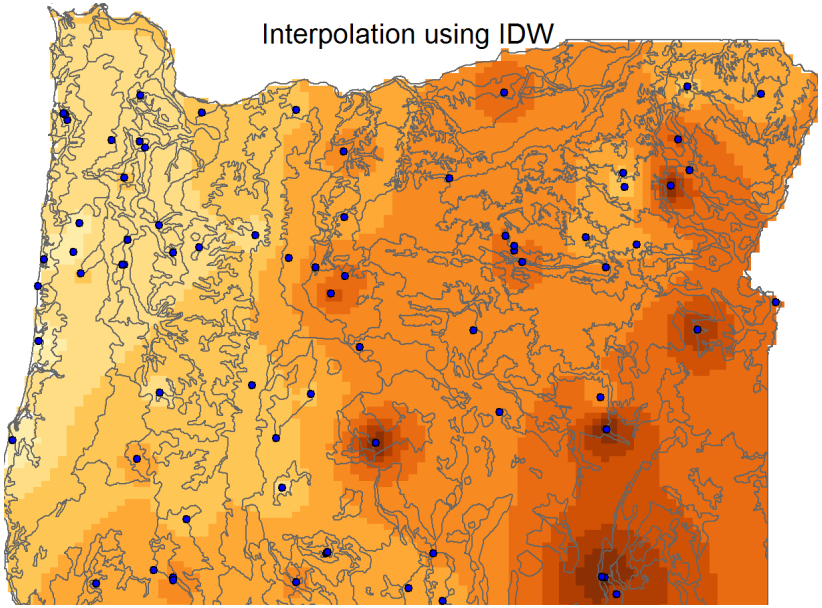
- 探索、或者计算新参数的损失方程可能成本很高（计算成本 或 资金成本）
- 例如，金矿探索钻一个孔可能花费数百万人民币、穷尽计算围棋的胜率变化几乎不可能（围棋的变化约为3的361次方，宇宙原子数约为3的168次方）。
- 可使用贝叶斯优化（Bayesian Optimization）解决该类问题：维纳-柯洛莫高夫滤波/高斯过程/克里金法

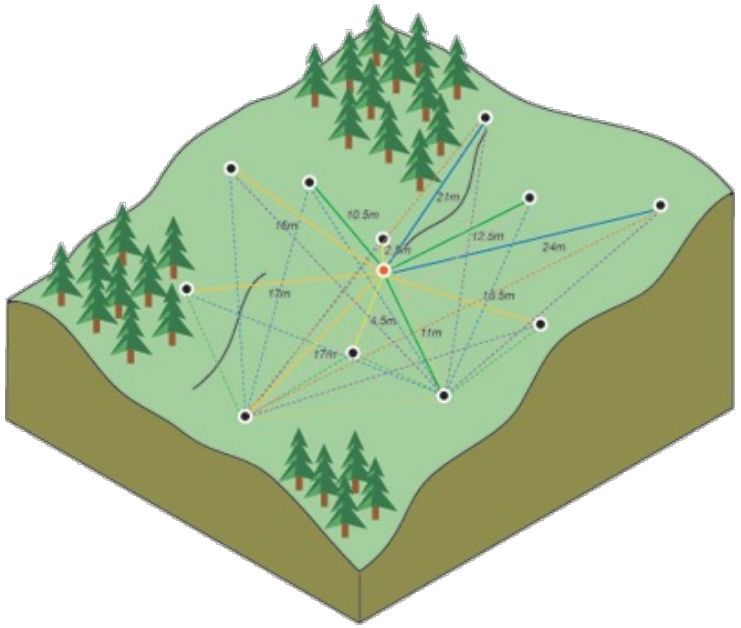


机器学习核心思想3: 贝叶斯优化

继续探索还是停止，这是一个问题；往哪个方向探索，又是一个问题。

Bayesian Optimization





克里金法（Kriging）勘探

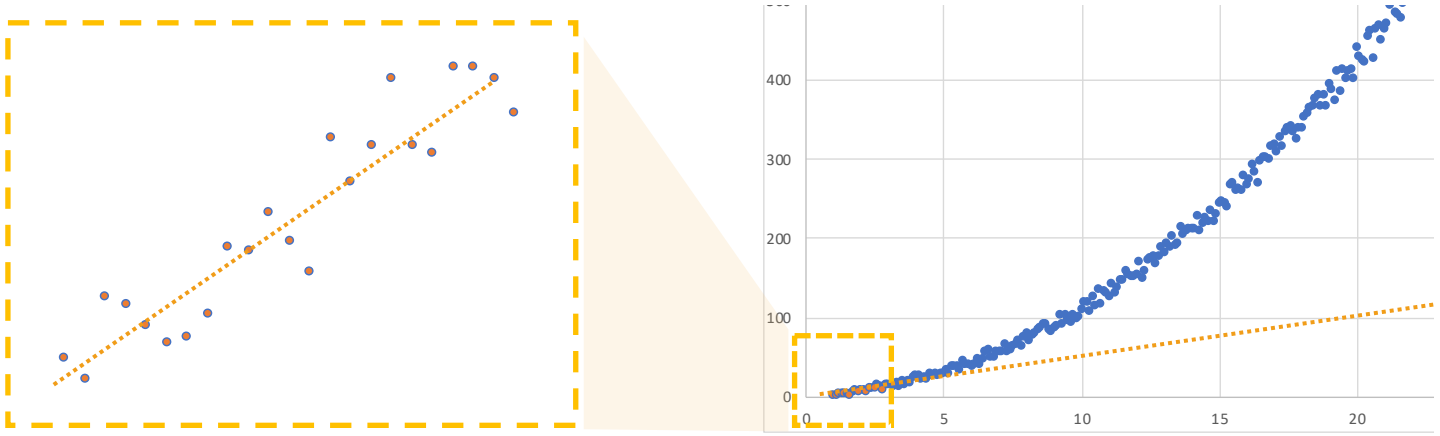
机器学习核心思想4: 训练-测试分开

能纸上谈兵、也要能真正带兵。

Train-Test Separation

“火鸡决然猜不出一直喂它的人类，会在感恩节扭断它的脖子。”--纳西姆·塔勒布《黑天鹅》

样本训练出来的模型得出的结论，可能无法适用于样本外的数据。



解决方案：将数据拆分成不同份，在部份数据上训练（训练集 training Set），其它数据上测试模型效果（测试集 test set），另外还可以准备（验证集 validation set）。

