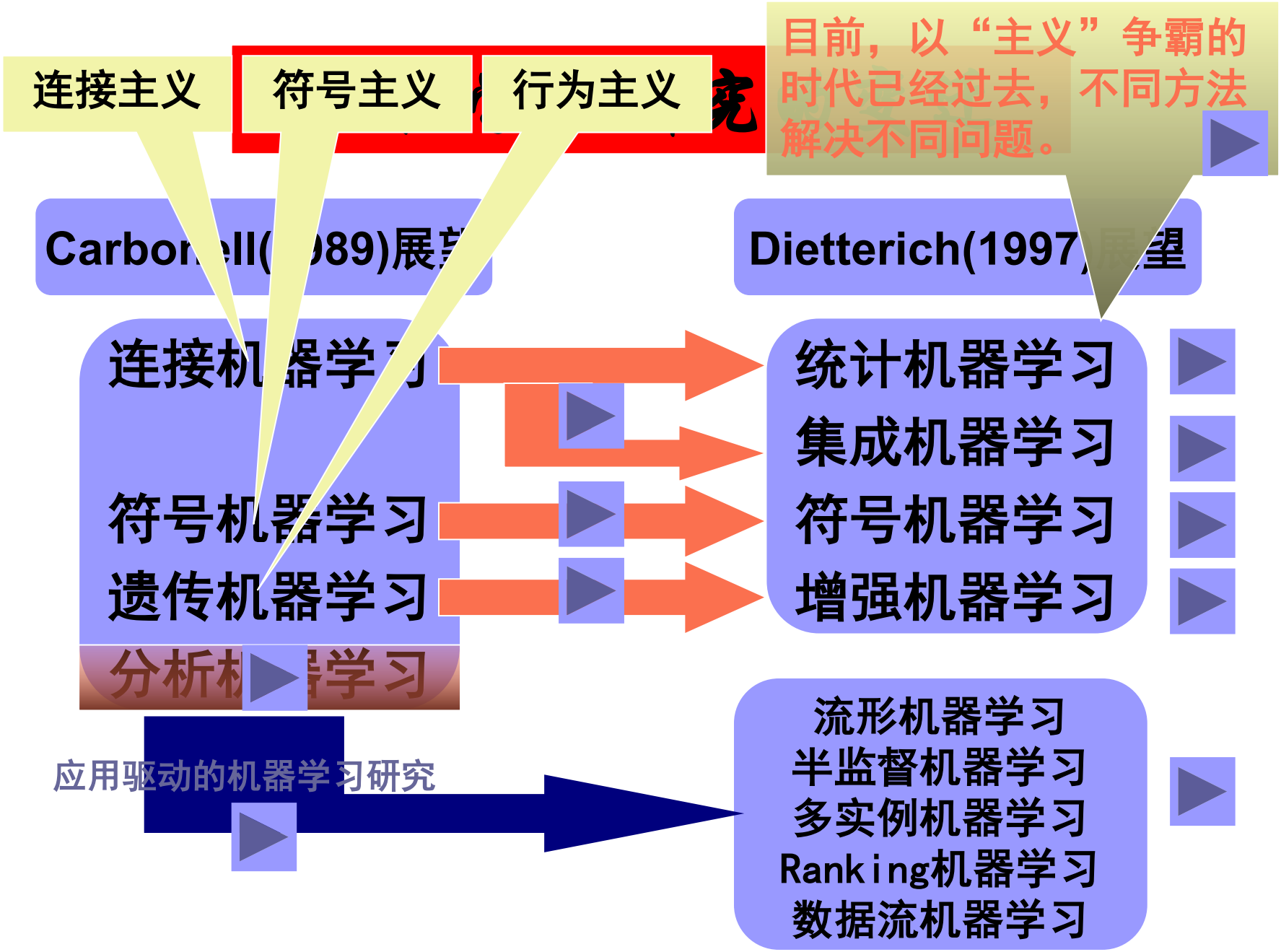
《机器学习》课件

机器学习研究进展

参考王珏老师的资料，表示感谢。


连接主义
符号主义
行为主义
目前，以“主义”争霸的时代已经过去，不同方法解决不同问题。
Carbonell(1989)展望
Dietterich(1997)展望
连接机器学习
符号机器学习
遗传机器学习
分析机器学习
统计机器学习
集成机器学习
符号机器学习
增强机器学习
应用驱动的机器学习研究
流形机器学习
半监督机器学习
多实例机器学习
Ranking机器学习
数据流机器学习

