



2021年秋季 《机器学习概论》课程

# 第一章：简介

主讲：连德富 特任教授 | 博士生导师

邮箱： [liandefu@ustc.edu.cn](mailto:liandefu@ustc.edu.cn)

手机：13739227137

主页： <http://staff.ustc.edu.cn/~liandefu>

# 课程讲师与助教

- 主讲人： 连德富 特任教授 | 博士生导师  
@大数据学院 | 大数据分析与应用安徽省重点实验室
- 办公室： 西区科技楼 东楼715室
- 邮箱： liandefu@ustc.edu.cn
- 研究方向： 数据挖掘、机器学习、深度学习

## 助教



冯超 (博士生@计算机学院)  
邮箱: chaofeng@mail.ustc.edu.cn  
电话: 18755102509  
研究方向: 机器学习



毕昊阳 (博士生@计算机学院)  
邮箱: bhy0521@mail.ustc.edu.cn  
电话: 17707168083  
研究方向: 数据挖掘

# 参考资料



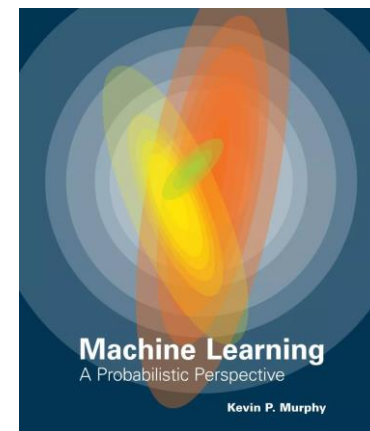
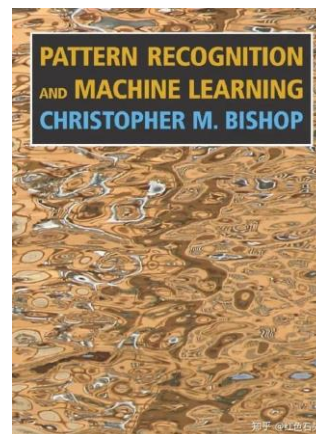
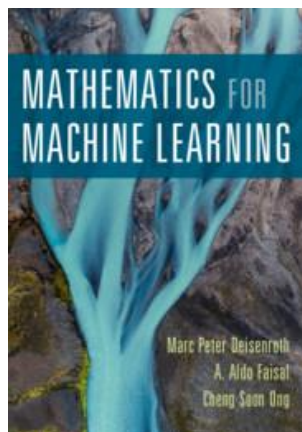
3

- 教科书

- 周志华《机器学习》清华大学出版社

- 参考书

- 李航《统计机器学习》清华大学出版社
  - Christopher M. Bishop 《[Pattern Recognition and Machine Learning](#)》 Springer
  - Kevin P. Murphy 《[Machine Learning: A probabilistic perspective](#)》 MIT Press
  - Marc Peter Deisenroth et al. 《[Mathematics for Machine Learning](#)》 Cambridge



# 阅读材料

- 机器学习领域的重要会议

- ICML (International Conference on Machine Learning)
- NeurIPS (Conference on Neural Information Processing Systems)
- ICLR (International Conference on Learning Representations)
- COLT (Conference on Learning Theory)
- ECML (European Conference on Machine Learning)
- ACML (Asian Conference on Machine Learning)
- CCML (中国机器学习大会)

- 机器学习领域的重要期刊

- Journal of Machine Learning Research (JMLR)
- Machine Learning Journal (MLJ)
- IEEE Transactions on Pattern Recognition and Machine Intelligence (PAMI)

# 课程安排与考核方式

- 笔试
  - 期中占比10%
  - 期末占比30%
- 作业
  - 占比30%
  - 大概15次作业，平均每次4-5道题，每2次提交一次作业
- 上机
  - 占比30%
  - 单独完成约6次上机小实验
- **总成绩=期末 (30%) +作业 (30%) +上机 (30%) +期中 (10%)**

# 课程安排与考核方式

- 笔试

- 期中占比10%
- 期末占比30%

- 作业

- 占比30%
- 大概15次作业，平均

- 上机

- 占比30%
- 单独完成约6次上机

- **总成绩=期末 (30%) + 期中 (10%)**





# 课程目标

- 理解机器学习的基本概念
- 理解机器学习基础理论，熟练掌握机器学习算法原理和工程实现
- 掌握机器学习在实际问题中的应用
  - 特征抽取与预处理
  - 模型选择与调参
  - 实验方法

# 人工智能

- 人工智能（artificial intelligence, AI）就是让机器具有人类的智能。
  - “计算机控制” + “智能行为”

人工智能就是要让机器的行为看起来就像是人所表现出的智能行为一样。

John McCarthy (1927-2011)

- 人工智能这个学科的诞生有着明确的标志性事件，就是1956年的达特茅斯（Dartmouth）会议。在这次会议上，“人工智能”被提出并作为本研究领域的名称。



# 弱人工智能 VS 强人工智能

- **弱人工智能**：限制领域人工智能（Narrow AI）或应用型人工智能（Applied AI），指的是专注于且只能解决特定领域问题的人工智能
- **强人工智能**：又称通用人工智能（Artificial General Intelligence）或完全人工智能（Full AI），指的是可以胜任人类所有工作的人工智能。

强人工智能具备以下能力

1. 存在不确定性因素时进行推理，使用策略，解决问题，制定决策的能力
2. 知识表示的能力，包括常识性知识的表示能力
3. 规划能力
4. 学习能力
5. 使用自然语言进行交流沟通的能力
6. 将上述能力整合起来实现既定目标的能力

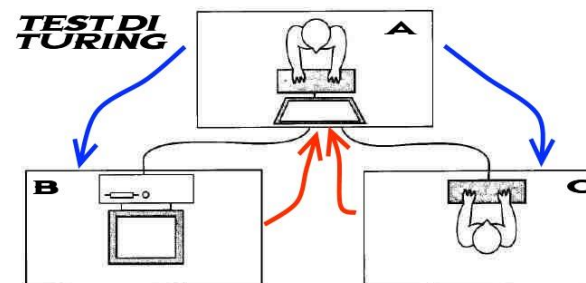
# 图灵测试

“一个人在不接触对方的情况下，通过一种特殊的方式，和对方进行一系列的问答。如果在相当长时间内，他无法根据这些问题判断对方是人还是计算机，那么就可以认为这个计算机是智能的”。

---Alan Turing [1950]  
《Computing Machinery and Intelligence》



Alan Turing



# 人工智能的三个阶段

## • 让机器具有人类的智能

### 感知智能

- 基于视觉、听觉及各种传感器的信息处理
- 表现为“能听会说、能看会认”
- 机器视觉、语音识别、文字识别

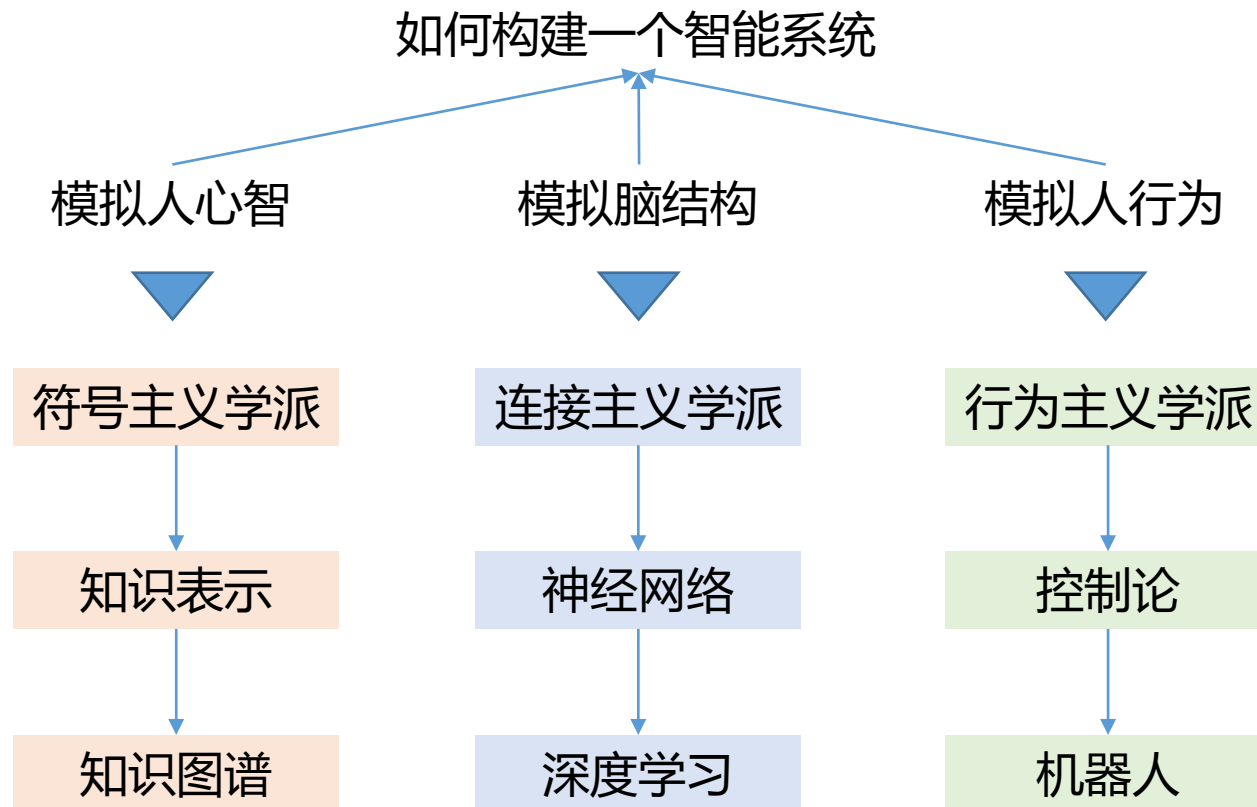
### 认知智能

- 更高层的语义处理、推理、规划、记忆、学习
- 表现为“能理解、会思考、有认知”
- 知识表示、模式识别、机器学习、自然语言处理

### 决策智能

- 复杂问题下，提升人机信任度，增强人类与智能系统交互协作智能
- 表现为“自主性”
- 规划、数据挖掘、强化学习

# 人工智能流派



# 什么是机器学习

Learning is any process by which a system improves performance from experience.

学习是系统从经验中提高性能的任何过程

美国著名学者、计算机科学家和心理学家

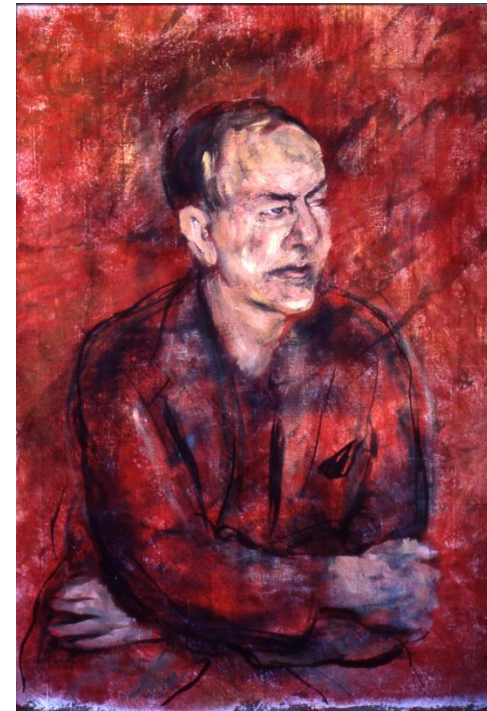
——Herbert Simon (司马贺)  
Carnegie Mellon University

