大数据与应用统计

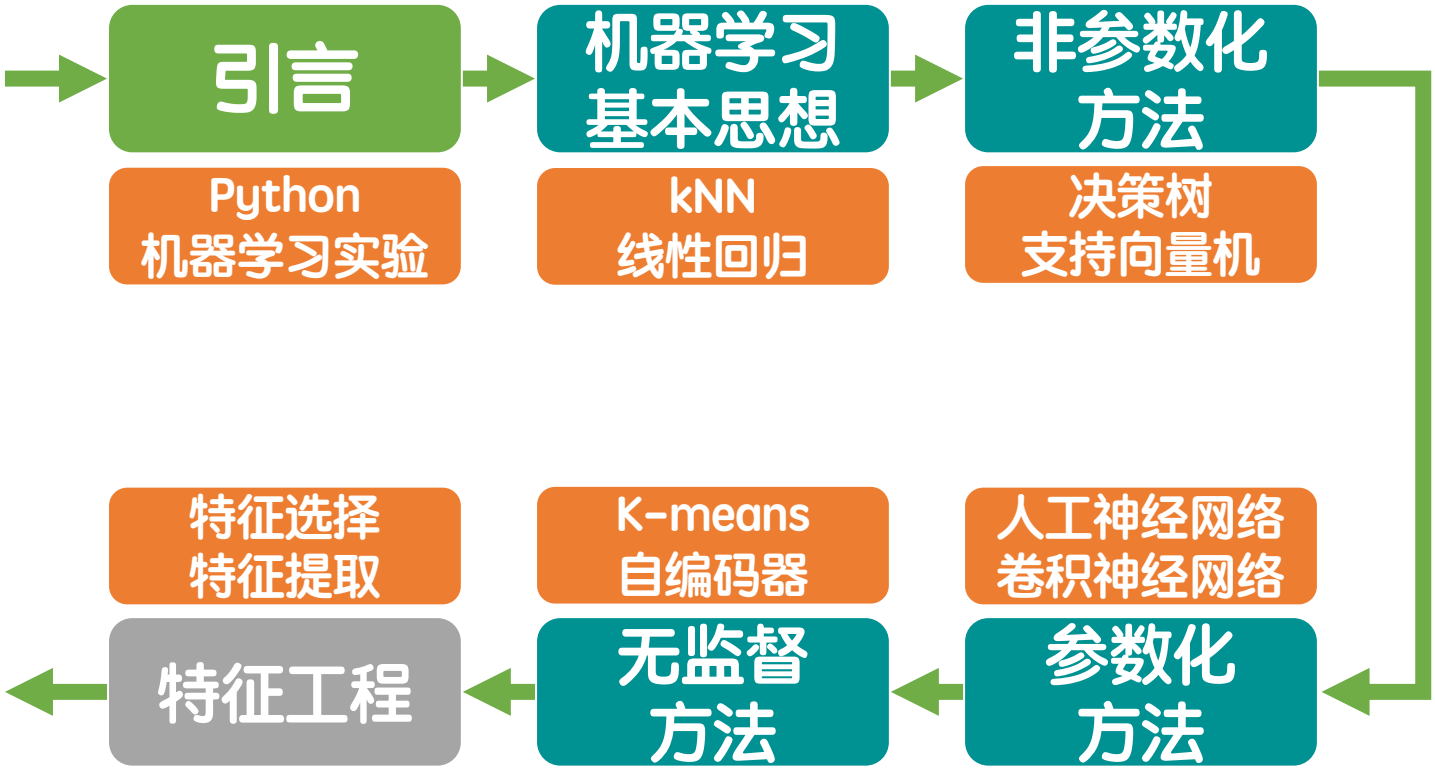
Big Data and Applied Statistics

北京工商大学

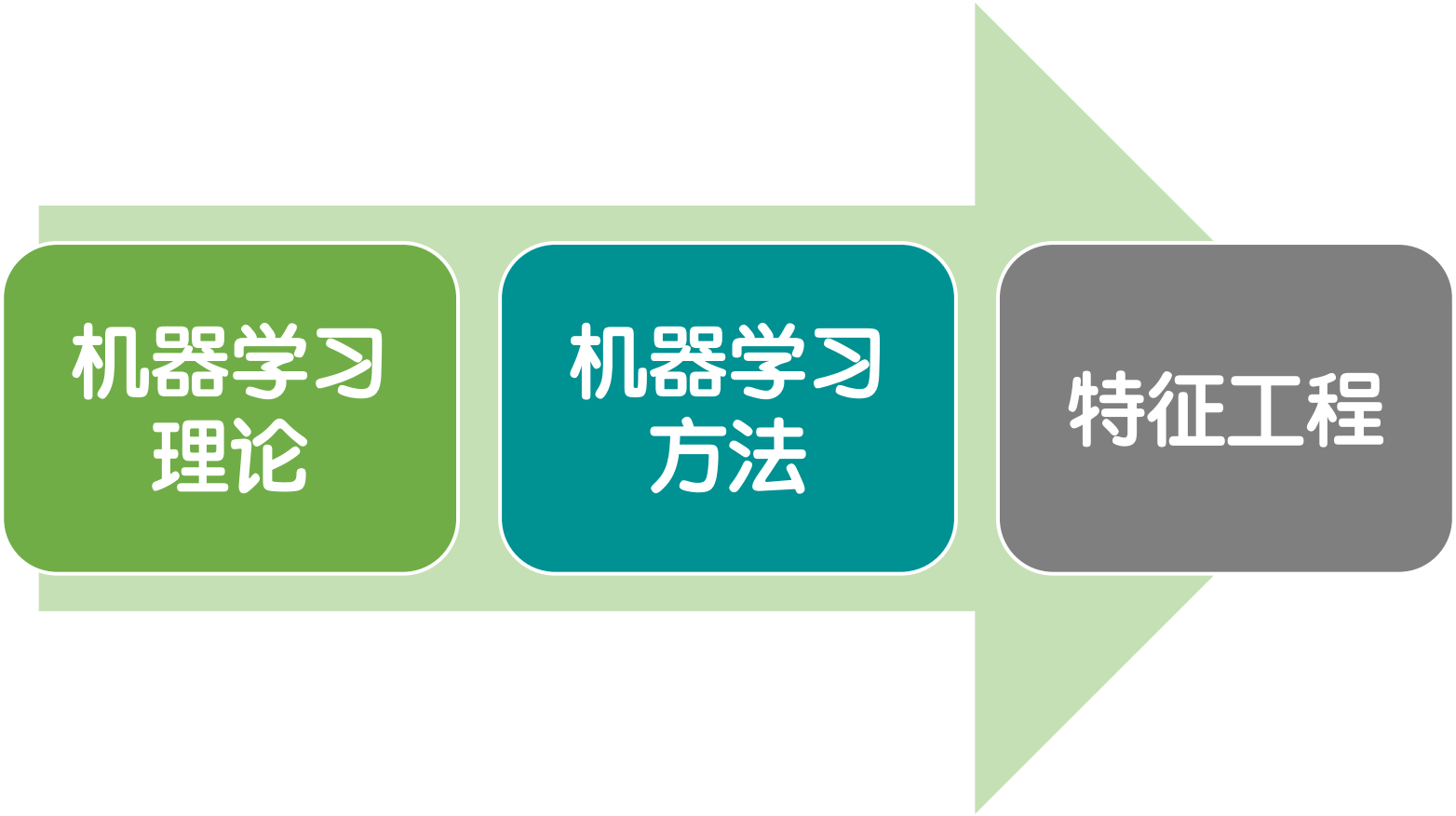
王晓川

课程说明

教学大纲


引言
机器学习
基本思想
非参数化
方法
Python
机器学习实验
kNN
线性回归
决策树
支持向量机
特征选择
特征提取
K-means
自编码器
人工神经网络
卷积神经网络
特征工程
无监督
方法
参数化
方法

课程体系



参考资料

- 课程来源
  - 主教材：Python机器学习：数据建模与分析，薛薇等
  - 其他引用的教材：
    - 机器学习方法，李航
    - 动手学机器学习，张伟楠等
  - 在线资源：