对统计机器学习的说明

- Dietterich将感知机类的连接机器学习分离出来，并根据划分机理，将其分为两种类型：统计机器学习与集成机器学习。这意味着，感知机类机器学习是重点
- 强调：

(1) 表示：非线性问题的线性表示

(2) 泛化：以泛化能力为基础的算法设计

对增强机器学习的说明

- “适应性”是控制理论中最重要的概念之一，以往在计算机科学中考虑较少
- 1975年，Holland首先将这个概念引入计算机科学。1990年左右，MIT的Sutton等青年计算机科学家，结合动态规划等问题，统称其为增强机器学习
- 这样，遗传学习成为实现增强机器学习的一种方法

对符号机器学习的说明

- 尽管经过十年，符号机器学习被保留，然而，其目标和内涵已发生很大的变化
- 改变泛化目标为符号描述(数据挖掘)。这意味着，符号机器学习已不是与统计机器学习竞争的研究，而是一个研究目标与其不同的研究范式

分析机器学习被放弃

- 分析机器学习所包含的类比、解释等问题对背景知识有更高的要求，这从表示到学习均需要考虑新的理论基础，在这些理论未出现之前，其淡出机器学习研究的视野是自然的

近几年的发展动向

- 由于真实世界的问题十分困难，现有的理论、方法，甚至理念已不能满足需要，由此，大量近代数学的研究结果被引入计算机科学，由此，形成新的机器学习范式

特点

- 从CarbonellI到Dietterich的特点是：

(1) 在算法设计理论上，基础代替随意的算法设计，具体地说，更为强调机器学习的数学基础

(2) 应用驱动代替理论驱动(认知科学与算法的Open问题)。具体地说，从AI中以“学习”机制驱动(智能)”的研究方式，改变为根据面临的实际问题发展新的理论与方法

统计机器学习的要点

- 目前，统计机器学习的研究主要集中在两个要点上：

表示问题

非线性问题
在线形空间的表示

泛化问题

对给定样本集合，
通过算法建立模型，
对问题世界为真的程度

线性表示

所以，手工构造的特征也有相当的作用。它辅助我们做转化

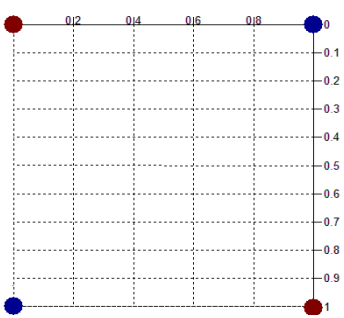
- 计算：非线性算法一般是NP完全的。
- 认识世界：只有在某个空间中可以描述为线性的世界，人们才说，这个世界已被认识(将问题变换为另一个问题)
- 数学方法：寻找一个映射，将非线性问题映射到线性空间，以便其可以线性表述

例子---XOR问题

特征工程的重要性

例子：XOR问题：

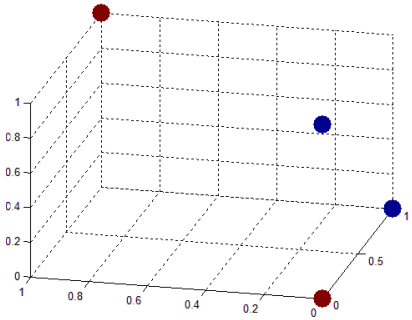
	x	y	d
1.	0	0	1
2.	0	1	0
3.	1	0	0
4.	1	1	1



映射：
(x, y)→(x, xy, y)

线形表示：

	x	y	xy	d
1.	0	0	0	1
2.	0	1	0	0
3.	1	0	0	0
4.	1	1	1	1



在机器学习中的方法

- 寻找具有一般意义的线性空间(方法)
- 目前，机器学习主要采用两种方法：
  (1) 整体线性，Hilbert空间(核映射)
  (2) 类似分段线性，Madaline或弱分类方法

Hilbert空间

- Hilbert空间是Von Neumman为量子力学数学基础提出的一类具有一般意义的线性内积空间
- 在机器学习中借助Hilbert空间构成特征空间

线性不可分机器学习问题

- 将线性不可分问题变为线性可分问题的关键是寻找一个映射，将样本集映射到特征空间，使其在特征空间线性可分
- 这样，我们只需以感知机为基础，研究统计机器学习问题。

困难—特征空间基的选择

- 选择特征空间的基
- 特征空间的基可以采用多项式基或三角函数基
- 寻找一般的方法描述特征空间存在根本性困难(维数灾)
- 与神经网络相比，核函数的选择可以借助领域知识，这是一个优点