Turing Award (1975)

artificial intelligence, the psychology of human cognition

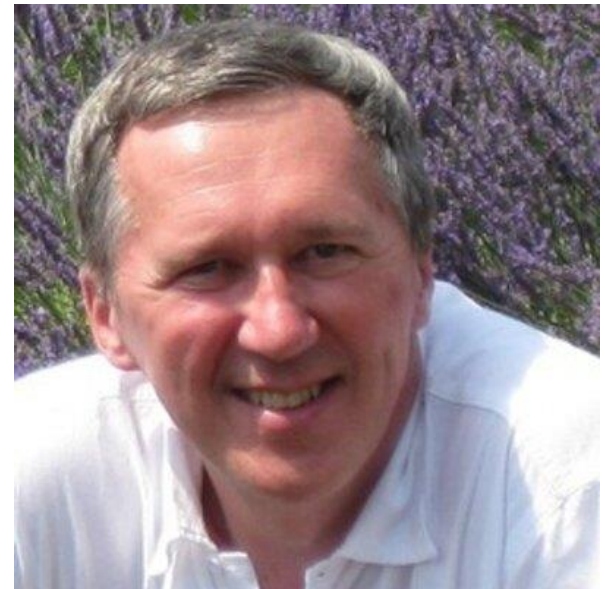
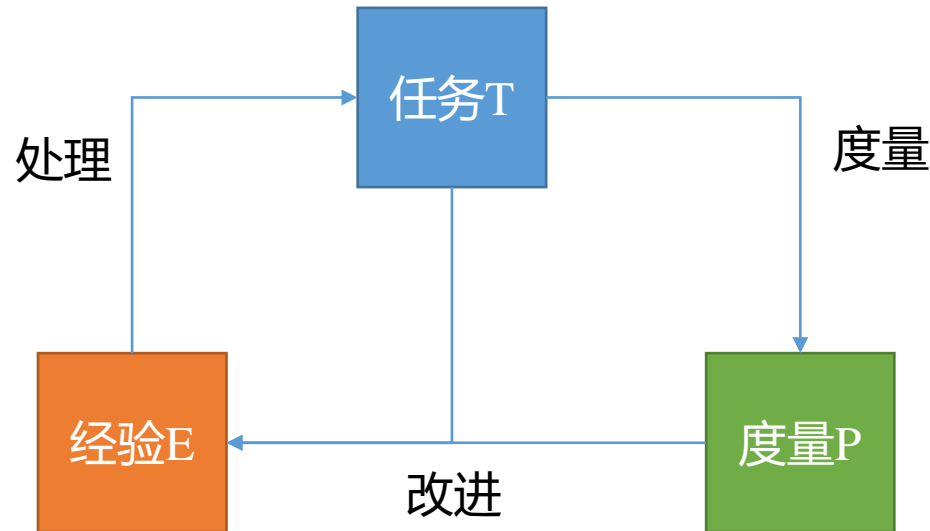
Nobel Prize in Economics (1978)

decision-making process within economic organizations



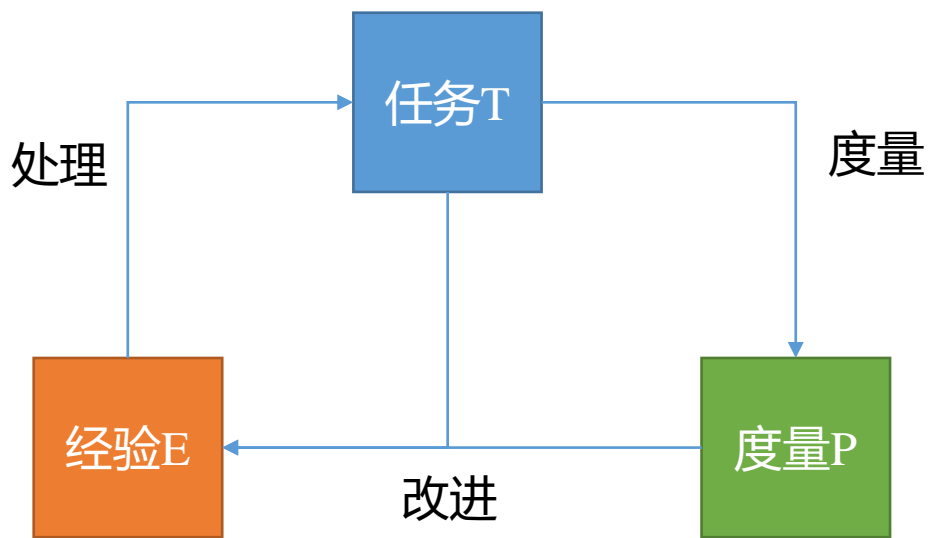
# 什么是机器学习

- 对于某类任务 $T$ 和性能度量 $P$ ，如果一个计算机程序在某些任务 $T$ 上以 $P$ 度量的性能随着经验 $E$ 的增加而提高，那么我们称这个计算机程序是在从经验 $E$ 中学习 —— Tom Mitchell



机器学习致力于研究如何通过计算的手段，利用经验来改善系统自身的性能，从而在计算机上从数据中产生“模型”，用于对新的情况给出判断。

# 什么是机器学习



## 邮件分类

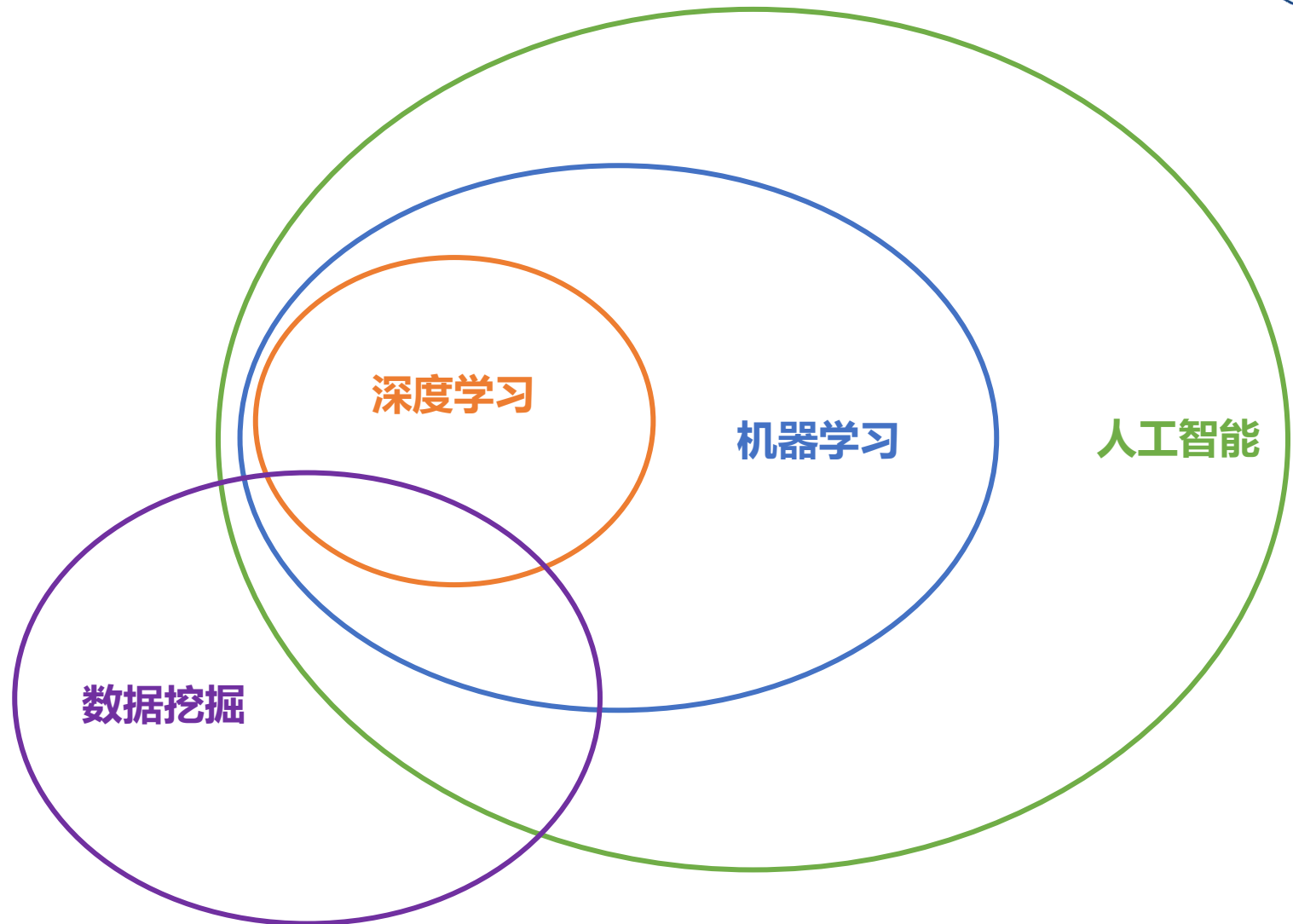
- T: 将电子邮件分类为垃圾邮件或合法邮件
- P: 被分类为垃圾邮件的比例
- E: 邮件数据库和人工标记