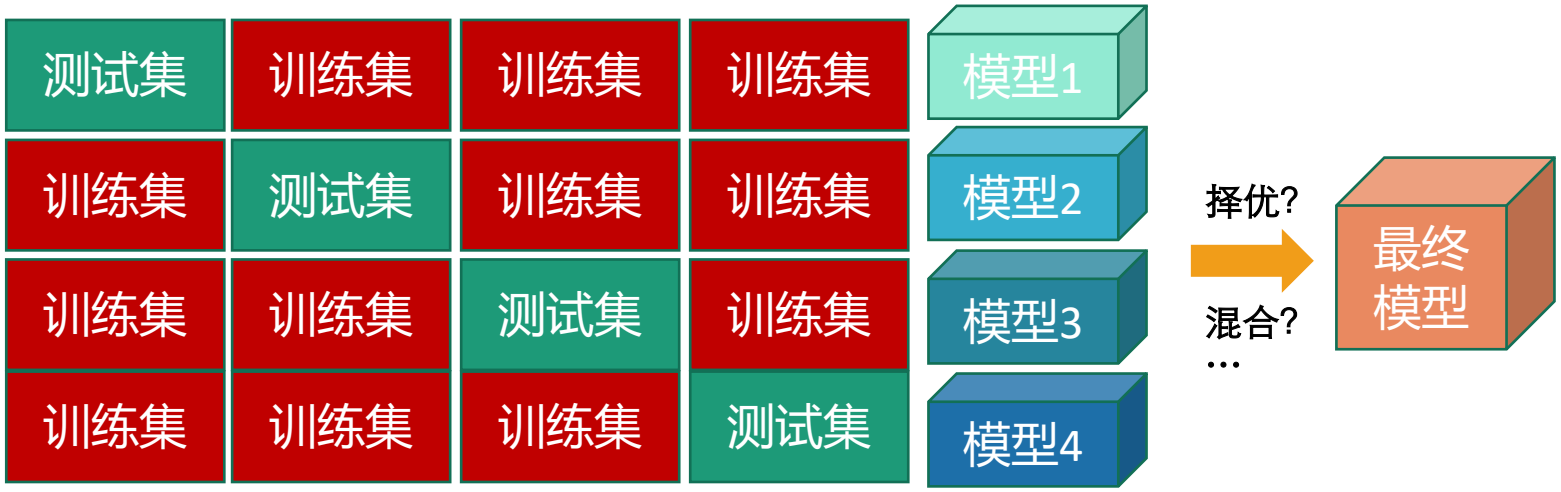
如果模型在两个数据里面都表现好，可能模型是稳定-正确的（鲁棒性 robustness）。

机器学习核心思想4: 训练-测试分开

能纸上谈兵、也要能真正带兵。

Train-Test Separation

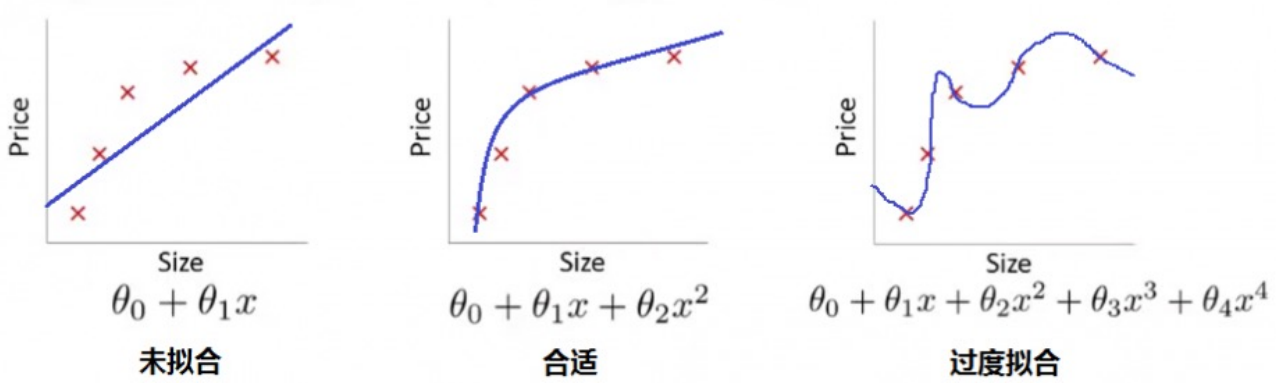
K-Fold 交叉验证 （K = 4） *K-fold Cross-Validation (K=4)*