理论描述

- 是否可以不显现地描述特征空间，将特征空间上描述变为样本空间上的描述？
- 如果不考虑维数问题，在泛函分析理论上，这是可行的
- 这就是核函数方法

泛化能力描述

	Duda(1973)	Vapnik(1971)
样本集：	样本个数趋近无穷大	有限样本，样本集内部结构(VC维)
泛化关系：	模型与泛化	随机选择样本集的随机变量 样本集、模型与泛化
泛化能力描述：	以概率为1成立	以概率1-δ成立
泛化不等式：	?(无法指导算法设计)	最大边缘(指导算法设计)

“泛化误差界”研究的演变

- PAC界(Valiant[1984])
- VC维界(Blumer[1989])
- 最大边缘(Shawe-Taylor[1998])

最大边缘(Shawe-Taylor[1998])

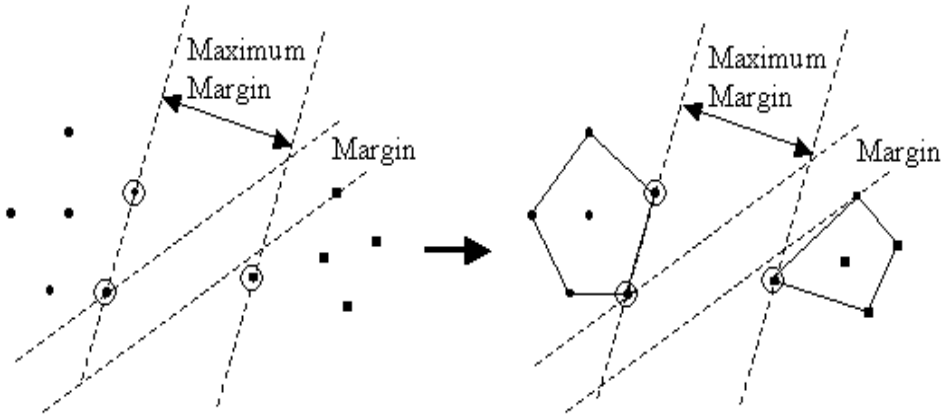
衡量分类器的性能的一个指标err(h)

这个不等式依赖于边缘M。贡献：给出了有几何直观的界描述，从而为算法设计奠定基础。

$$err(h) \le \sqrt{\frac{c}{l}\left(\frac{R^2}{M^2}\log^2 l - \log\delta\right)}$$

M>0，边缘不能等于零。这意味着，样本集合必须是可划分的。

边缘最大，误差界最小，泛化能力最强。泛化能力可以使用样本集合的边缘刻画。



研究趋势

- 算法的理论研究基本已经完成，根据特定需求的研究可能是必要的
- 目前主要集中在下述两个问题上：

(1) 泛化不等式需要样本集满足独立同分布，这个条件太严厉，可以放宽这个条件？

(2) 如何根据领域需求选择核函数，有基本原则吗？

集成机器学习的来源

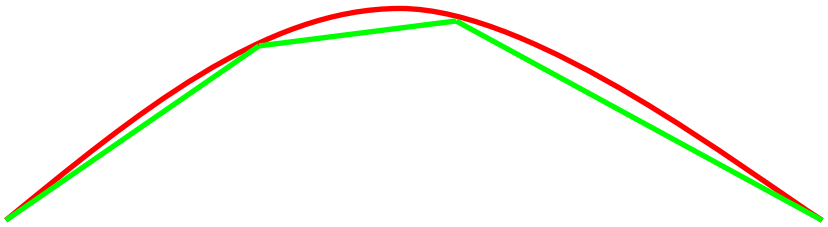
- 神经科学：Hebb神经细胞工作方式
- 数学方法：非线性问题的分段化(类似)
- 计算技术：Widrow的Madaline模型
- 统计理论：PAC的弱可学习理论

Ensemble(集成)

- 1954年，Hebb使用这个单词来说明视觉神经细胞的信息加工方式
- 假设信息加工是由神经集合体来完成

Madaline模型

- Widrow的Madaline模型
- 在数学上，其本质是放弃感知机对样本空间划分的超平面需要满足连续且光滑的条件，代之分段的超平面



Schapire的理论

1990年，Schapire证明了一个关键定理，由此，奠定了集成机器学习的理论基础

定理：如果一个概念是弱可学习的，充要条件是它是强可学习的

这个定理证明是构造性的，派生了弱分类器的概念，即，比随机猜想稍好的分类器

这个定理说明：
多个弱分类器可以集成为一个强分类器

问题

- 集成机器学习的研究还存在着大量未解决的问题，关于泛化能力的估计(不等式)还存在问题
- 目前，这类机器学习的理论研究主要是观察与积累，大量的现象还不能解释

符号机器学习

Gold证明，这是不可能的实现的(1967)。

最早的符号机器学习：

Solomonoff的文法归纳方法(1959)