## 光学字符识别

- T: 识别手写字符
- P: 有多少比例字符被识别
- E: 手写字符数据库和人工标记

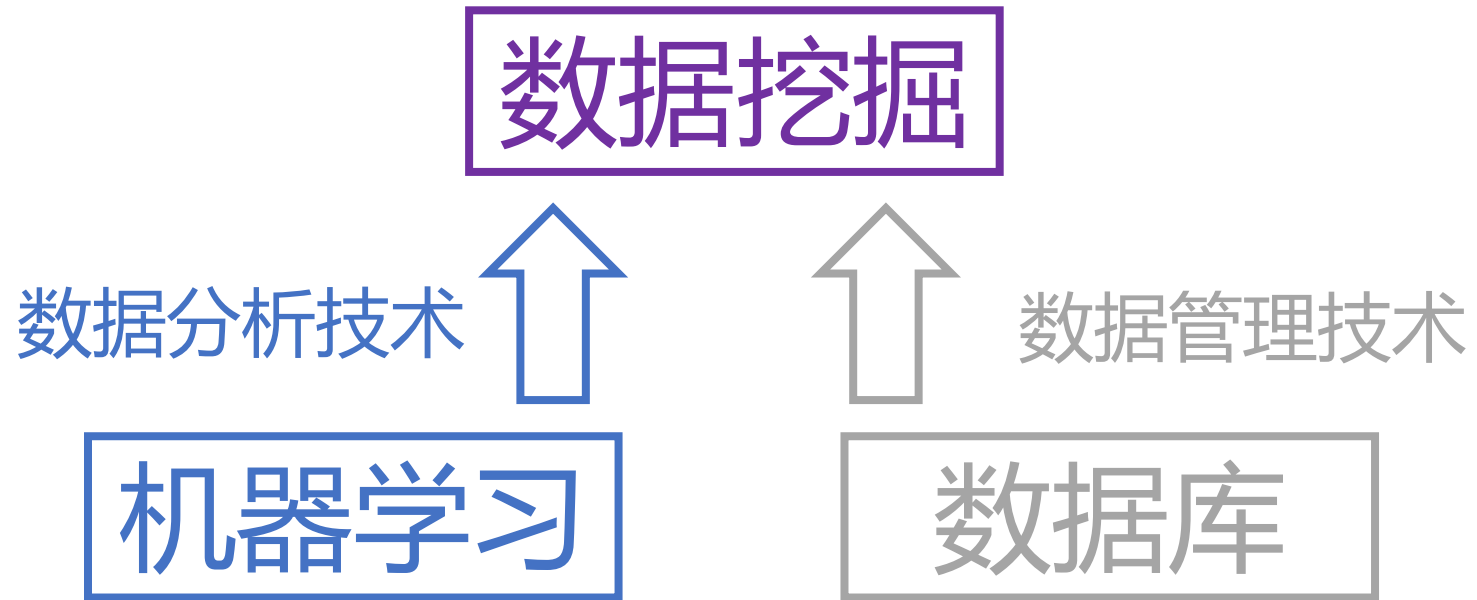


# 什么是机器学习



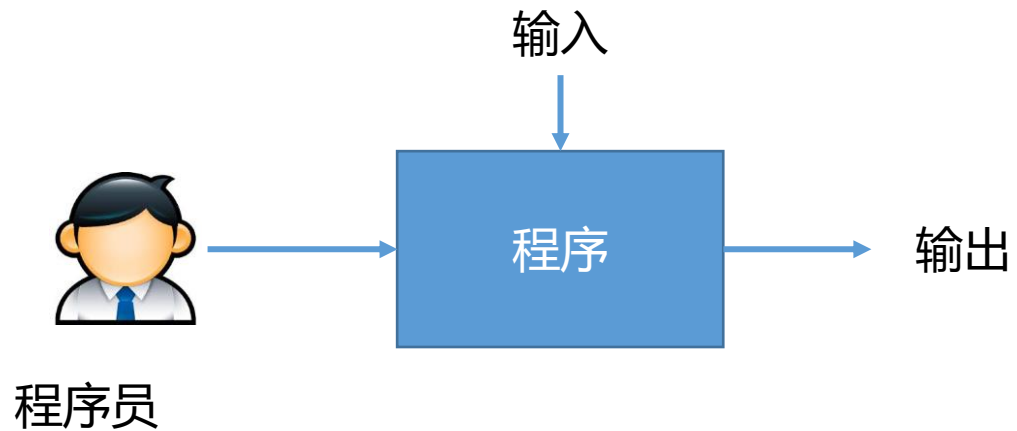


# 什么是机器学习

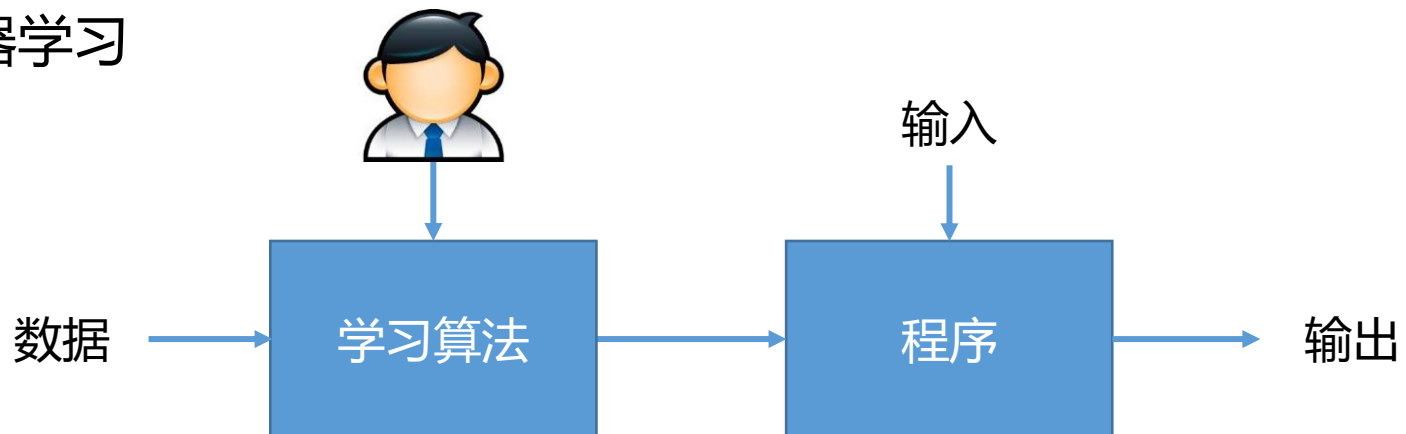


# 机器学习是什么

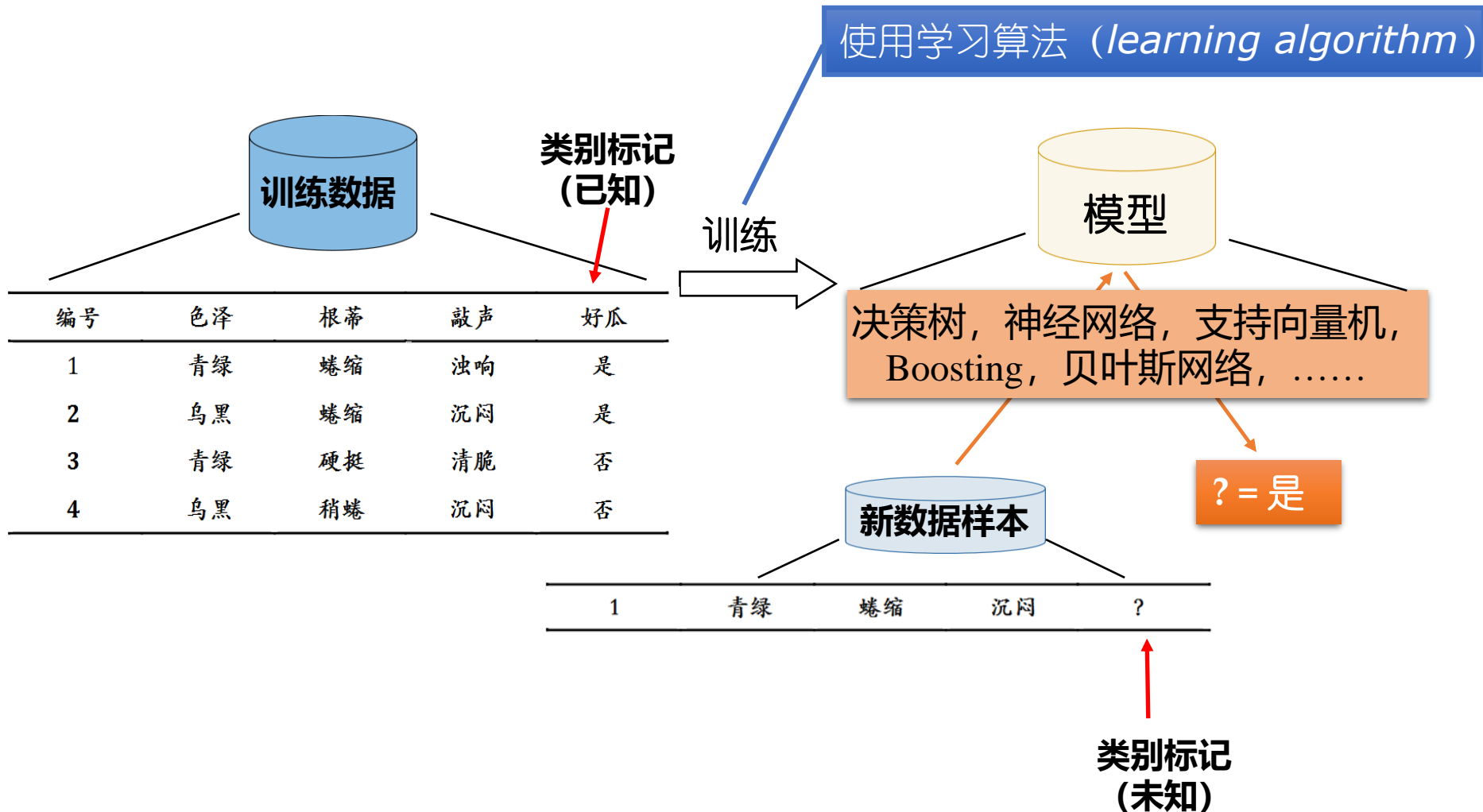
- 传统编程



- 机器学习



# 典型机器学习过程



# 机器学习的基本概念—数据（经验E）

		特征			标记	
		↑			↑	
		编号	色泽	根蒂	敲声	好瓜
		1	青绿	蜷缩	浊响	是 → 训练样本
训练集 ←		2	乌黑	蜷缩	沉闷	是
		3	青绿	硬挺	清脆	否
		4	乌黑	稍蜷	沉闷	否
测试集 ←	1	青绿	蜷缩	沉闷	?	

# 机器学习的基本概念—任务T

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	沉闷	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否

↑  
标记

- 按照**标记**区分
  - 分类：**标记**为离散值（二分类、多分类）
  - 回归：**标记**为连续值（瓜的成熟度）
  - 聚类：没有**标记**

# 机器学习的基本概念—任务T

- 按照标记区分

- 分类：标记为离散值（二分类、多分类）
- 回归：标记为连续值（瓜的成熟度）

监督学习  
Supervised Learning

- 聚类：没有标记

无监督学习  
Unsupervised Learning

监督学习  
Supervised Learning

+

无监督学习  
Unsupervised Learning

=

半监督学习  
Semi-supervised Learning

# 机器学习的基本概念—泛化能力

- 机器学习的目标是使得学到的模型能很好的适用于“新样本”，而不仅仅是训练集合，我们称模型适用于新样本的能力为泛化 (generalization) 能力。
- 通常假设样本空间中的样本服从一个未知分布  $\mathcal{D}$ ，样本从这个分布中独立获得，即“独立同分布” (i.i.d)。一般而言训练样本越多越有可能通过学习获得强泛化能力的模型

# 机器学习的基本概念—假设空间


编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	沉闷	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否



归纳学习 inductive learning

**假设**  
**Hypothesis**

$(\text{色泽}=?)\wedge(\text{根蒂}=?)\wedge(\text{敲声}=?)\leftrightarrow\text{好瓜}$

- 
1. 青绿
  2. 乌黑
  3. \* (通配符)

**在假设空间中搜索不违背训练集的假设**

**假设空间大小:  $3*4*4+1=49$**



$\emptyset$  表示好瓜概念不成立



# 机器学习的基本概念—归纳偏好

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	沉闷	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否



归纳学习 inductive learning

假设空间中有三个与训练集一致的假设

(色泽= \* ; 根蒂=蜷缩 ; 敲声= \*)

好瓜

(色泽= \* ; 根蒂= \* ; 敲声=浊响)

坏瓜

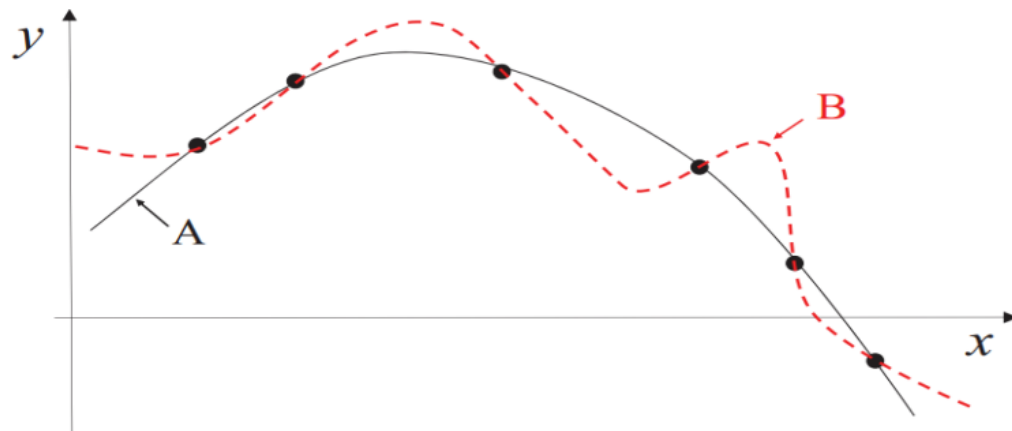
(色泽= \* ; 根蒂=蜷缩 ; 敲声=浊响)

坏瓜

他们对(色泽=青绿;根蒂=蜷缩;敲声=沉闷)的瓜会预测结果不同

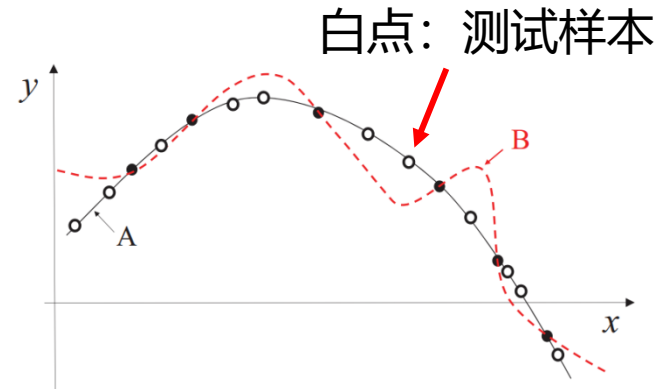
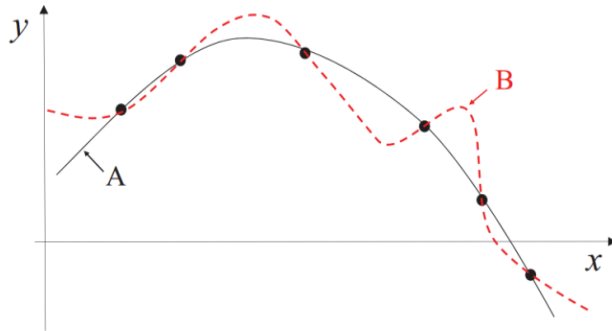
# 机器学习的基本概念—归纳偏好

- 归纳偏好可看作学习算法自身在一个可能很庞大的假设空间中对假设进行选择的启发式或“价值观”。
- “奥卡姆剃刀”是一种常用的、自然科学研究中最基本的原则，即“若有多个假设与观察一致，选最简单的那个”。



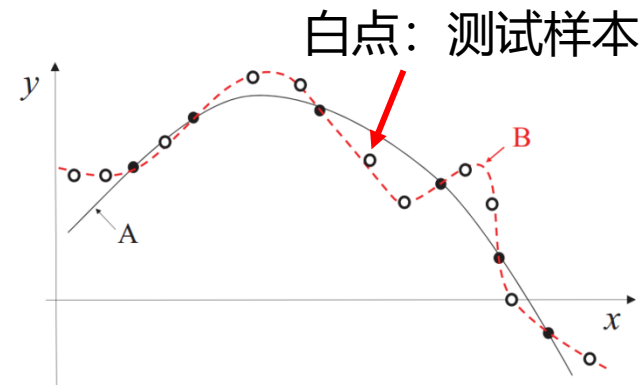
存在多条曲线与有限样本训练集一致

# 机器学习的基本概念—NFL



## 没有免费的午餐 No Free Lunch

一个算法 $\Omega_a$ 如果在某些问题上比另一个算法 $\Omega_b$ 好, 必然存在另一些问题,  $\Omega_b$  比  $\Omega_a$  好



# 机器学习的基本概念—NFL

- 假设样本空间  $\mathcal{X}$  和假设空间  $\mathcal{H}$  离散, 令  $P(h|\mathbf{X}, \mathcal{Q}_a)$  代表算法  $\mathcal{Q}_a$  基于训练数据  $\mathbf{X}$  产生假设  $h$  的概率, 在令  $f$  代表要学的目标函数, 在训练集之外所有样本上的总误差为

$$E_{ote}(\mathcal{Q}_a|\mathbf{X}, f) = \sum_h \sum_{\mathbf{x} \in \mathcal{X} - \mathbf{X}} P(\mathbf{x}) \mathbb{I}(h(\mathbf{x}) \neq f(\mathbf{x})) P(h|\mathbf{X}, \mathcal{Q}_a)$$

$\mathbb{I}(\cdot)$  为指示函数, 若  $\cdot$  为真取值1, 否则取值0

考虑二分类问题, 目标函数可以为任何函数  $\mathcal{X} \mapsto \{0,1\}$ , 函数空间为  $\{0,1\}^{|\mathcal{X}|}$ . 对所有可能  $f$  按**均匀分布**对误差求和, 有如下结论:

$$\sum_f E_{ote}(\mathcal{Q}_a|\mathbf{X}, f) = 2^{|\mathcal{X}|-1} \sum_{\mathbf{x} \in \mathcal{X} \setminus \mathbf{X}} P(\mathbf{x}) \quad \text{总误差与学习算法无关!}$$

# 机器学习的基本概念—NFL

考虑二分类问题，目标函数可以为任何函数  $\mathcal{X} \mapsto \{0,1\}$ ，函数空间为  $\{0,1\}^{|\mathcal{X}|}$ 。  
对所有可能  $f$  按**均匀分布**对误差求和，有如下结论：

$$\sum_f E_{ote}(\mathcal{Q}_a | \mathbf{X}, f) = 2^{|\mathcal{X}|-1} \sum_{\mathbf{x} \in \mathcal{X} \setminus \mathbf{X}} P(\mathbf{x})$$

• 证明如下

$$\begin{aligned} \sum_f E_{ote}(\mathcal{Q}_a | \mathbf{X}, f) &= \sum_f \sum_h \sum_{\mathbf{x} \in \mathcal{X} - \mathbf{X}} P(\mathbf{x}) \mathbb{I}(h(\mathbf{x}) \neq f(\mathbf{x})) P(h | \mathbf{X}, \mathcal{Q}_a) \\ &= \sum_{\mathbf{x} \in \mathcal{X} - \mathbf{X}} P(\mathbf{x}) \sum_h P(h | \mathbf{X}, \mathcal{Q}_a) \sum_f \mathbb{I}(h(\mathbf{x}) \neq f(\mathbf{x})) \\ &= \sum_{\mathbf{x} \in \mathcal{X} - \mathbf{X}} P(\mathbf{x}) \sum_h P(h | \mathbf{X}, \mathcal{Q}_a) \frac{1}{2} 2^{|\mathcal{X}|} \\ &= 2^{|\mathcal{X}|-1} \sum_{\mathbf{x} \in \mathcal{X} \setminus \mathbf{X}} P(\mathbf{x}) \end{aligned}$$