机器学习核心思想5: 复杂度与鲁棒性

“差不多对”比“准确地错”好，复杂本身是一种成本和风险。

Complexity and Robustness

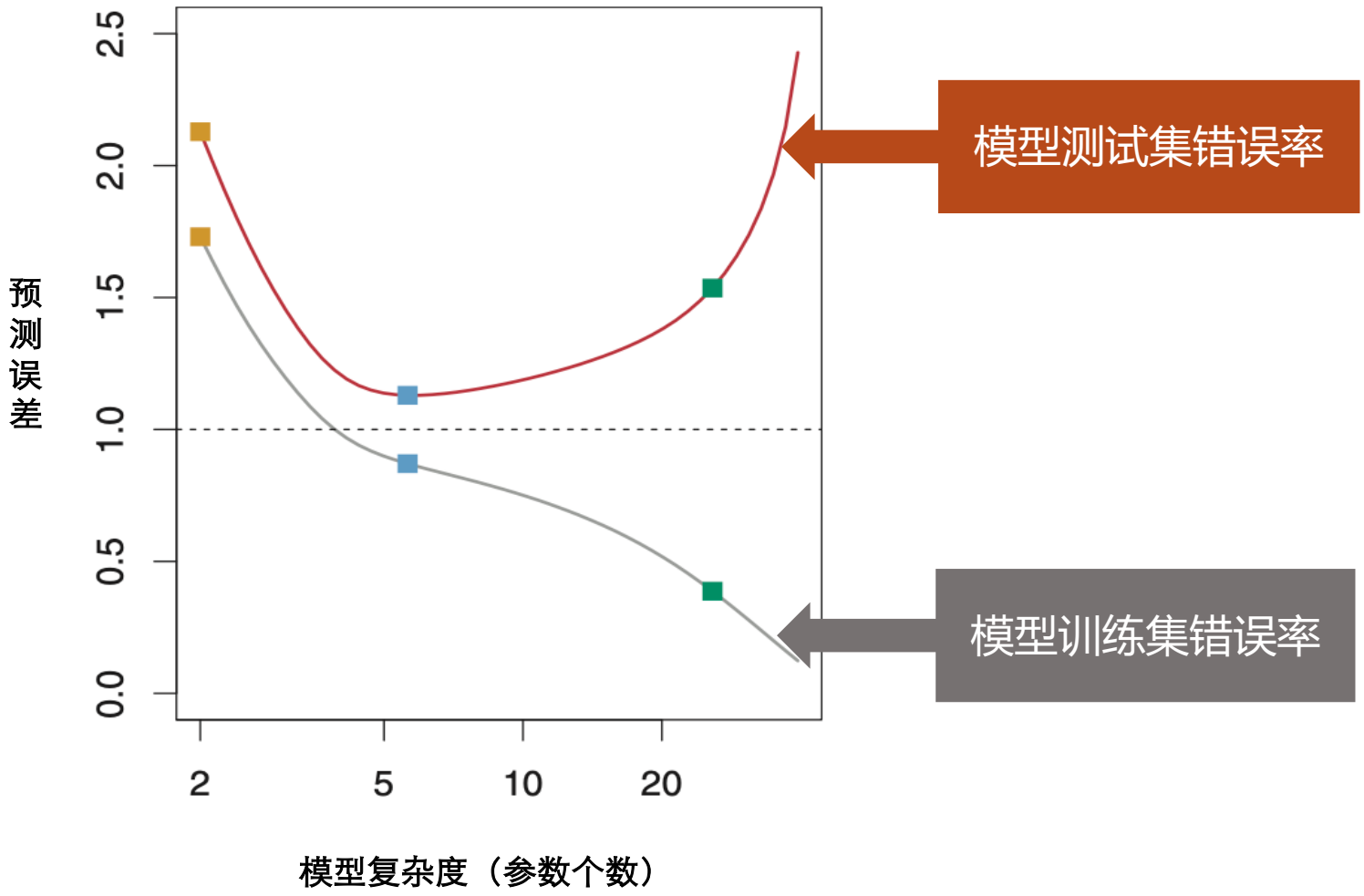


- 过于复杂（参数过多）的模型容易出现在训练集表现优异，而测试集外不灵的现象（低鲁棒性），我们称这样的模型**“过度拟合”（overfitting）**了。
- 如参数数量大于训练数据数，模型必然出现“过度拟合”问题，除非有更高一级的方程限制各个参数间关系，而这样更高一级的方程中的参数我们称之为**“超参数”（hyper-parameters）**。
- 在训练中我们也可以将非零参数数量作为损失方程的“惩罚”而迫使模型将一些不重要变量的参数归零，这样的方法我们称之为**“正则化”（regularization）**，正则化方法可以帮助训练更鲁棒的模型。
- **金融交易算法训练的最大问题是模型过度拟合。**

机器学习核心思想5: 复杂度与鲁棒性

“实践是检验真理的唯一标准”

Complexity and Robustness



机器学习核心思想6: 两种错误、查全、查准

不放走坏人、不冤枉好人 、全面与准确难两全。

Type I ,II Error , Recall, Precision,

- True Positive TP: 正确，预测为正，真实数据也为正。(抓住特务)
- False Negative FN:错误，预测为负，真实数据为正。（一类错误，Type I Error，放走坏人）
- Ture Negative TN: 正确，预测为负，真实数据也为负。(释放群众)
- False Positive FP：错误，预测为正，真实数据为负。（二类错误，Type I Error，冤枉好人）

查全率 (Recall) $= \frac{TP}{TP + FN}$，（抓住特务的比例）

查准率 (Precision) $= \frac{TP}{TP + FP}$，（抓住的人里，真是特务的比例）

- 一般来说，查全率和查准率很难兼顾：

- *以交易算法为例，胜率高（高查准率）的信号往往触发频率低（查准率），导致交易次数过少，利润较低。而提高交易次数往往需要牺牲胜率。*
- *以信用审核算法为例，风险阈值高（高查准率）的模型往往通过率低（查准率），导致市场开拓不利，收入降低。而提高通过率往往意味着更高的违约风险。*