符号机器学习的主流：

Samuel限制机器学习在结构化符号数据集合上(1967)，约简算法。

值得注意的动向：

文法归纳方法引起人们的重视。

Hebb路线：每个规则可以理解为一个弱分类器。

符号机器学习的数学基础

- 符号机器学习不同于统计机器学习，划分样本集合的等价关系是学习所得，符号机器学习是事先定义等价关系，学习只是在这个等价关系下约简样本集合
- 等价关系为：

$$\{(x, y) : a(x)=a(y), x, y \in U\}$$

符号机器学习的泛化问题

- 一个无矛盾规则越短，其覆盖对象越多，因此，符号机器学习的泛化是以信息长度描述的。这样，“最小”树或规则集合就是其目标函数
- 两个因素影响这个目标：其一，从实域到符号域的映射，其二，在符号域上的约简。对“最小”两者都是NP完全的。因此，近似算法是必然的
- 但是，只有在符号域上的约简是符号机器学习特有，因此，其泛化能力受到限制
- 不必与统计机器学习竞争，设立新目标

数据分析与传统机器学习区别

- 传统机器学习假设所有用户有相同的需求，其目标函数确定，而数据分析，不同用户有不同需求，目标函数随用户需求而定
- 传统机器学习是“黑箱”，模型无须可解释，但是，数据分析必须考虑对用户的可读性
- 传统机器学习将“例外”考虑为噪音，而数据分析则认为“例外”可能是更有意义的知识

符号机器学习的特点

- 由于这类机器学习主要处理符号，因此，如果获得一个长度较短的数据集合的描述，可以将其翻译为人可以阅读的文本。人通过阅读这个文本就可以了解数据集合的内容
- 这个目标与泛化能力无关，计算结果只是给定数据集合根据特定需求的一个可以被人阅读的缩影
- 这与传统数据分析的目标一致

符号数据分析(数据挖掘)

- 数据分析的主要工具是统计，“统计显现”是分析的主要指标
- 符号数据分析，尽管统计工具是必要的，但是，主要是通过将符号数据集合约简为简洁形式

符号机器学习的最新进展

- Rough sets中的reduct理论是近几年符号机器学习最重要的研究结果之一
- 这个理论理论可以作为符号机器学习的数学基础
- 这个理论可以作为符号数据分析的基础(数据挖掘)

Reduct与符号数据分析

- 在任务上，association rules派生于统计相关分析，其方法可以使用reduct理论来刻画
- Reduct具有很多重要的数学性质，可以保证根据不同需求识别不同的例外
- 我们建议，将符号数据分析建立在reduct理论之上

什么是outlier/exception

- 不能被模型(统计分布、规则集合等)概括的某些观察称为相对这个模型的outlier/exception
- 注释：

  在统计学中，这类观察称为outlier，在认知科学中，有意义的outlier称为exception

方法

- R是给定<U, C∪{d}>的reduct，使用R构造新信息系统<U, R∪{d}>，R是这个信息系统的Core
- 只要删除一个核属性，必然产生例外，同时缩短规则集

例外研究的意义

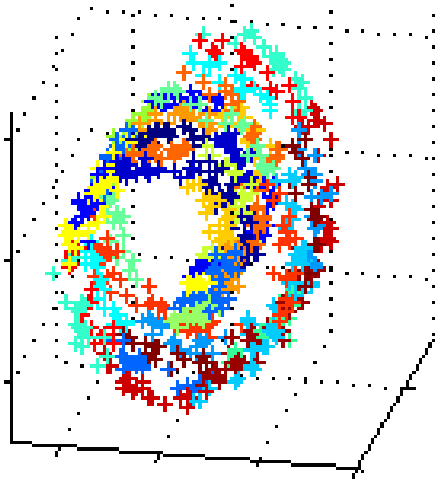
- 删除例外，可以使得规则更为简洁，从而突出信息的重点。例外是噪音
- 例外是比可以覆盖大多数样本的规则更为有趣的知识
- 对专家，规则是“老生常谈”，而例外则是应该引起注意的事件或知识
- 例外是新研究与发现的开始