函数空间大小  $2^{|\mathcal{X}|}$   
均匀分布，一半的  $f$  对  $\mathbf{x}$   
的预测和  $h$  不一致

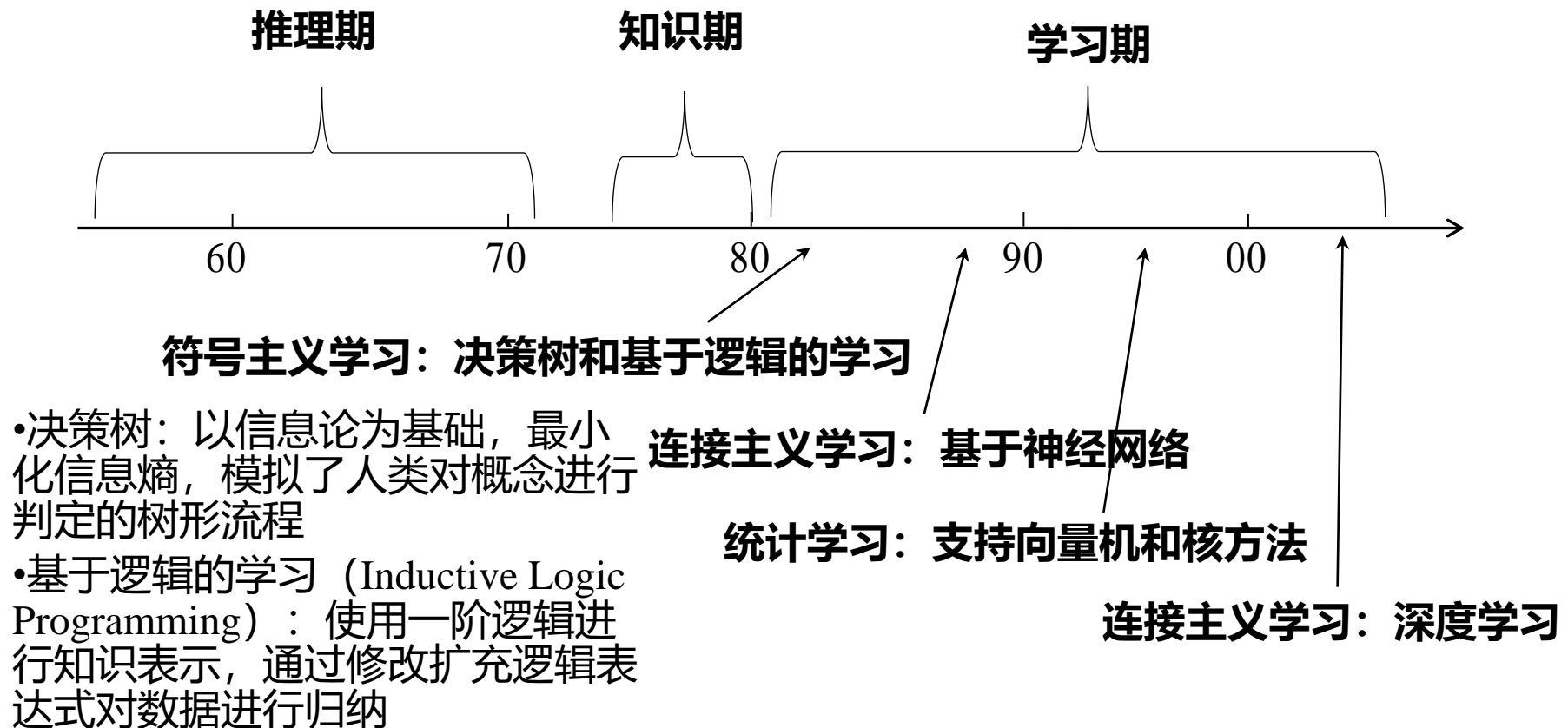
# 机器学习发展历程—推理期

- 基于符号知识表示，通过演绎推理技术取得了巨大成绩
- A. Newell和H. Simon的“逻辑理论家” (Logic Theorist)程序以及伺候的“通用问题求解” (General Problem Solving)程序等在当时取得了令人振奋的结果。
  - 1952年证明了著名数学家罗素和怀特海名著《数学原理》38条定理
  - 1963年证明了全部52条定理
  - 定理2.85比罗素和怀特海证明的更巧妙
  - 获得了1975年的图灵奖

# 机器学习发展历程—知识期

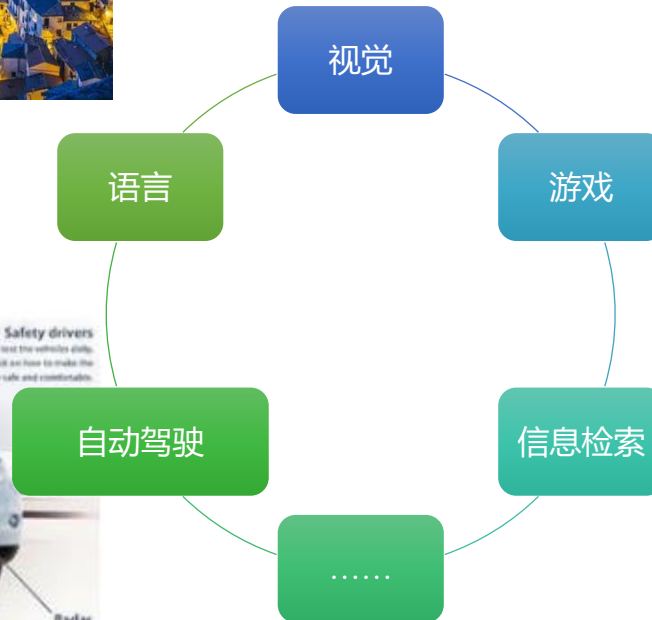
- 基于符号知识表示，通过获取和利用领域知识建立专家系统取得大量成果
- “知识工程”之父Edward A. Feigenbaum 研制了世界上第一个专家系统DENDRAL，并获得了1994年的图灵奖
  - DENDRAL输入的是质谱仪的数据，输出是给定物质的化学结构。
  - 捕捉化学家的化学分析知识，把知识提炼成规则。
  - 质谱数据就是关于原子重量的信息，能够帮助化学家确定一种化合物的结构和性质。
- 但是由人来总结知识再交给计算机相当困难。

# 机器学习发展历程



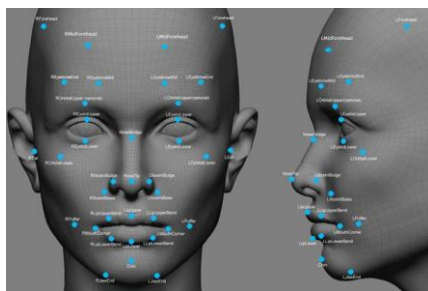


# 机器学习应用



# 机器学习取得了巨大的进展

- **语音识别**：微软英语语音识别实现词错率5.9%的突破，第一次**超越人类**
- **人脸识别**：Facebook的人脸识别系统DeepFace达到97.53%的准确率，**达到人类水平**
- **图像识别**：微软在ImageNet图像数据库上，达到4.94%的错误率，**低于人类5.1%的错误率**
- **人机对弈**：AlphaGo以4：1的战绩击败李世石，Master在围棋快棋上击败柯杰，聂卫平等高手，取得**60胜0负**的战绩

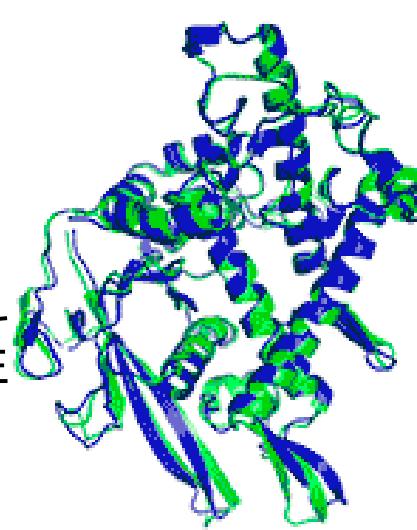


# 人工智能取得了巨大的进展

- 蛋白质结构预测  
( AlphaFold ) :

AlphaFold在数天内通过蛋白质序列来预测蛋白质结构，此前科学家识别蛋白质形状需花费数年时间。

蛋白质通过卷曲折叠会构成三维结构，蛋白质的功能正由其结构决定，了解蛋白质结构有助于开发治疗疾病的药物。



T1037 / 6vr4  
90.7 GDT  
(CRNA polymerase domain)



T1049 / 6y4f  
93.3 GDT  
(adhesin tip)

● Experimental result  
● Computational prediction

# 机器学习应用—图片搜索

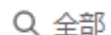


## 外观类似的图片



924f8...5a5f7.jpeg ×

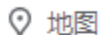
蔡徐坤 打篮球



全部



图片



地图



购物

找到约 576 条结果 (用时 0.83 秒)