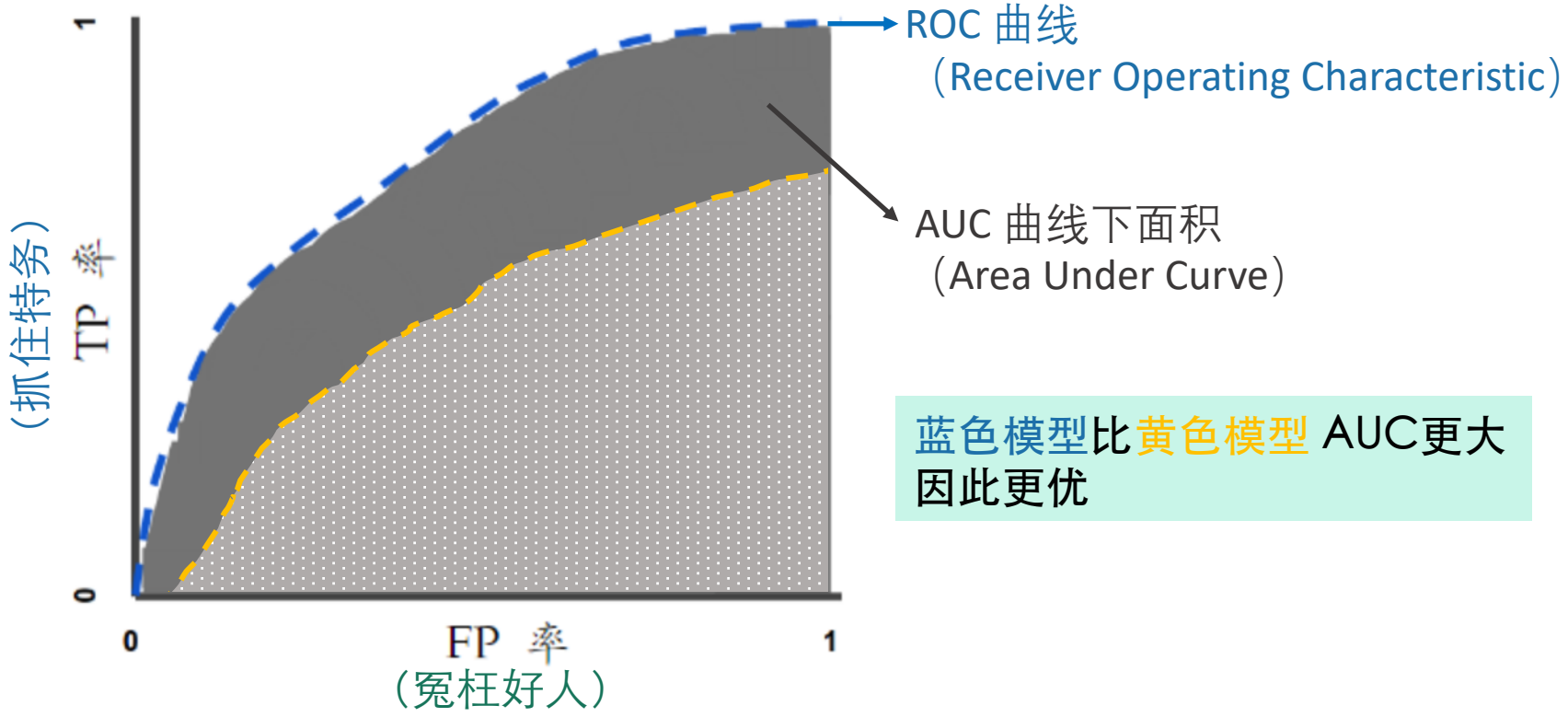
机器学习核心思想6: 两种错误、查全、查准

不放走坏人、不冤枉好人 、全面与准确难两全。

Type I ,II Error , Recall, Precision,

可使用AUC（Area Under Curve）工具解决该类问题:

- 信号处理工具
- 模型AUC越大越好



蓝色模型比黄色模型 AUC更大
因此更优

机器学习核心思想7: 贝叶斯公式

常识与样本该信哪一个？

Bayesian Formula

- A调查某市人口，发现采样的100人里，90个都是男性，因此该市男性比例为 0.9?
- *可能A是在大学男生宿舍采的样?*
- 常识：一般男性占比为0.5
- 贝叶斯派：真实的人口比例是0.9（证据 Evidence）到 0.5 (先验 Prior) 之间的某个数。
- 贝叶斯斯公式（后验 Posterior）：

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

	下雨 Rain（0.20，73天）	不下雨 （0.80，292天）
多云 Cloudy（0.36，132天）	44天	88天
少云（0.64，234天）	30天	204天

似然 likelihood　　先验

$$P(\text{下雨}|\text{多云}) = \frac{P(\text{多云}|\text{下雨})P(\text{下雨})}{P(\text{多云})} = \frac{0.603 \times 0.2}{0.362} = 0.333$$

后验：新证据下的条件概率　　证据

验算：44/(44+88)=0.333

机器学习核心思想7: 贝叶斯公式

“当市场变化时，我的判断也随之变化“ – 凯恩斯

Bayesian Formula

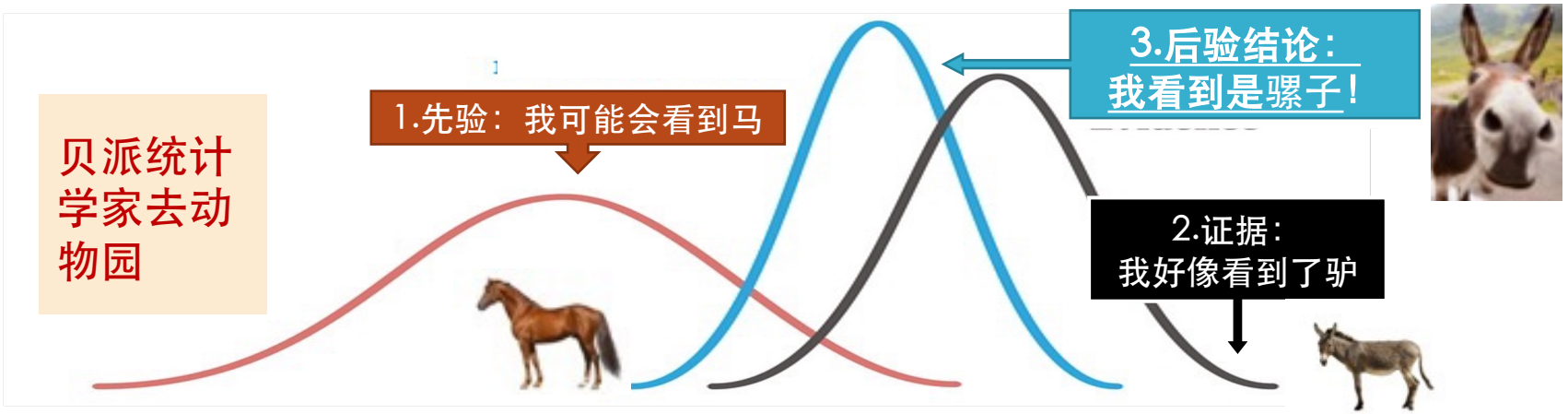
贝叶斯后验概率更新：

- $P(\text{感冒}|\text{头疼}) = \frac{P(\text{头疼}|\text{感冒})P(\text{感冒})}{P(\text{头疼})} = \frac{0.8*0.1}{0.2} = 0.4$

新证据出现时，旧的后验概率成为新的先验概率

- $P(\text{感冒}|\text{发烧、头疼}) = \frac{P(\text{发烧}|\text{感冒、头疼})P(\text{感冒}|\text{头疼})}{P(\text{发烧})} = \frac{0.45*0.4}{0.2} = 0.9$

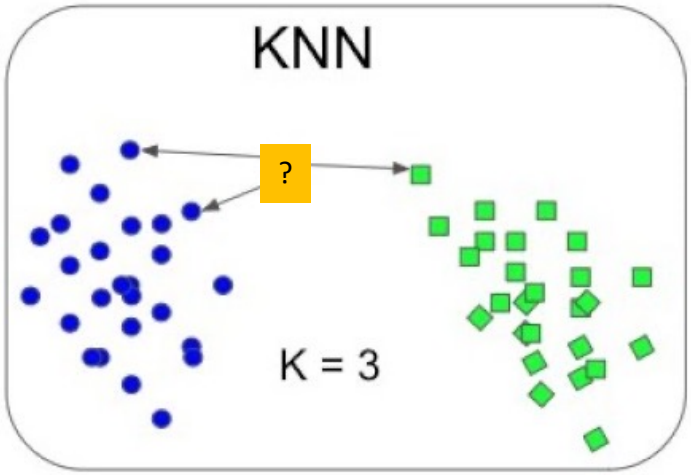
贝叶斯后验概率更新可以看做一种“学习过程”