kaggle TIANCHI 天池

考核方式

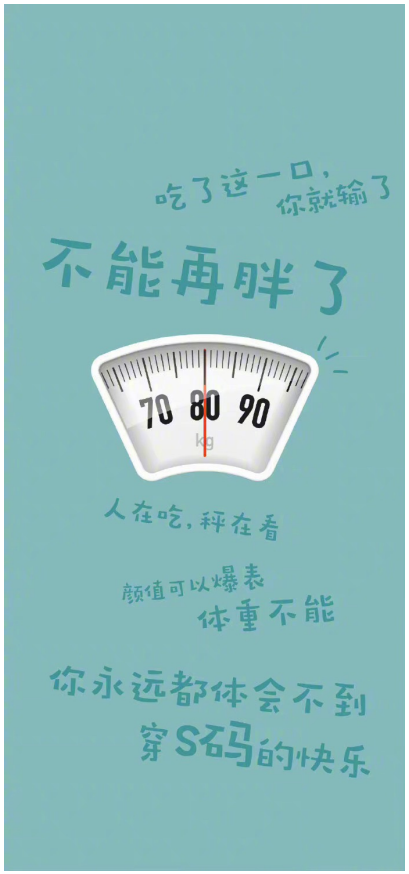
- 课后实验，占比40%。
  - 共6次作业，任选其中5次完成即可。
- 期末考试，占比60%。
- 联系方式：
  - E-mail：wangxc@btbu.edu.cn
  - 微信群

机器学习

什么是机器学习？

自律的一天：

- 10点半，准时起床。刷牙的时候忽然想听音乐，于是打开网易云，致郁的音乐瞬间填充你的大脑。
- 11点半，吃饭。感觉食堂吃得有点腻了，于是打开饿了么，发现第一页推荐的饭也是常吃的，刷了半天，也不知道要吃啥。
- 12点半，导师让你帮忙寄个快递，啰里八嗦说了一段。你懒得打字，先把这段语音转成文字，然后粘贴到顺丰app里，自动把收件人、联系方式、地址给填好了。
- 1点半，自律的一天结束了。睡觉前想打一把LOL，结果发现不知道为什么匹配的队友都很菜，连输几把以后决定打一局人机，瞬间找回感觉。
- ……

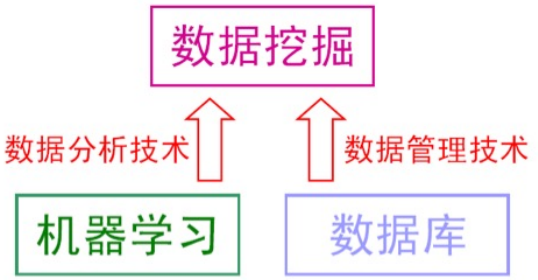


机器学习

机器学习，也叫统计机器学习，统计学习，是基于数据构建概率统计模型，并运用不同的模型对数据进行预测与分析的一门学科。

机器学习vs.数据挖掘

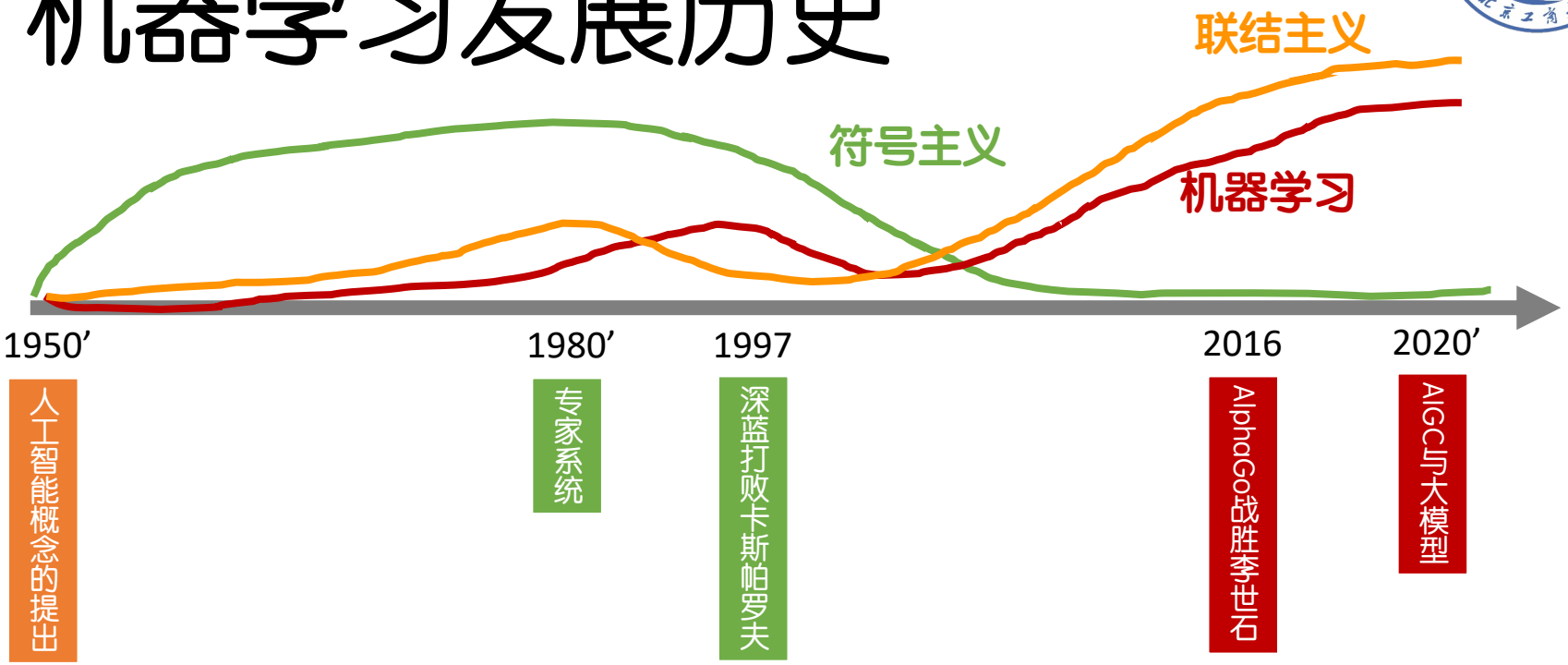
机器学习可以看作数据挖掘的工具之一。数据挖掘除利用机器学习之外，还需要解决一系列数据相关的问题，如数据清洗、数据存储、大数据处理等。