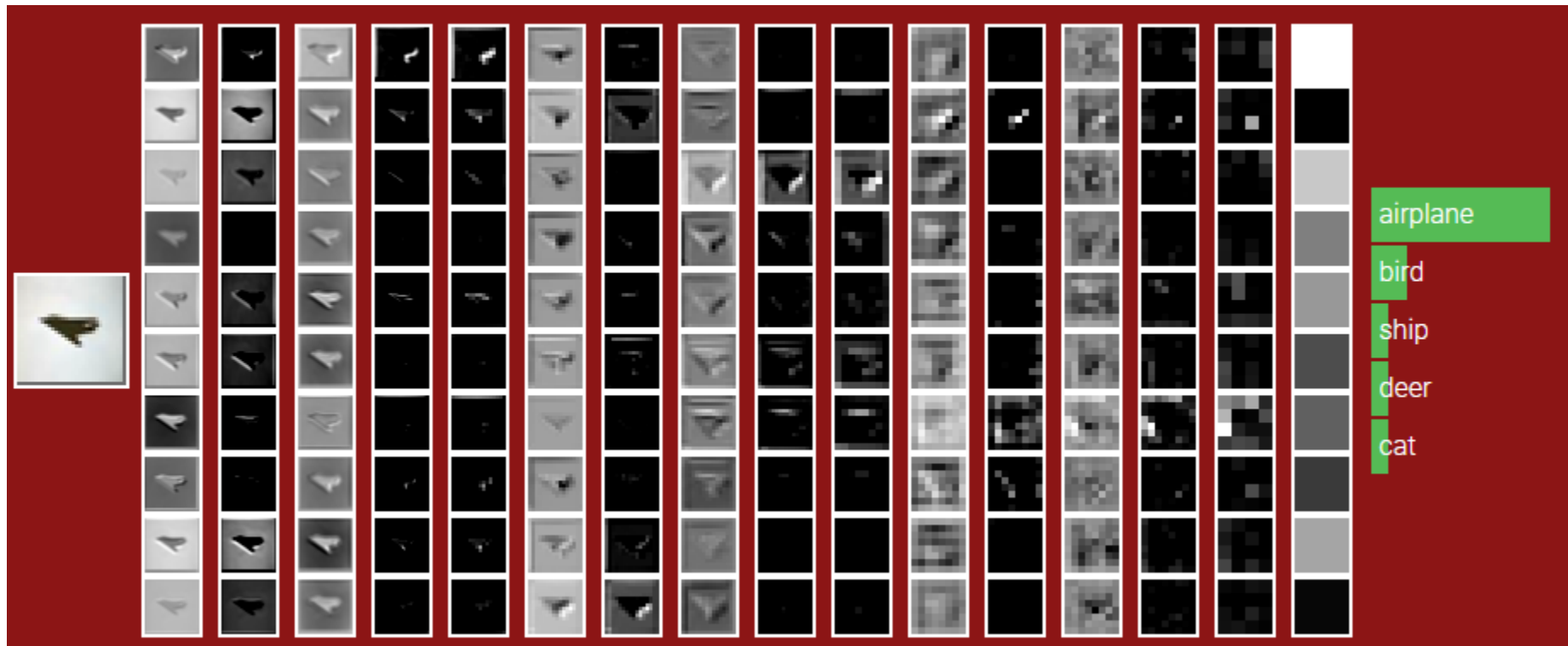
图片尺寸:  
440 × 262

查找该图片的其他尺寸  
全部尺寸 - 小尺寸 -

可能相关的搜索查询: 蔡徐坤 打篮球

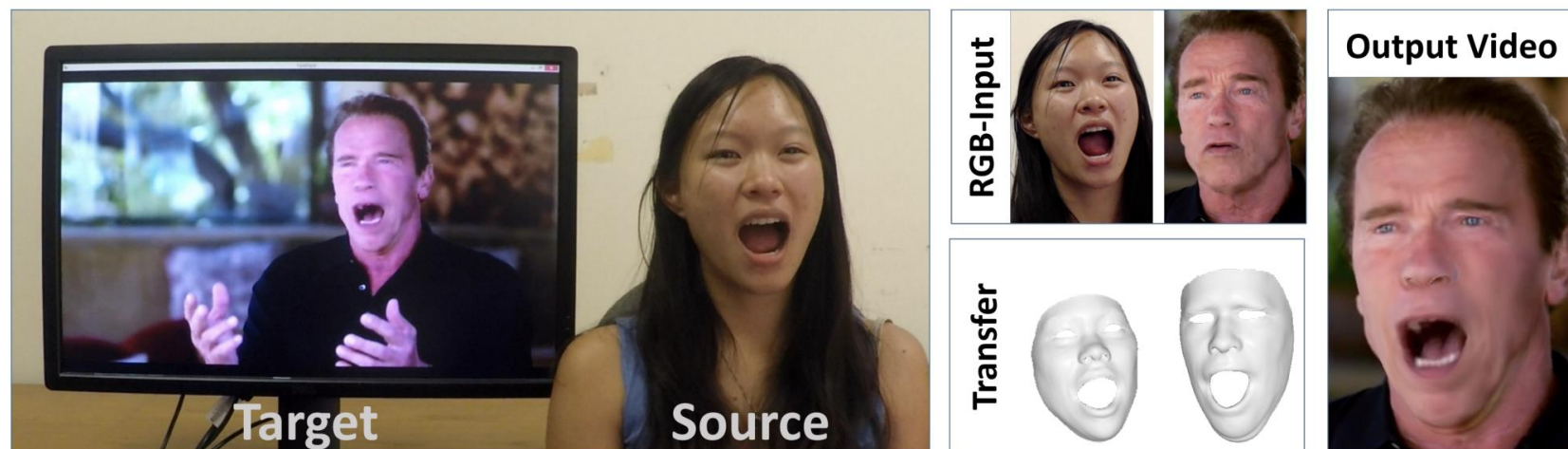


# 机器学习应用—图片分类





# 机器学习应用—换脸



学习算法利用人脸捕捉，让你在视频里实时扮演另一个人，简单来讲，就是可以把你的面部表情实时移植到视频里正在发表演讲的美国总统身上。



# 机器学习应用—翻译



翻译

关闭即时翻译

中文 英语 德语 检测语言



英语 中文(简体) 日语

翻译

Google神经机器翻译系统面世。该系统克服在大型数据集上工作的挑战，不再将句子分解为词和短语独立翻译，而是翻译完整的句子，使得误差降低了55%-85%以上。目前这项技术已经运用于Google Translate的汉英翻译。



112/5000

The Google Neuro Machine Translation System is available. The system overcomes the challenge of working on large data sets, no longer decomposing sentences into independent translations of words and phrases, but translating complete sentences, reducing errors by more than 55%-85%. This technology is currently used in Chinese-English translation of Google Translate.



提出修改建议

Google shénjīng jīqì fānyì xìtǒng miànshì. Gāi xìtǒng kèfú zài dàxíng shùjù jí shàng gōngzuò de tiǎozhàn, bù zài jiāng jùzǐ fēnjiē wéi cí hé duǎnyǔ dúlì fānyì, ér shì fānyì wánzhěng de jùzǐ, shǐdé wùchā jiàngdīle 55%-85%yǐshàng. Mùqián zhè xiàng jìshù yǐjīng yùnyòng yú Google Translate de hàn yīng fānyì.

文言文



中文

翻译



知之为知之，不知为不知，是知也。  
扁鹊见蔡桓公，立有间。  
长太息以掩涕兮，哀民生之多艰。



知道就知道，不知道就是知道，这才是聪明的。  
扁鹊见蔡桓公，站了一会儿。  
止不住的叹息擦不干的泪水啊，可怜人生道路多么艰难不顺利。



您输入的可能是：中文



拼音



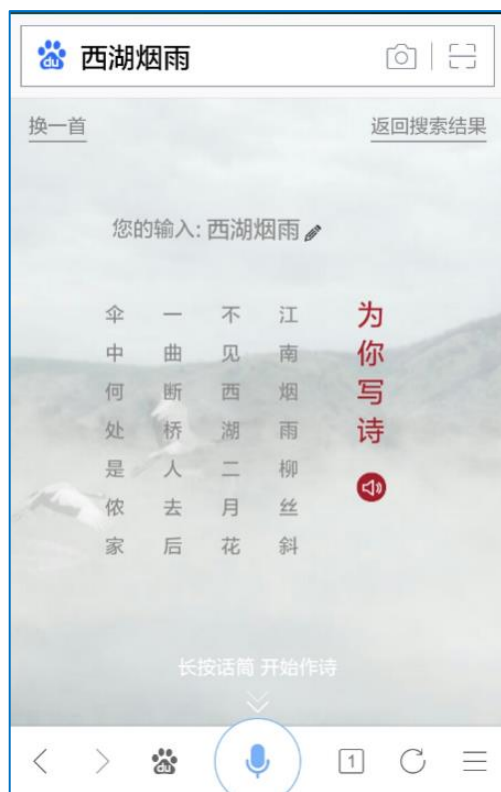
双语对照

# 机器学习应用—聊天机器人





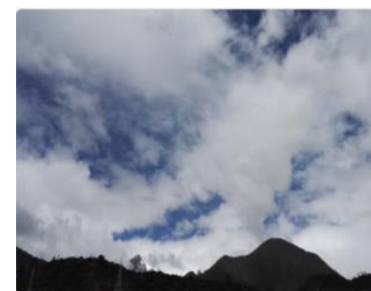
# 机器学习应用—写诗



语音作诗



藏头诗



天空云淡飞鸿远  
人间何处觅芳踪  
红尘难解相思意  
一曲清风月下逢

度秘写诗  
DUER

看图作诗

# 机器学习应用—作曲编曲

## 念奴娇·赤壁怀古

【作者】苏轼【朝代】宋

大江东去，浪淘尽，千古风流人物。故垒西边，人道是，三国周郎赤壁。乱石穿空，惊涛拍岸，卷起千堆雪。江山如画，一时多少豪杰。遥想公瑾当年，小乔初嫁了，雄姿英发。羽扇纶巾，谈笑间，檣櫓灰飞烟灭。故国神游，多情应笑我，早生华发。人生如梦，一尊还酹江月。



# 机器学习应用—下棋



AlphaGo以3：0战胜柯洁



# 机器学习应用—德州扑克

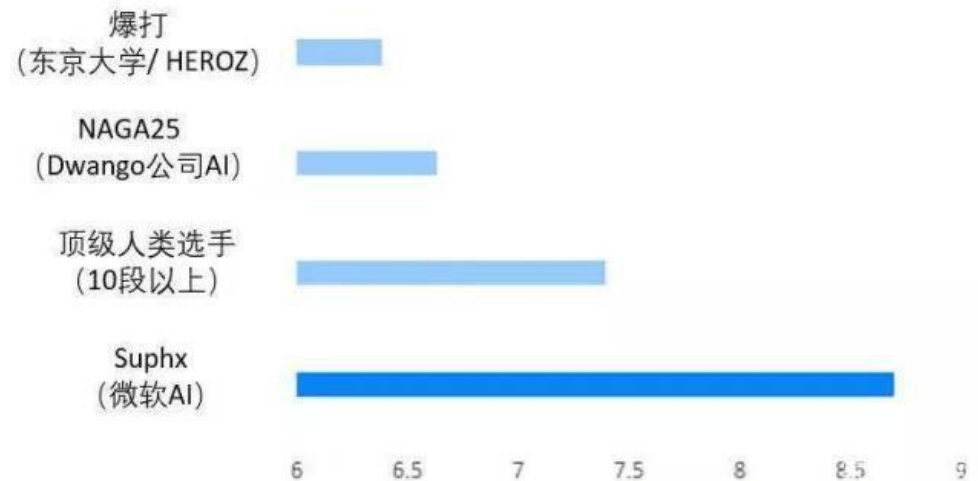


- 在无限限制德州扑克六人对决的比赛中，由 Facebook 与CMU共同开发的德扑 AI Pluribus 成功战胜了五名专家级人类玩家

# 机器学习应用—麻将



天凤平台“特上房”稳定段位对比

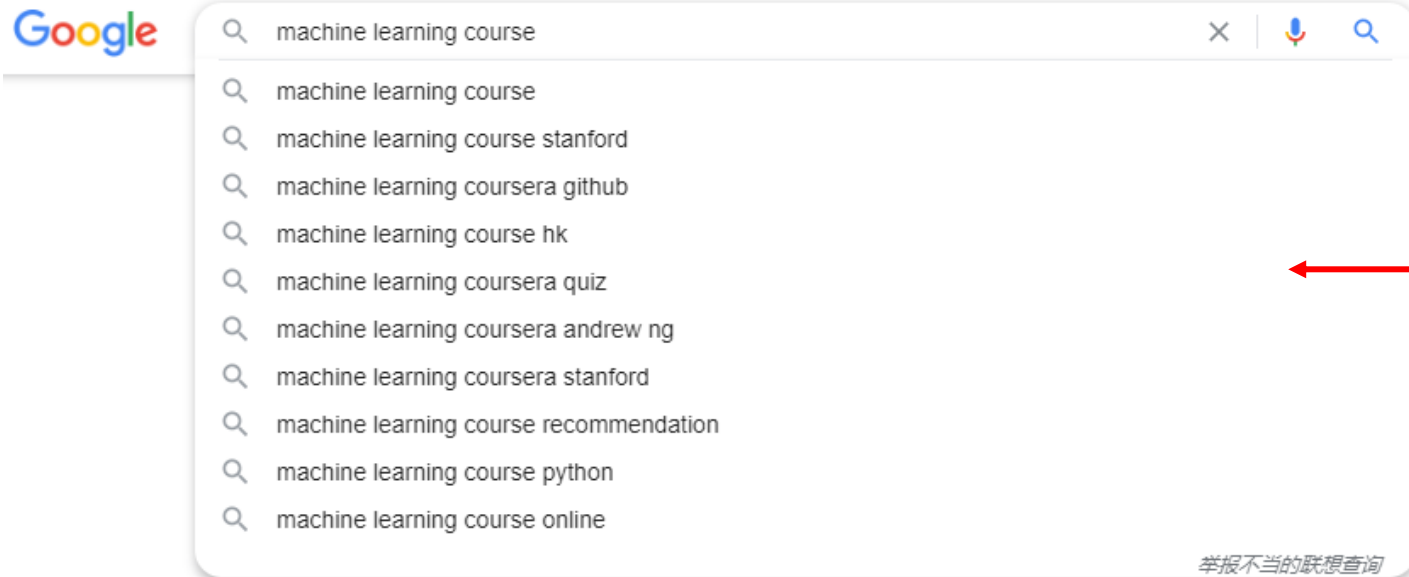


微软亚洲研究院研发的麻将AI Suphx，在国际知名麻将平台“天凤”上荣升十段，稳定段位显著超越人类顶级选手

# 游戏AI历史



# 机器学习应用—搜索



查询词建议

www.coursera.org > ... > Machine Learning ▾

**Machine Learning by Stanford University | Coursera**

About this **Course**. 9,957,819 recent views. **Machine learning** is the science of getting computers to act without being explicitly programmed. In the past decade, ...

[Machine learning and data ...](#) · [What is Machine Learning?](#) · [Unsupervised Learning](#)

相关性计算  
&PageRank



# 机器学习应用—推荐



## 机器学习

击败AlphaGo的武林秘籍，赢得人机大战的必由之路：人工智能大师周志华教授巨著，全面揭开机器学习的奥秘

周志华 著

京东价 **¥68.40** [7.8折] [定价 ¥88.00] (降价通知)