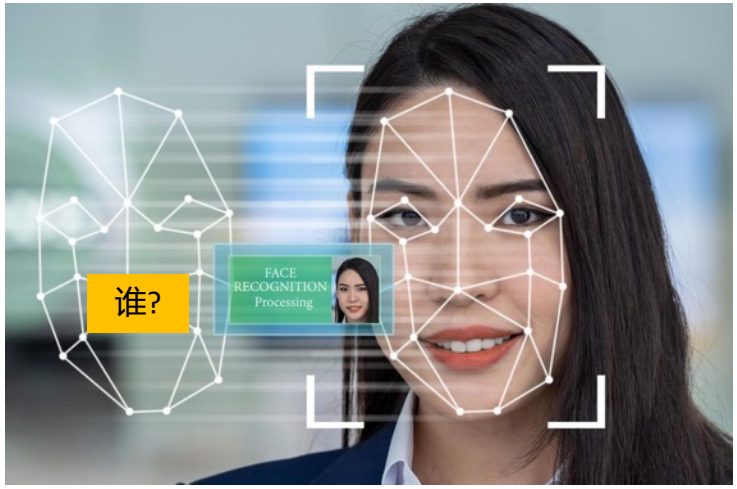


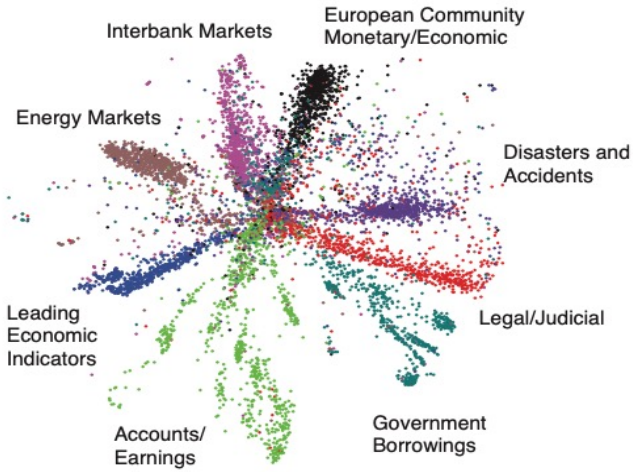
机器学习核心思想8: 距离

近朱者赤。

Distance



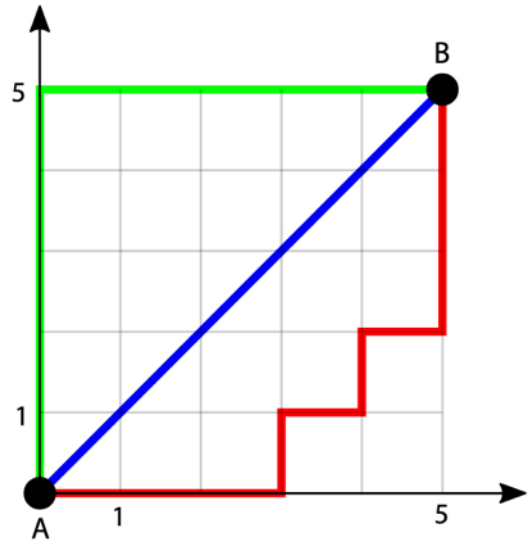




机器学习核心思想8: 距离

近朱者赤，但什么是“近”？

Distance



欧几里得距离、曼哈顿距离…

内积： $X \cdot Y = \sum_{i=0}^{n} x_i \times y_i$

佩奇排序： $PR(A) = \frac{1-d}{N} + d\left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \cdots\right).$

汉明距离： 1011101与1001001之间的汉明距离是2

余弦相似性： $\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$

- 思考：两个公司的相似度如何衡量？它有什么用？

“我这一生都用在了度量美国梦与美国现实的距离上。”