增强机器学习

- 增强机器学习最早提出是考虑“从变化环境中”学习蕴含在环境中知识，其本质是对环境的适应
- 开始的动机主要是为了解决机器人规划、避障与在环境中适应的学习问题
- 目前，由于网络用户是更为复杂的环境，例如，如何使搜索引擎适应用户的需求，成为更为重要的应用领域

流形机器学习

- 很多问题的表示方法，使得信息十分稀疏，如何将信息稠密化是一个困难的问题（“维数灾难”），主成分分析是一种方法，但是，只对线性情况有效
- 流形学习是解决上述问题的非线性方法
- 由于流形的本质是分段线性化，因此，流形学习需要解决计算开集、设计同胚映射等问题

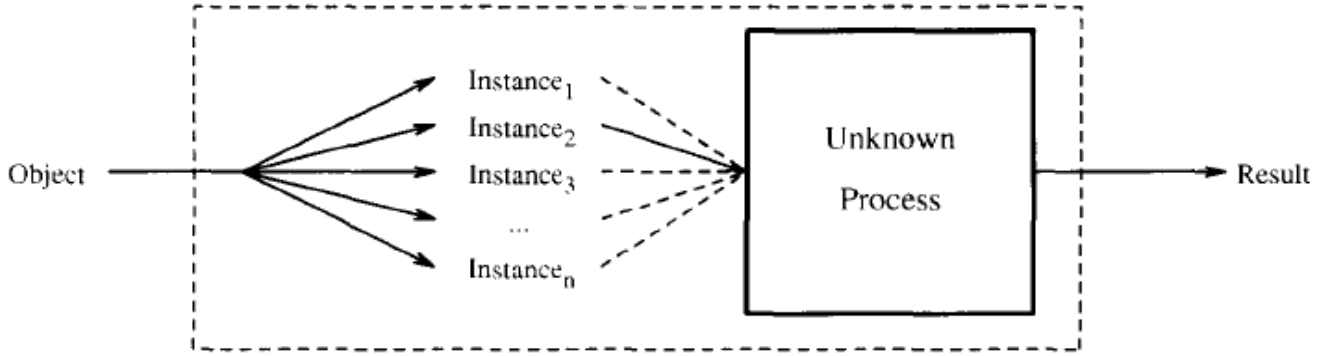


半监督机器学习

- 在观测数据中，可能有很多观测不能决定其类别标号。这需要根据数据中已知类别标号的样本与领域知识来推测这些样本的类别标号，并建立问题世界的模型，这就是半监督学习
- 这类问题直接来自于实际应用：例如，大量医学影像，医生把每张片子上的每个病灶都标出来再进行学习，是不可能的，能否只标一部分，并且还能利用未标的部分？

多示例机器学习

- 传统的机器学习中，一个对象有一个描述，而在一些实际问题中，一个对象可能同时有多个描述，到底哪个描述是决定对象性质(例如类别)的，却并不知道。解决这种“对象：描述：类别”之间1:N:1关系的学习就是多示例学习



Ranking机器学习

- 其原始说法是learning for ranking
- 问题主要来自信息检索，假设用户的需求不能简单地表示为“喜欢”或“不喜欢”，而需要将“喜欢”表示为一个顺序，问题是如何通过学习，获得关于这个“喜欢”顺序的模型。

数据流机器学习

- 在网络数据分析与处理中，有一类问题，从一个用户节点上流过的数据，大多数是无意义的，由于数据量极大，不能全部存储，因此，只能简单判断流过的文件是否有用，而无法细致分析
- 如何学习一个模型可以完成这个任务，同时可以增量学习，以保证可以从数据流中不断改善(或适应)用户需求的模型

研究现状

- 上述的五类机器学习范式还处于实验观察阶段，没有坚实的理论基础！
- 这些范式主要以任务为驱动力，大多数采用的方法是传统机器学习的方法
- 应用效果还不十分明显

总结

- 目前，我们所面临的问题是：数据复杂，需求多样。这要求：

(1) 需要考虑科学原理解决表示问题，特别需要借用近代数学的研究结果，“拍脑袋式”研究的时代可能已经过去

(2) 一种范式独步天下的时代已经过去，进入“多极世界” 时代

(3) 应用驱动成为必然，这意味着，任何方法需要在应用中检验

课程主要内容

- 常用不等式，概率可学习理论
- 增强学习，*ALphaGO*
- 归纳学习，决策树
- 统计机器学习，神经网络，支持向量机，深度学习