累计评价  
10万+

促销信息 **换购** 购买1件可优惠换购热销商品 立即换购 >>

**加价购** 满10元另加26.90元，或满12元另加16.90元，或满15元另加9.90元。

即可在购物车换购热销商品 详情 >>

以上促销可在购物车任选其一

增值业务  助力环保，传递知识，旧书换新

排 名 自营 计算机与互联网销量榜 第 4 位

配 送 至 安徽合肥市蜀山区笔架山街道 有货

由 京东 发货，并提供售后服务。23:10前下单，预计明天(09月14日)送达

重 量 0.92kg



❤ 关注  分享

举报

猜你喜欢

猜你喜欢

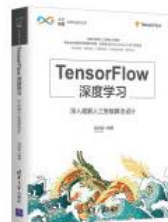
1/4



¥56.80



¥118.00



¥84.60



¥56.60



¥66.30

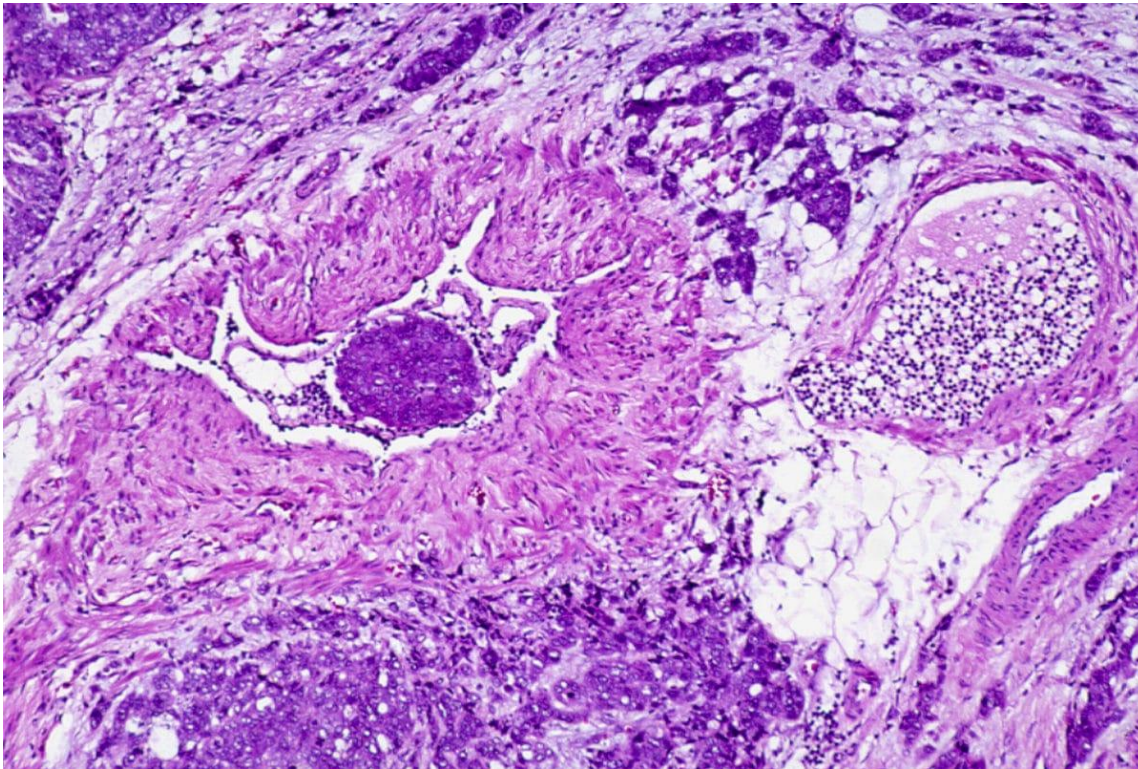


¥122.90

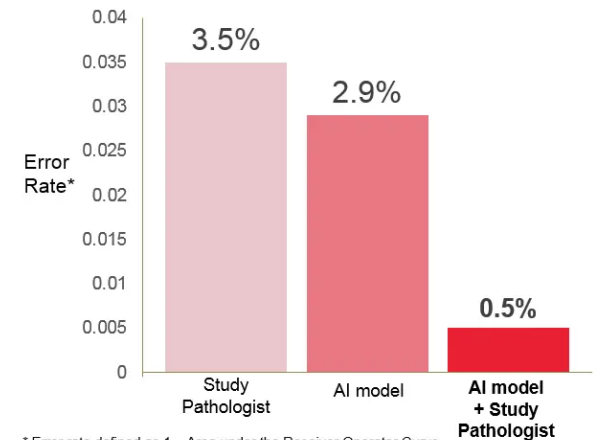


# 机器学习应用—疾病诊断

## • 乳腺癌诊断



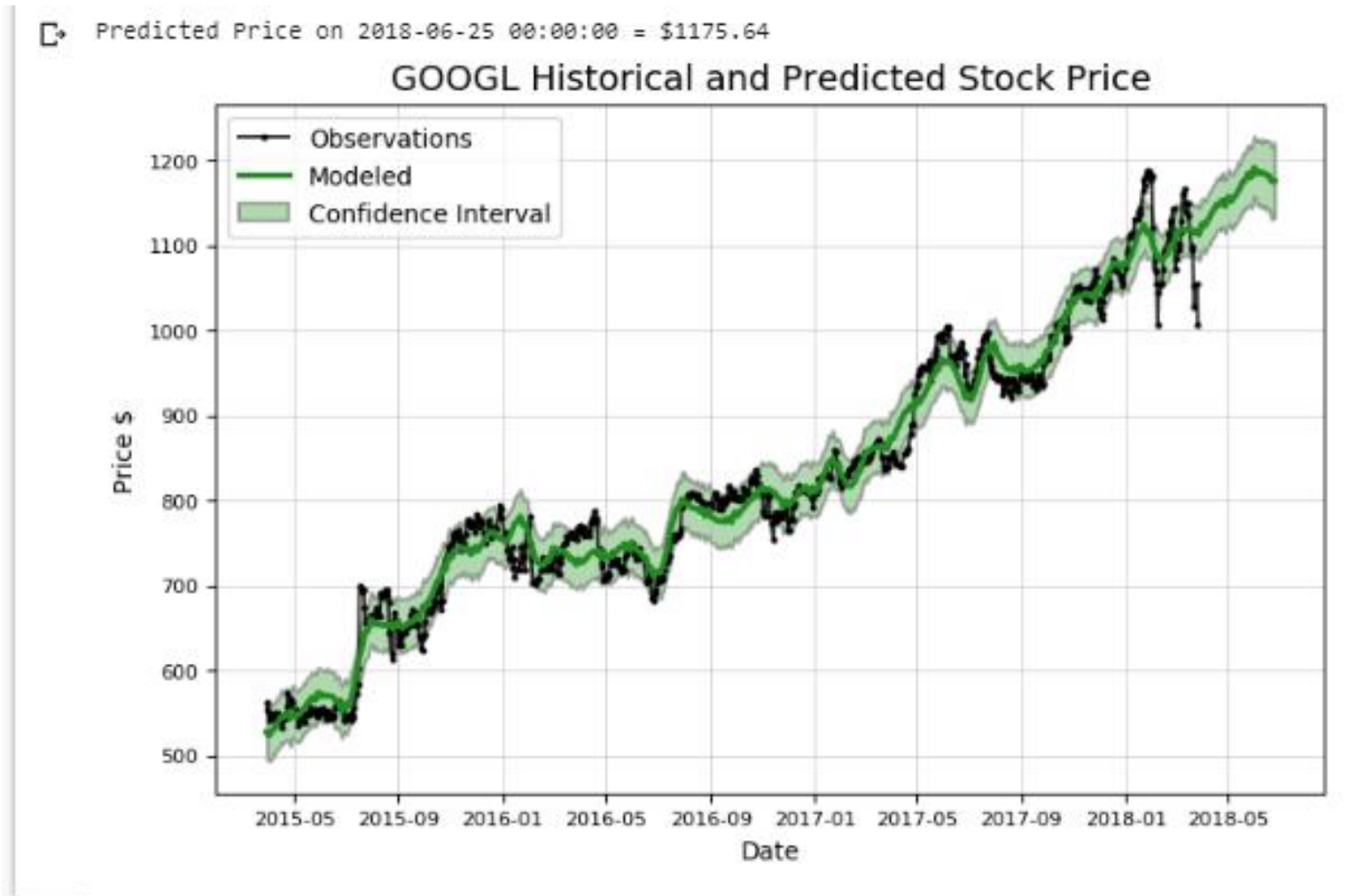
(AI + Pathologist) > Pathologist



\* Error rate defined as 1 - Area under the Receiver Operator Curve  
\*\* A study pathologist, blinded to the ground truth diagnoses, independently scored all evaluation slides.

© 2016 PathAI

# 机器学习应用—股价预测

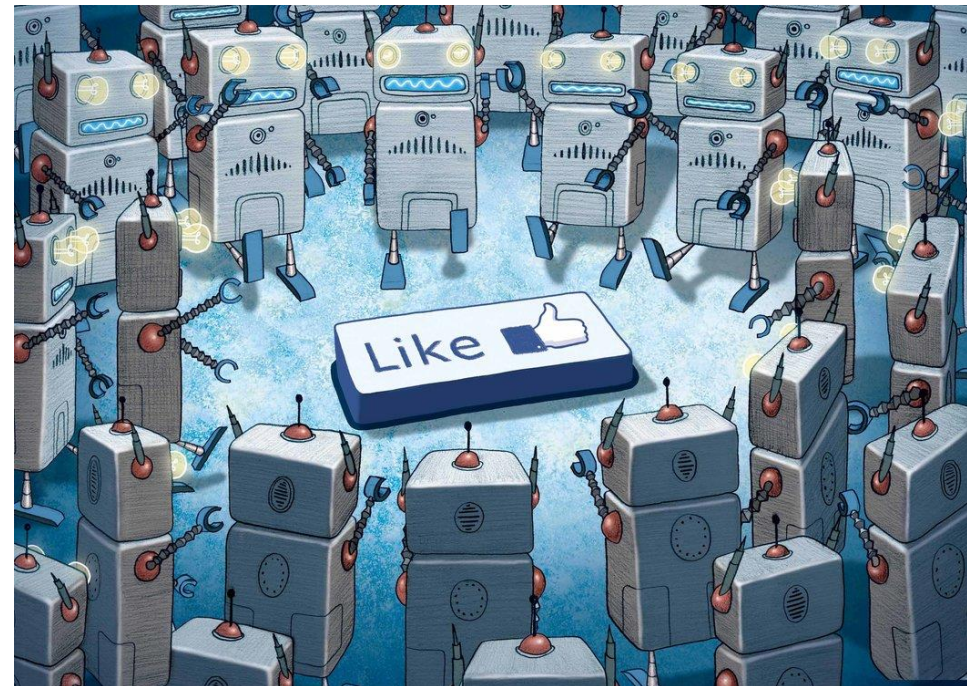




# 机器学习应用—异常检测

## • 社交网络中异常账号检测

<input type="checkbox"/> 货到付款 <input type="checkbox"/> 海外商品 <input type="checkbox"/> 二手 <input type="checkbox"/> 天猫 <input type="checkbox"/> 正品保障 <input type="checkbox"/> 旺旺在线			
	微博刷量转发评论点赞/直发@人1元1万话题阅读量讨论大号推广	¥1.00	356人付款 227条评论
	【皇冠信誉】微博营销/话题直发/微博转发/评论点赞/活动推广	¥0.50	807人付款 152条评论
	微博营销/话题直发/微博转发/微博点赞/微博评论活动推广参与	¥0.50	738人付款 180条评论
	微博转发包月评论点赞话题试听阅读量参与/1个QB1q币0.5元	¥0.50	598人付款 1条评论



# 机器学习应用—虚假新闻检测



提问较真



疫情数据



北京疫情

新冠肺炎男性患者的死亡率更高 **确实如此**

2020-09-10



新冠死者平均只损失了几个月的寿命 **谣言**

2020-09-09



# 机器学习应用—自动驾驶





# 机器学习应用—自动驾驶

优酷

# 机器学习的部分数学基础

- 矩阵
- 概率分布
- 优化



# 机器学习的部分数学基础—矩阵

- 矩阵为二维数组，用大写粗斜体表示  $A \in \mathbb{R}^{m \times n}$

- $A_{:,i}$  表示第*i*列
- $A_{i,:}$  表示第*i*行
- $A_{i,j}$  表示第*i*行第*j*列元素

- 矩阵范数

- 函数 $f(A)$ 衡量矩阵的大小，满足三个条件

- $f(A) \geq 0$ ，等号成立当且仅当 $A = 0$
- $f(\alpha A) = \alpha f(A)$
- $f(A + B) \leq f(A) + f(B)$

- Frobenius norm

$$\|A\|_F = \sqrt{\sum_i \sum_j |A_{i,j}|^2}$$



$$\|A\|_F^2 = \text{tr}(A^T A)$$

- p-诱导范数

$$\|A\|_p = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}$$

$$\|x\|_p = \sqrt[p]{x_1^p + \dots + x_d^p}$$



$$\|A\|_2 = \sigma_{\max}(A)$$

	$A_{11}$	$A_{12}$	$A_{13}$	行
	$A_{21}$	$A_{22}$	$A_{23}$	
	$A_{31}$	$A_{32}$	$A_{33}$	
列				



# 机器学习的部分数学基础—矩阵导数

$$\bullet (\nabla f(\mathbf{x}))_i = \frac{\partial f(\mathbf{x})}{\partial x_i} \quad (\nabla^2 f(\mathbf{x}))_{ij} = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} \quad (\text{海森矩阵})$$

$$\bullet \left( \frac{\partial x}{\partial \mathbf{A}} \right)_{ij} = \frac{\partial x}{\partial A_{ij}}$$

$$\bullet \frac{\partial \text{tr}(\mathbf{A}\mathbf{B})}{\partial A_{ij}} = B_{ji} \quad \longrightarrow \quad \frac{\partial \text{tr}(\mathbf{A}\mathbf{B})}{\partial \mathbf{A}} = \mathbf{B}^\top \quad \frac{\partial \text{tr}(\mathbf{A}^\top \mathbf{B})}{\partial \mathbf{A}} = \mathbf{B}$$

$$\bullet \frac{\partial \|\mathbf{A}\|_F^2}{\partial \mathbf{A}} = \frac{\partial \text{tr}(\mathbf{A}^\top \mathbf{A})}{\partial \mathbf{A}} = \mathbf{A} + \mathbf{A}^\top$$

# 机器学习的部分数学基础—矩阵导数

- $\left(\frac{\partial A}{\partial x}\right)_{ij} = \frac{\partial A_{ij}}{\partial x} \quad \left(\frac{\partial x}{\partial A}\right)_{ij} = \frac{\partial x}{\partial A_{ij}}$
- $A^{-1}A = I \Rightarrow \frac{\partial A^{-1}A}{\partial x} = \frac{\partial A^{-1}}{\partial x}A + A^{-1}\frac{\partial A}{\partial x} = 0 \Rightarrow \frac{\partial A^{-1}}{\partial x} = -A^{-1}\frac{\partial A}{\partial x}A^{-1}$
- 试计算  $\frac{\partial \det(A)}{\partial A} \Rightarrow \frac{\partial \det(A)}{\partial A} = \mathbf{C} = \text{adj}(A)^\top \Rightarrow \frac{\partial \ln \det(A)}{\partial A} = (A^{-1})^\top$   
伴随矩阵
- 回忆  $\det(A) = \sum_i A_{ij}C_{ij}$ ,  $C_{ij}$  表示方阵  $A$  关于  $A_{ij}$  的代数余子式
- 回忆矩阵逆的运算  $A^{-1} = \det(A)^{-1} \text{adj}(A)$
- 试计算  $\frac{\partial \ln \det(A)}{\partial x}$