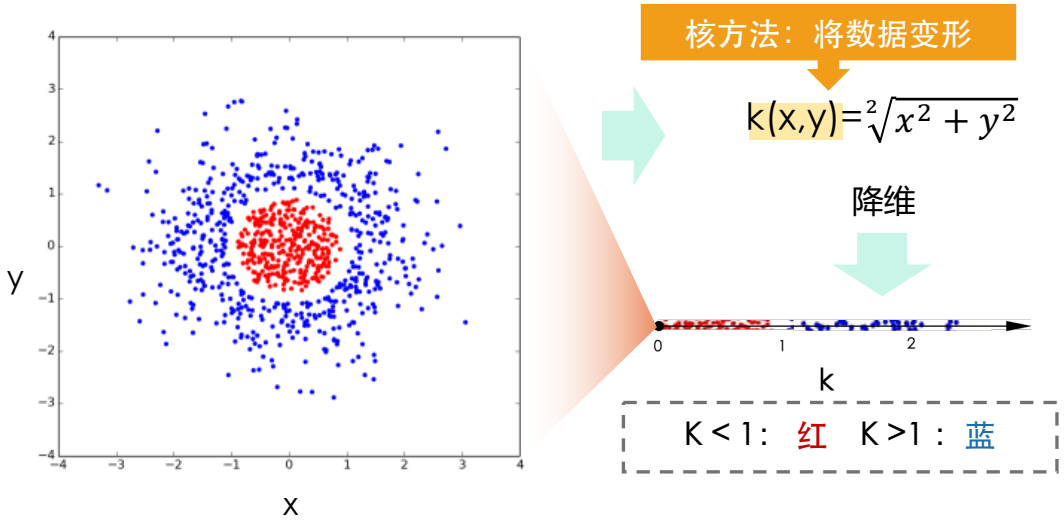
-- 布鲁斯·斯普林斯汀

美国政治批评家、歌手

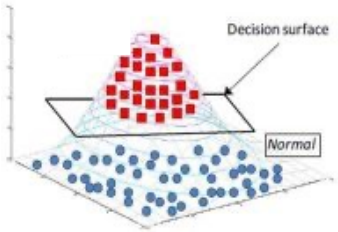
机器学习核心思想9: 核方法

数据变形、换个角度看问题

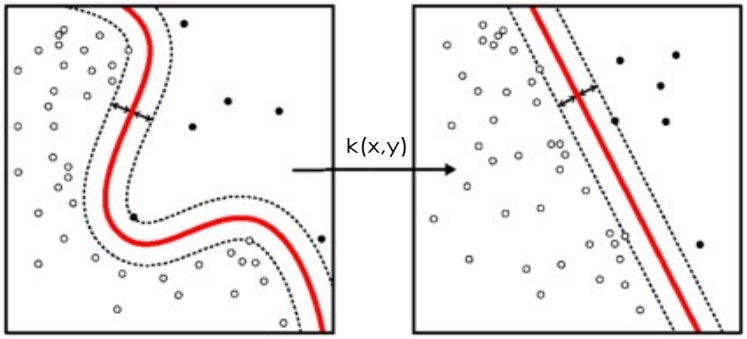
Kernel Methods



升维视角



- 指数、信用分等都可以看做是核方法。
- 使用好的核方程（kernel function）能够显著加大模型训练的成功率和模型适用范围。
- 例：支持向量机（Support Vector Machine，SVM）使用核方法后可以处理非线性分类问题。
- SVM是网络信用贷款平台的常用决策算法之一



机器学习核心思想10: 维度灾难

维数越多，学习成本越大，成功越难。

Curse of Dimensionality

是否可以通过收集越来越多种类（维度）的数据增加模型预测能力?

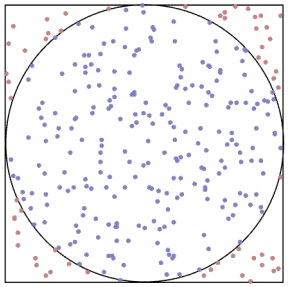
- 是也不是。
- 高纬度会给机器学习算法带来种种困难
    1. 如果模型训练1维数据需要10的数据量，训练2维数据可能需要$(10)^2$= 100 的数据量（也取决于模型本身）。
    2. 以KNN为例，如果1维空间最近点的距离是2， 2维空间可能变成、$2\sqrt{2}$， 3维空间可能变成$2\sqrt{3}$
    3. 高维数据的计算量几何上涨。
- 一味通过增加维度改进模型是得不偿失的，必须有所选择（如使用正则法）。

机器学习核心思想11: 抽样与实验

一叶可以知秋，百叶更可以知秋

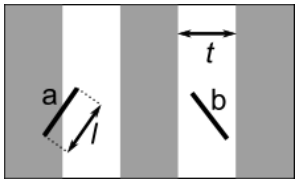
Sampling

- 实验是找到答案的最佳最直接方案之一，抽样是核心。
  - 同直径和边长的园和正方形面积的比可以通过抽样样本数的比来算：



$$\frac{\text{圆形面积}}{\text{方形面积}} \approx \frac{\text{蓝点数}}{\text{红点数} + \text{蓝点数}}$$

  - 同理圆周率可以用“丢针”的办法来算（布丰投针实验, 1733年）



$$\text{圆周率} \approx \frac{2\times\text{针长}\times\text{总针数}}{\text{地板砖宽度}\times\text{与地缝相交的针数}}$$

- 这类通过（计算机辅助）抽样-实验求解的方法叫做**蒙地卡罗法（Monte Carlo Methods）**为计算困难或不可求积分（贝叶斯后验计算经常涉及积分）的主力算法，最早期被用于计算核裂变，而最著名的金融应用例子是基于蒙地卡罗的期权价格计算。

机器学习核心思想11: 抽样与实验

“均匀”不等于“代表性”