数据科学家vs.计算机科学家

- 数据中心论vs.方法中心论；
- 结果的合理vs.出结果的效率；
- 关注数据偏差的来源vs.防止数据偏差；
- 关注数据的含义vs.关注数据的精度。

机器学习发展历史



符号主义

规则
数据
→ 经典程序设计 → 答案

机器学习

数据
答案
→ 机器学习模型 → 规则

机器学习的核心：数据

可将机器学习看作一种新的编程范式，其核心任务是发现“数据”和“答案”背后的“规则”。

从数学角度看，就是发现隐藏在自变量和因变量之间的映射关系；其中，统计学习是发现隐藏关系的有力工具

数据集（样本集）

自变量（数据）

	姓名	物理攻击强度	法术攻击强度	控制技能	攻击距离	种属
1	奥拉夫	100	20	无	近距离	战士
2	武器大师	90	30	有	近距离	打野
3	阿卡丽	20	100	无	中距离	刺客
4	诡术妖姬	10	110	有	中距离	法师
5	璐璐	40	80	有	远距离	辅助
6	卢锡安	110	10	无	远距离	射手

样本容量

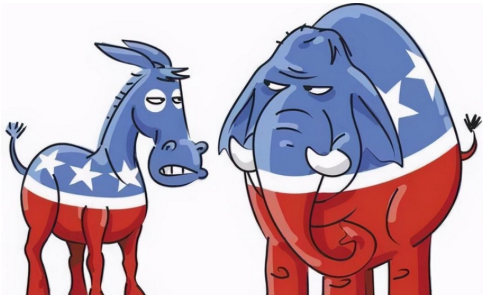
因变量（答案）

大数据的挑战

- 一个分析周期所需要的时间随着数据规模的增长而增长；
- 大型数据集的可视化过程非常复杂；
- 简单的模型不需要大量的数据来匹配或评估。

美国总统选举如何获取选民偏好?

- 分析大量X或Meta的网络数据并从中推断选民的观点；
- 通过民意调查，对特定问题进行调查问卷。



根据所需要完成的任务来选择数据，不用盲目追求大型数据集。

变量

数据集从列的角度看，每一列对应一个变量（也称特征），用于描述某种属性或状态。

自变量（内生变量）

因变量（目标变量、外生变量）

序号	姓名	物理攻击强度	法术攻击强度	控制技能	攻击距离	种属
1	奥拉夫	100	20	无	近距离	战士
2	武器大师	90	30	有	近距离	打野
3	阿卡丽	20	100	无	中距离	刺客
4	诡术妖姬	10	110	有	中距离	法师
5	璐璐	40	80	有	远距离	辅助
6	卢锡安	110	10	无	远距离	射手