# 机器学习的部分数学基础——特征值分解

- 对于方阵 $A$ , 特征向量方程
  - $Av = \lambda v$ ,  $v$  特征向量,  $\lambda$ 为特征值  $A(\alpha v) = \lambda(\alpha v)$ , 考虑单位特征向量
- 可对角化矩阵的特征值分解为
  - $A = V \text{diag}(\lambda) V^{-1}$ ,  $\lambda$ 对应特征值
  - $V$ 中的每一列为特征向量
- 机器学习算法常常涉及 实对称矩阵
  - 可对角化的
  - 特征值是实数, 特征值为正数的矩阵为**正定阵**, 非负的为**半正定矩**
  - $V$  为正交矩阵, 满足  $V^T V = V V^T = I$

# 机器学习的部分数学基础—奇异值分解

- 奇异值分解类似于特征值分解，但是对任意矩阵都成立
- 对于任意大小为  $m \times n$  的矩阵  $A$ ， $A^T A v = \lambda v$ 
  - 令  $A v = \sigma u$ ，那么  $A^T \sigma u = \lambda v$ ，分别左乘  $A$  得到  $AA^T u = \frac{\lambda}{\sigma} A v = \lambda u$

$u$  对应  $AA^T$  的特征值为  $\lambda$  特征向量

- 在  $A v = \sigma u$  两边分别乘以  $u$ ，那么  $u^T A v = \sigma$
  - 在  $A^T \sigma u = \lambda v$  两边分别乘以  $v$ ，那么  $v^T A^T u = \frac{\lambda}{\sigma}$
- $\sigma^2 = \lambda$

- 将  $A v = \sigma u$  写成矩阵形式为  $AV = U\Sigma$ 

$\Sigma = \text{diag}([\sigma_1, \dots, \sigma_r, 0, \dots, 0])$   
 $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r \geq 0$   
 $r = \text{rank}(A)$

- 由于  $V$  是正交矩阵，所以  $A = AVV^T = U\Sigma V^T$ 

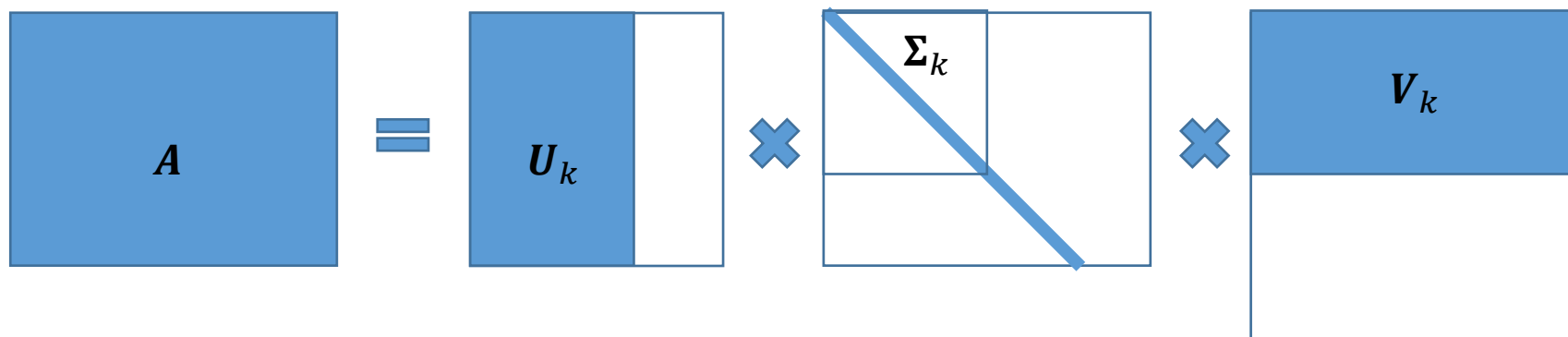
$\Sigma$  的大小为  $m \times n$   
 $U$  的大小为  $m \times m$   
 $V$  的大小为  $n \times n$

$U$  的列向量为左奇异向量， $V$  的列向量为右奇异向量

# 机器学习的部分数学基础—奇异值分解

- 截断奇异值分解

$$A \approx U_k \Sigma_k V_k^T$$



$U_k$  的大小为  $m \times k$

$\Sigma_k$  的大小为  $k \times k$

$V_k$  的大小为  $n \times k$

$$U_k^T U_k = I_k$$

$$\Sigma_k = \text{diag}([\sigma_1, \dots, \sigma_k])$$

$$V_k^T V_k = I_k$$

# 机器学习的部分数学基础—概率分布

## • 高斯分布、正态分布

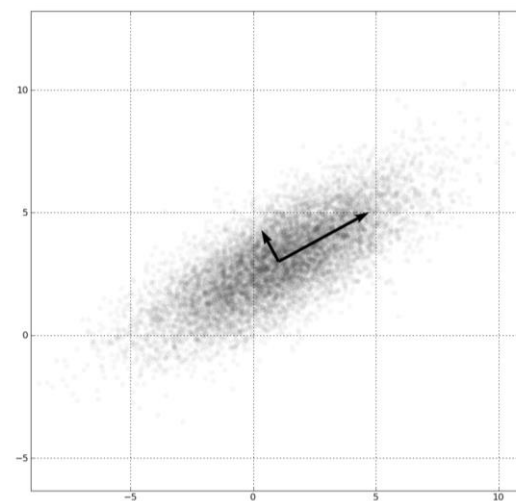
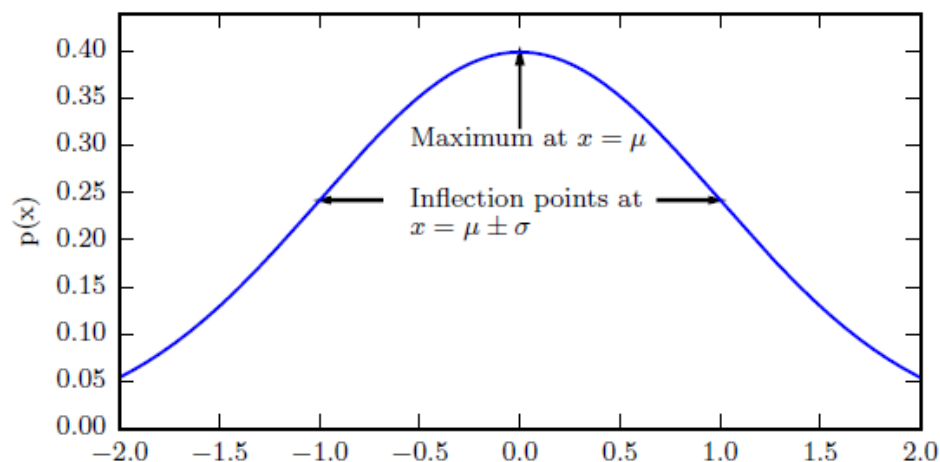
$$\bullet \mathcal{N}(x; \mu, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) \quad \text{均值}\mu \text{ 标准差}\sigma$$

$$\bullet \mathcal{N}(x; \mu, \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{\beta}{2}(x - \mu)^2\right) \quad \text{均值}\mu \text{ Scale } \beta$$

## • 多元正态分布

$$\bullet \mathcal{N}(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right)$$

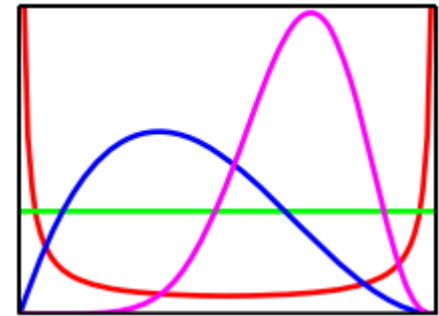
$$\bullet \mathcal{N}(x; \mu, \beta^{-1}) = \sqrt{\frac{\det(\beta)}{(2\pi)^n}} \exp\left(-\frac{1}{2}(x - \mu)^\top \beta(x - \mu)\right)$$



# 机器学习的部分数学基础—概率分布

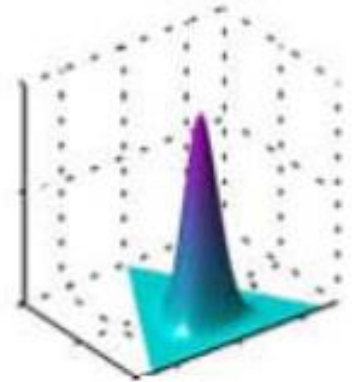
- 贝塔分布 (Beta Distribution) 是随机变量  $\mu \in [0,1]$  上的分布

- $Beta(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1 - \mu)^{b-1}$
- $\mathbb{E}[\mu] = \frac{a}{a+b}$
- $var[\mu] = \frac{ab}{(a+b)^2(a+b+1)}$



- 狄利克雷分布 (Dirichlet Distribution) 是 贝塔分布的多元扩展

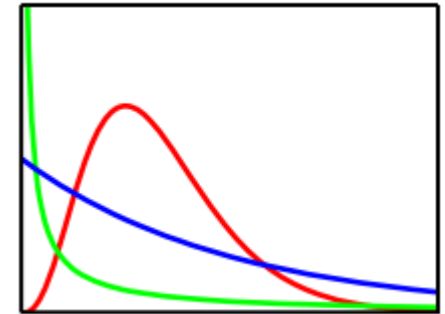
- 多个连续变量  $\mu_i \in [0,1]$  的概率分布, 满足  $\sum_i \mu_i = 1$
- $Dir(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i \mu_i^{\alpha_i-1}$
- $\mathbb{E}[\mu_i] = \frac{\alpha_i}{\sum_i \alpha_i}$



# 机器学习的部分数学基础—概率分布

- 伽玛分布 (Gamma Distribution) 是正数随机变量  $\tau > 0$  上的分布

- $Gam(\tau|a, b) = \frac{1}{\Gamma(a)} b^a \tau^{a-1} e^{-b\tau}$
- $\mathbb{E}[\tau] = \frac{a}{b}$
- $var[\tau] = \frac{a}{b^2}$





# 机器学习的部分数学基础—KL散度

- KL散度：衡量两个分布的差异

- $D_{KL}(P||Q) = \mathbb{E}_{x \sim P} \left[ \log \frac{P(x)}{Q(x)} \right]$

- 非负,  $P=Q$ 时为零

- $D_{KL}(P||Q) \neq D_{KL}(Q||P)$ , 但理论上最小值均当 $P=Q$

- $D_{KL}(P||Q) = \mathbb{E}_{x \sim P} \left[ \log \frac{P(x)}{Q(x)} \right] = \mathbb{E}_{x \sim P} [\log P(x)] - \mathbb{E}_{x \sim P} [\log Q(x)]$

$-H(P)$

P的熵

$H(P, Q)$

P和Q的交叉熵

# 机器学习的部分数学基础—优化

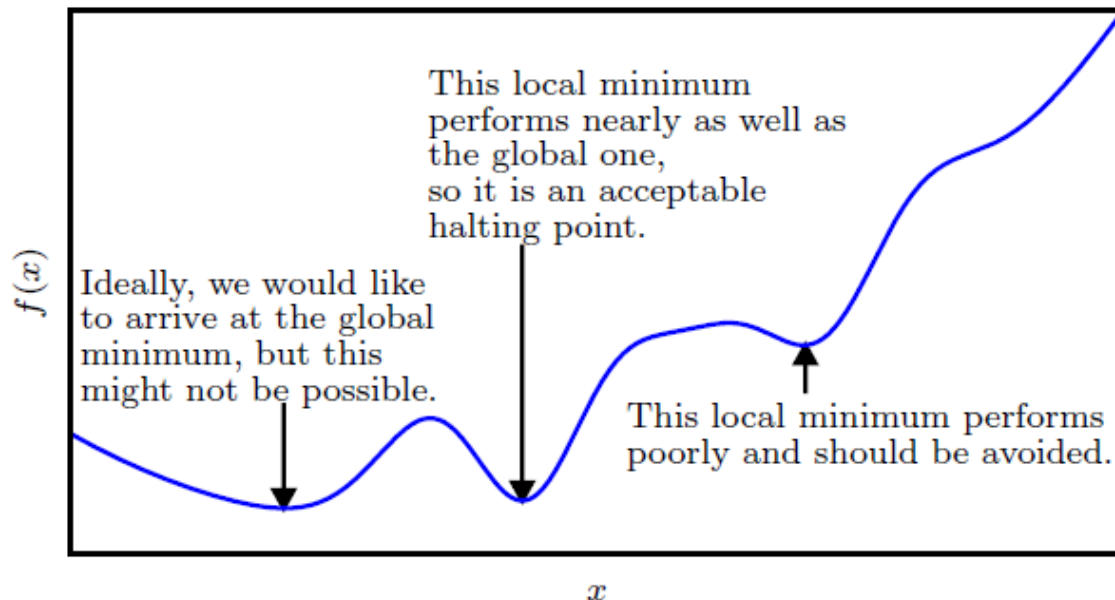
- 优化是什么？

求解目标函数在约束内的最小值

$$\begin{aligned} & \min_x f(x) \\ \text{s.t. } & g_i(x) \leq 0, i = 1, 2, \dots, m \\ & h_j(x) = 0, j = 1, 2, \dots, n \end{aligned}$$