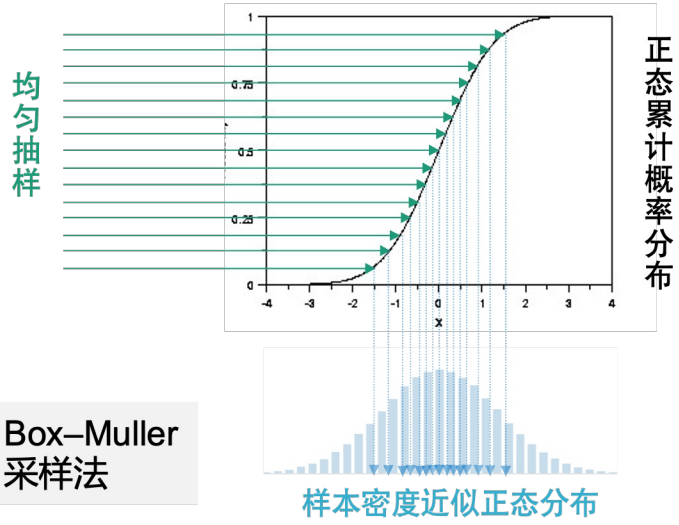
Sampling

“随机抽样”并不不简单，抽样的重点在于样本要有“代表性”**-- 样本特性分布要和总特性分布近似（概率密度高的地方样本多，反之亦然）。**

- 反例：在男生宿舍抽样学校性别比例
- 反例（均匀抽样）：每个年龄段选一人来分析人口年龄分布

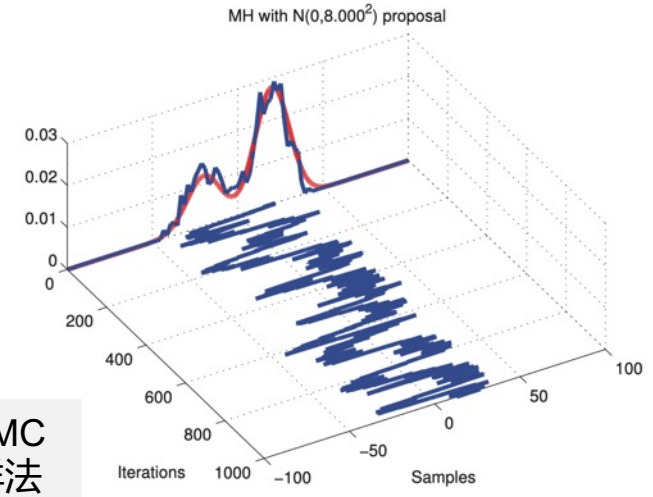
剑桥大学通过Slice采样法找到了肉眼不可见的新星系

例1：从正态分布采样



Box–Muller 采样法

例2：其它分布采样

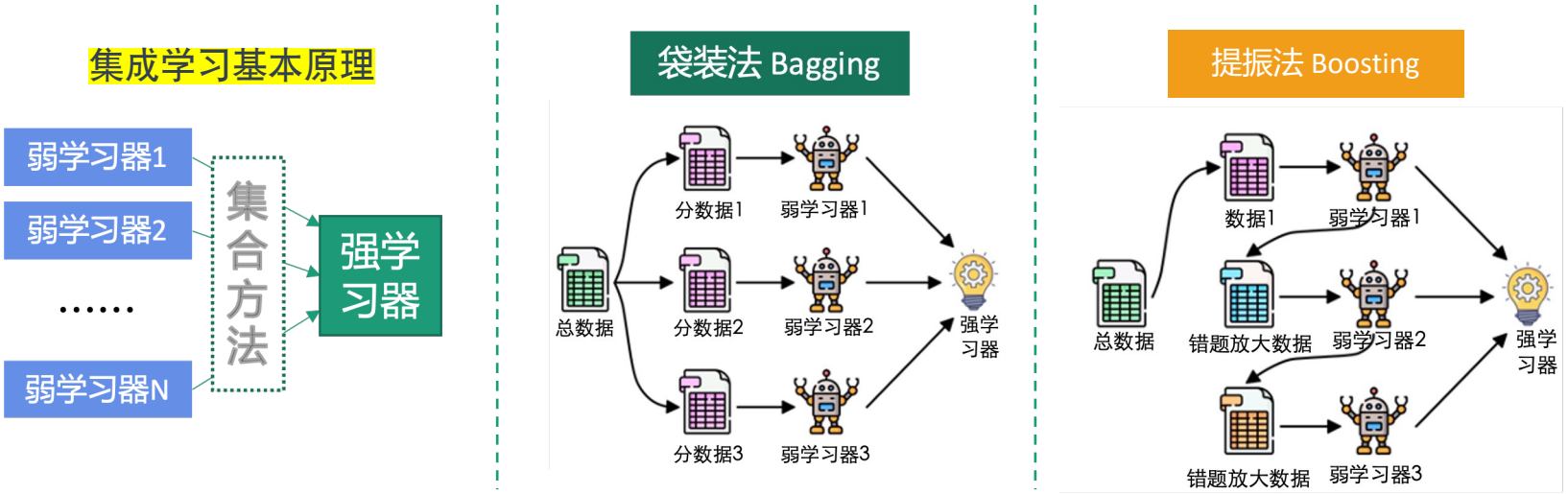


MCMC 采样法

机器学习核心思想12: 集成学习

三个臭皮匠，赛过诸葛亮。

Ensemble Learning

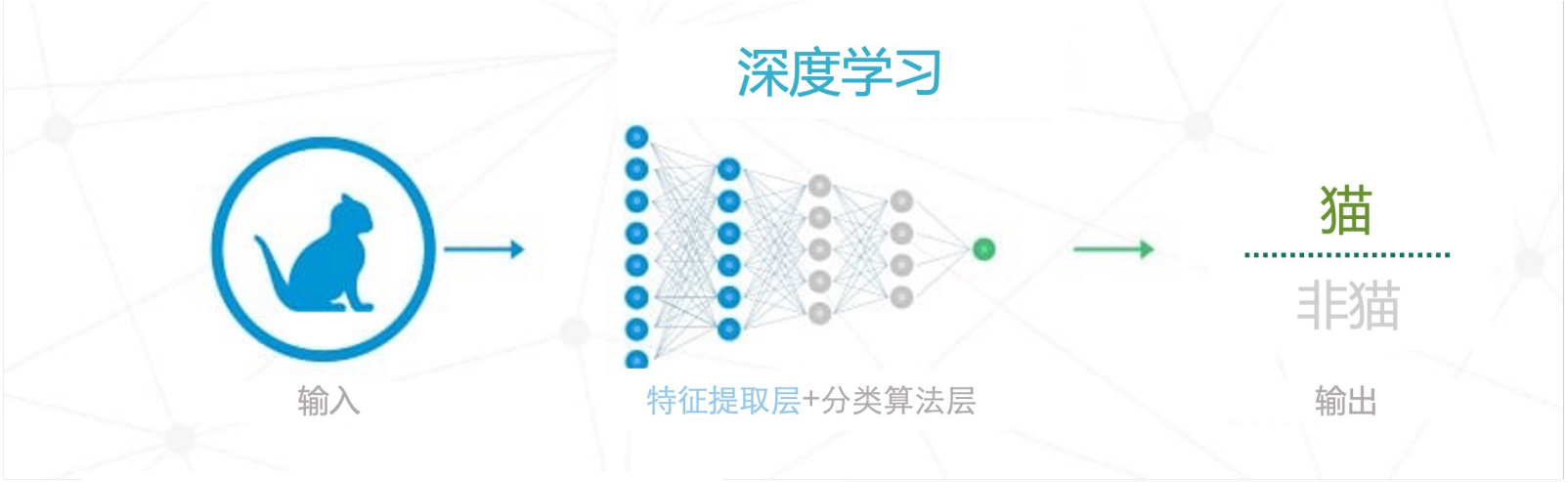


- 业界最实用的算法之一，几乎所有的复杂人工智能程序都用到了集成算法。
- 基本思路是“三个臭皮匠，赛过诸葛亮”。
- 但臭皮匠也需要比“瞎猜”强（>50%的正确率）
- 多种集成方式：袋装法、提振法、随机森林、贝叶斯委员会、堆叠法…

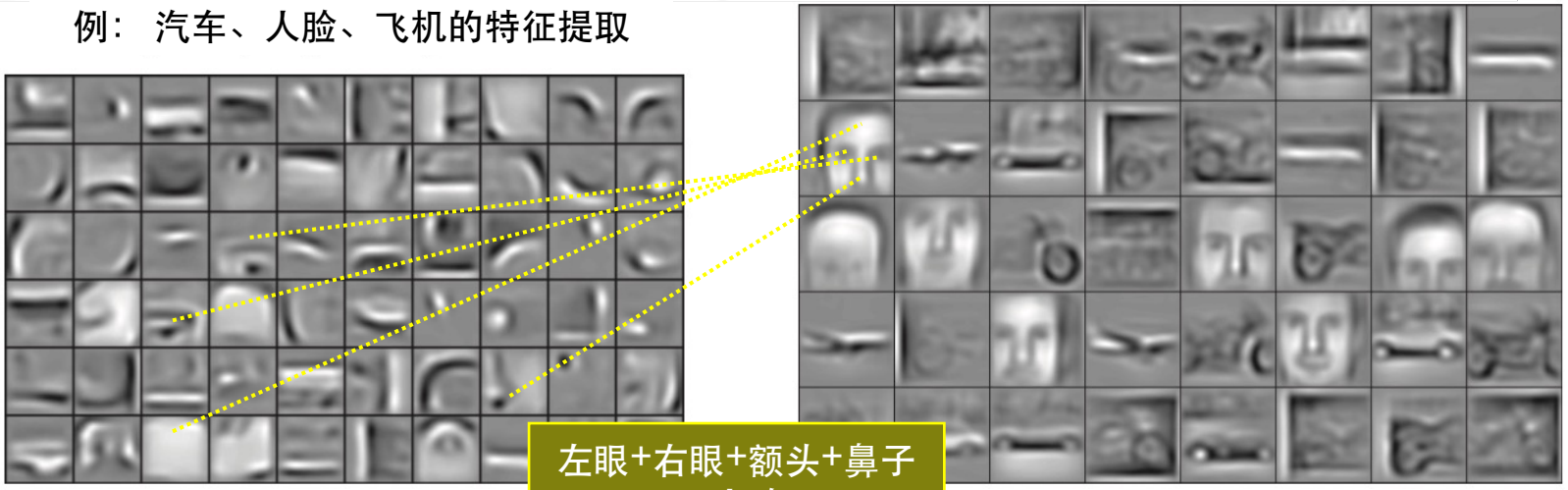
机器学习核心思想13: 深度学习

层次产生智能。

Deep Learning



例： 汽车、人脸、飞机的特征提取



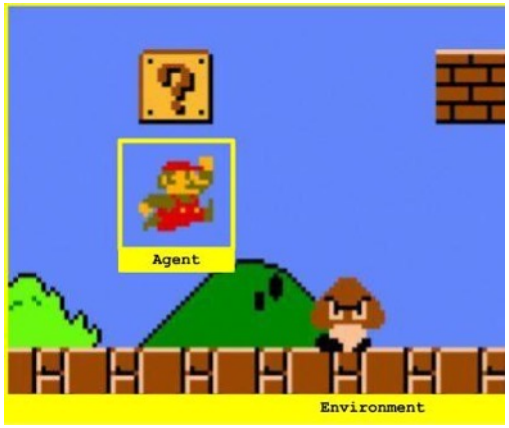
机器学习核心思想14: 增强学习

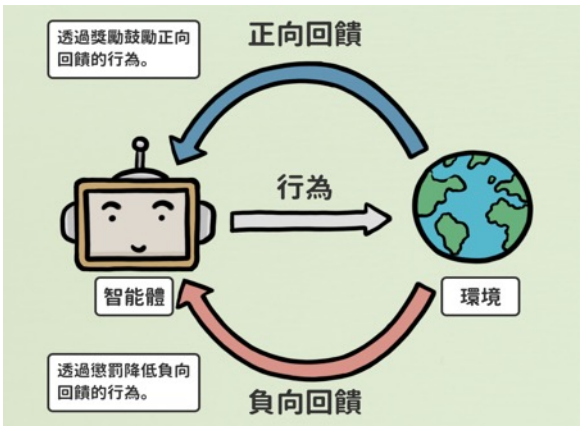
电子达尔文主义（固定环境+确定赏罚+随机变异+优胜劣汰）

Reinforcement Learning

用计算机建立个固定虚拟环境、设立随机行动的主体、定下赏罚规则、淘汰最差的策略或主体，让其进行“达尔文式进化”，是否能训练出优秀的人工智能?

- 能，学会了玩超级玛丽、学会了走楼梯。





- 还没完全学会炒股。

  因为增强学习需要环境比较稳定。

- 已经应用在部分银行的服务推荐系统。

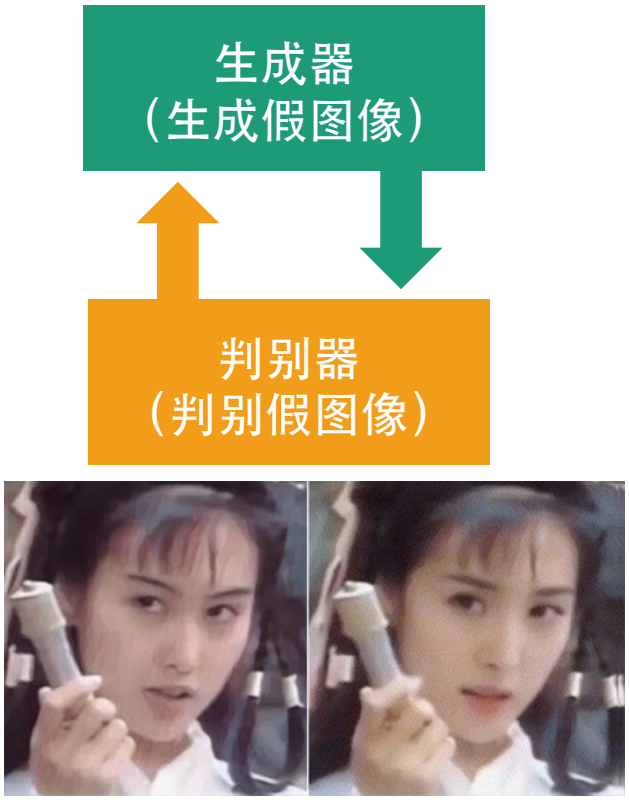
机器学习核心思想15: 对抗学习

对抗是训练的有效方法之一: 目的相同的竞争 以及 目的相反的对抗

Adversarial Learning

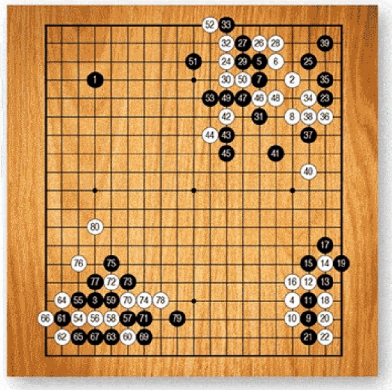
两个或多个人工智能模型互相对抗、并从对抗数据中训练和迭代可以不断改进算法。

对抗神经网络



AlphaGo: 大部分从人类对局数据中训练。

AlphaGo Zero:
(强化学习+对抗学习) 自己跟自己训练。



AlphaGo Zero

从零开始，70个小时超过人类最高水平