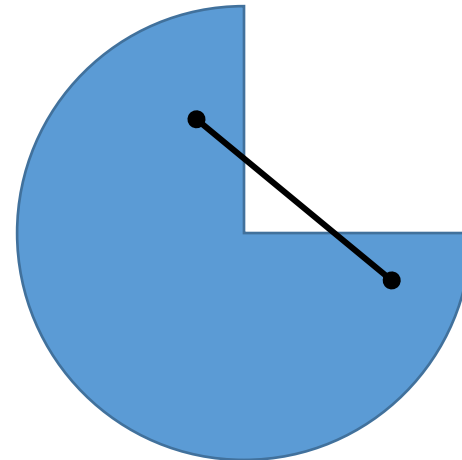
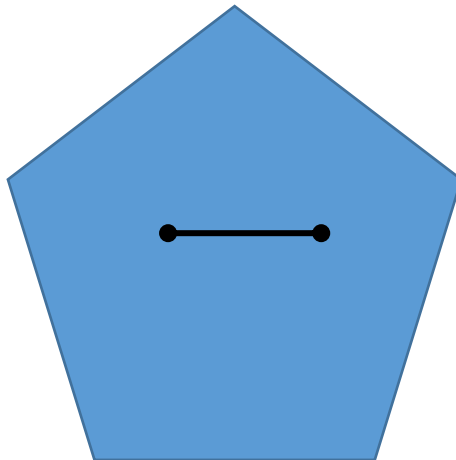
# 机器学习的部分数学基础—优化

- 临界点 (critical point)、驻点 (stationary point)
  - $f'(x) = 0$
  - 可能是局部最小点、局部最大点、鞍点
- 局部最小点  $x$ 
  - 对于 $x$ 的 $\epsilon$ 邻域上任意的 $c$ ,  $|x - c| < \epsilon$ ,  $f(x) < f(c)$
- 局部最大点  $x$ 
  - 对于 $x$ 的 $\epsilon$ 邻域上任意的 $c$ ,  $|x - c| < \epsilon$ ,  $f(x) > f(c)$



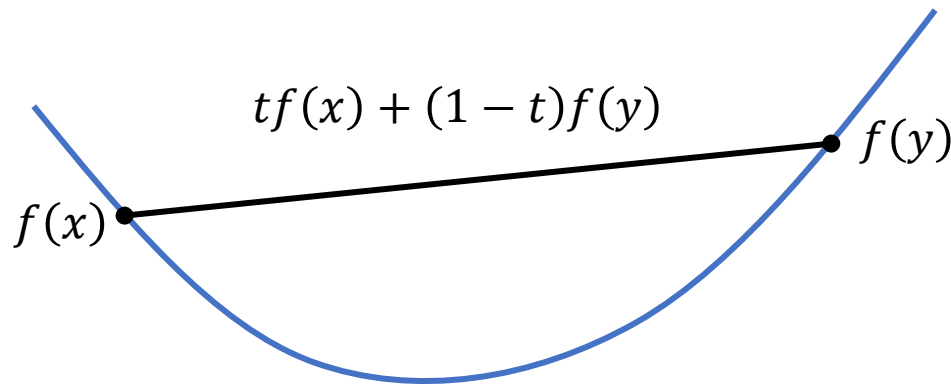
# 机器学习的部分数学基础—凸函数

给定集合  $C \subseteq \mathbb{R}^n$ 。若  $\forall x, y \in C$  满足  
 $\forall t \in [0, 1], tx + (1 - t)y \in C$   
那么集合  $C$  为凸集



# 机器学习的部分数学基础—凸函数

给定一个函数  $f: \mathbb{R}^n \mapsto \mathbb{R}$ 。如果满足  $\text{dom}(f)$  是凸集而且  $\forall x, y \in \text{dom}(f)$ ,  
 $\forall t \in [0, 1], f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y)$   
那么函数  $f$  是凸函数



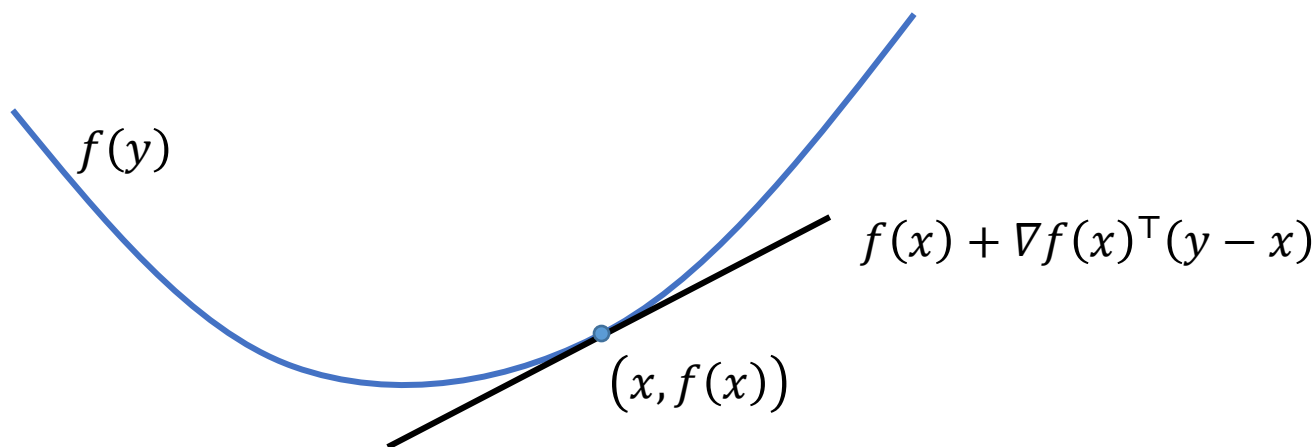
# 机器学习的部分数学基础—凸函数

- 指数函数  $\exp(ax)$
- 负对数函数  $-\log(x)$
- 反射函数  $\mathbf{a}^\top \mathbf{x} + b$
- 二次函数  $\mathbf{x}^\top \mathbf{A} \mathbf{x} + 2\mathbf{b}^\top \mathbf{x} + c$  ( $\mathbf{A}$ 半正定)
- 范数  $\|\mathbf{x}\|_p = \sqrt[p]{\sum_i |x_i|^p}$
- 最大函数  $f(\mathbf{x}) = \max\{x_1, \dots, x_n\}$
- Softplus  $\log(1 + \exp(x))$
- LogSumExp  $\log(\sum_i \exp(x_i))$
- LogDeterminant  $-\log \det(\mathbf{X})$ 在半正定矩阵定义域上

# 机器学习的部分数学基础—凸函数

## 一阶条件

假设函数 $f$ 可微, 那么 $f$ 是凸函数当且仅当  $\forall x, y \in \text{dom}(f)$ ,  
$$f(y) \geq f(x) + \nabla f(x)^\top (y - x)$$



# 机器学习的部分数学基础—凸优化

## 二阶条件

假设函数  $f$  二阶可微, 那么  $f$  是凸函数当且仅当  $\forall x \in \text{dom}(f)$ ,  $\nabla^2 f(x) \succeq 0$ , 即海森矩阵半正定



# 机器学习的部分数学基础—凸优化

- 凸优化问题

$$\begin{aligned} & \min_x f(\mathbf{x}) \\ \text{s.t. } & g_i(\mathbf{x}) \leq 0, i = 1, 2, \dots, m \\ & \mathbf{a}_i^\top \mathbf{x} = b, j = 1, 2, \dots, n \end{aligned}$$

其中 $f(\mathbf{x}), g_i(\mathbf{x})$ 是凸函数

# 机器学习的部分数学基础—凸优化

- 凸优化中，局部最优等价于全局最优

假设函数 $f$ 可微凸函数，那么 $x$  是  $f$  的全局最优当且仅当，  
$$\nabla f(x) = 0$$

证明：因为 $\nabla f(x) = 0$ . 所以

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) = f(x)$$

# 机器学习的部分数学基础—优化算法

- 无约束优化：梯度下降

- 目标：  $\min_x f(x)$

```
while  $\|\nabla f(x_t)\| > \delta$  do  
     $x_{t+1} \leftarrow x_t - \alpha \nabla f(x_t)$   
end while
```

# 机器学习的部分数学基础—优化算法

- 有约束优化：拉格朗日乘子法

$$\begin{aligned} & \min_{\mathbf{x}} f(\mathbf{x}) \\ \text{s.t. } & g_i(\mathbf{x}) \leq 0, i = 1, 2, \dots, m \\ & h_j(\mathbf{x}) = 0, j = 1, 2, \dots, n \end{aligned}$$

- 引入拉格朗日函数

$$L(\mathbf{x}, \mathbf{u}, \mathbf{v}) = f(\mathbf{x}) + \sum_i u_i g_i(\mathbf{x}) + \sum_j v_j h_j(\mathbf{x})$$

其中  $u_i \geq 0$

# 机器学习的部分数学基础—优化算法

拉格朗日函数

$$L(\mathbf{x}, \mathbf{u}, \mathbf{v}) = f(\mathbf{x}) + \sum_i u_i g_i(\mathbf{x}) + \sum_j v_j h_j(\mathbf{x})$$

其中  $u_i \geq 0$

• 有如下结论

$$\forall \mathbf{u} \geq 0, \mathbf{v}, \text{ 和可行解 } \mathbf{x}, \text{ 满足} \\ L(\mathbf{x}, \mathbf{u}, \mathbf{v}) \leq f(\mathbf{x})$$

# 机器学习的部分数学基础—优化算法

原问题

$$\begin{aligned} & \min_x f(\mathbf{x}) \\ \text{s.t. } & g_i(\mathbf{x}) \leq 0, i = 1, 2, \dots, m \\ & h_j(\mathbf{x}) = 0, j = 1, 2, \dots, n \end{aligned}$$

对偶问题  
凸优化

$$\begin{aligned} & \max_{\mathbf{u}, \mathbf{v}} g(\mathbf{u}, \mathbf{v}) \\ \text{s.t. } & \mathbf{u} \geq 0 \end{aligned}$$

$$\text{其中 } g(\mathbf{u}, \mathbf{v}) = \min_x L(\mathbf{x}, \mathbf{u}, \mathbf{v})$$

# 机器学习的部分数学基础—优化算法

- 假设可行解集为 $C$ ,  $f^*$ 为原问题最优解, 那么满足

$$f^* \geq \min_{x \in C} L(x, u, v) \geq \min_x L(x, u, v) = g(u, v)$$

- 进一步得到弱对偶性

$$f^* \geq g^* = \max_{u, v} g(u, v)$$

# 机器学习的部分数学基础—优化算法

- 强对偶性

$$f^* = g^* = \max_{\mathbf{u}, \mathbf{v}} g(\mathbf{u}, \mathbf{v})$$

## Slater条件

原问题为凸优化问题，且可行域中至少有一个点使得不等式约束严格成立



# 机器学习的部分数学基础—优化算法

- 最优解的必要条件：KKT条件

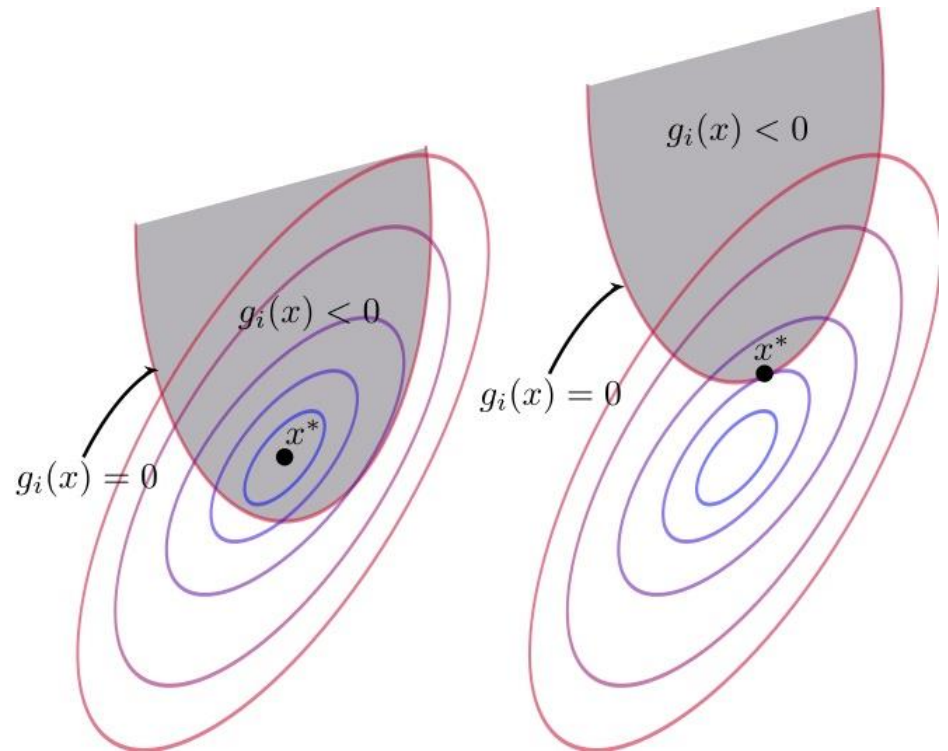
$$\nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{u}, \mathbf{v}) = \mathbf{0}$$

$$g_i(\mathbf{x}) \leq 0$$

$$h_j(\mathbf{x}) = 0$$

$$\mu_i \geq 0$$

$$\mu_i g_i(\mathbf{x}) = 0$$



# 作业

- 1. 计算  $\frac{\partial \ln \det(A)}{\partial x}$
- 2. 书习题1.2
- 3. 已知随机变量  $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2] \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , 计算  $P(\mathbf{x}_1), P(\mathbf{x}_1|\mathbf{x}_2)$
- 4. 证明范数  $\|\mathbf{x}\|_p$  是凸函数
- 5. 证明判定凸函数的0阶和1阶条件相互等价

$$\forall x, y \in \text{dom}(f), \forall t \in [0, 1], f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y)$$



$$\forall x, y \in \text{dom}(f), f(y) \geq f(x) + \nabla f(x)^\top (y - x)$$