变量的类型

变量根据取值类型可大致分为：

- 数值型，可以直接带入代数公式计算
  - 连续数值和非连续数值

大多数时候处理的数据都是非连续型数值，以整数或浮点数的方式存储。

- 类别型，不能像数值型一样处理，一般要进行编码
  - 顺序型
  - 类别型

即便如此，也不能忽略类别型数据本身的含义。例如，将头发按照颜色编码后，难道你的头发颜色减去我的头发颜色是有意义的吗？最大头发颜色和最小头发颜色又有什么意义呢？

序号	姓名	物理攻击强度	法术攻击强度	控制技能	攻击距离	种属
1	奥拉夫	100	20	无	近距离	战士

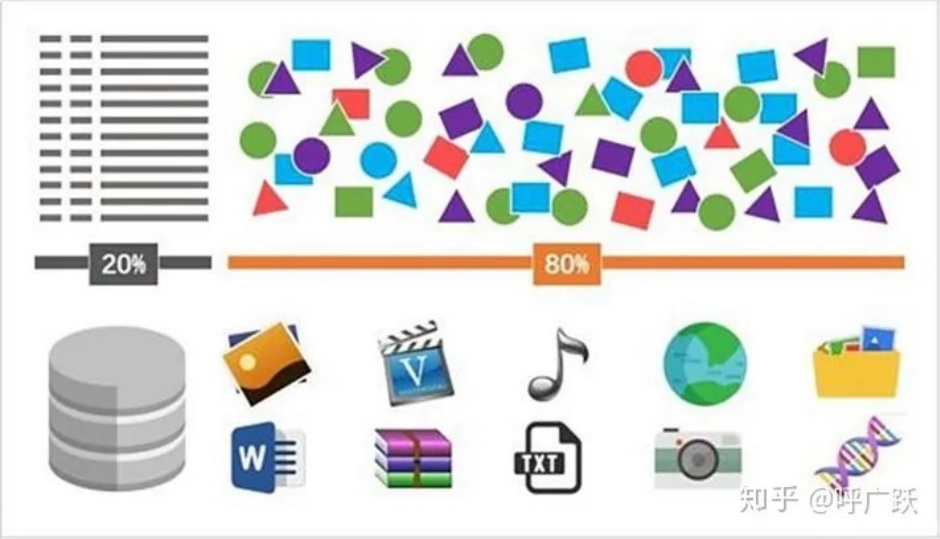
数据的类型

根据数据的特点，可大致分为**结构化数据**和**非结构化数据**。

- 结构化数据的特点是：属性（变量）的值是可定长的；各样本都具有共同的、确定性的属性。

从某个角度看，结构化数据是便于数据库存储的。

- 非结构化数据的特点是：属性往往是不定长的，且很难直接确定属性，需要进行必要的数字化处理和格式转换。



机器学习的任务

预测任务（建模）： $y = f(x; \Theta)$

以**数据预测（监督学习）**为核心的任务：

- 从数据集出发，归纳出输入变量和输出变量之间的数量关系。基于这种关系，一方面可以发现对输入变量产生重要影响的输入变量；另一方面，在数量关系具有普适性和未来不变的假设下，可用于对新数据的输出变量的取值进行预测。
- 常见的预测任务有**分类和回归**。

对数值型输出变量的预测是回归；对类别型输出变量的预测就是分类了。

日期	AQI	质量	PM2.5	PM10	SO_2	CO	NO_2	O_3
2019/1/4	40	优	18	40	5	0.5	26	61
2019/1/5	47	优	17	34	7	0.5	37	49
2019/1/6	88	良	64	95	12	1.4	70	13

Q1: SO_2, CO, NO_2, O_3，哪些是影响PM2.5的重要因素？ 回归

Q2: 哪些污染物的减少能有效改善空气质量等级？ 分类

机器学习的任务

聚类任务（建模）： $x \in c_i,\ \cup c_i = C$

以**数据聚类（无监督学习）**为核心的任务：

- 数据聚类的目的是发现数据中可能存在的小类（簇、子类），并通过小类刻画和揭示数据的内在组织结构。数据聚类的结果是给每个样本指派一个属于哪个小类的标签。

聚类和分类的区别在于，聚类得到的标签，是不属于数据集本身的；而分类预测的类别是数据集本身存在的变量

物理攻击强度	法术攻击强度	控制技能	攻击距离	种属
100	20	无	近距离	战士
90	30	有	近距离	打野
20	100	无	中距离	刺客
10	110	有	中距离	法师
40	80	有	远距离	辅助
110	10	无	远距离	射手

机器学习的任务

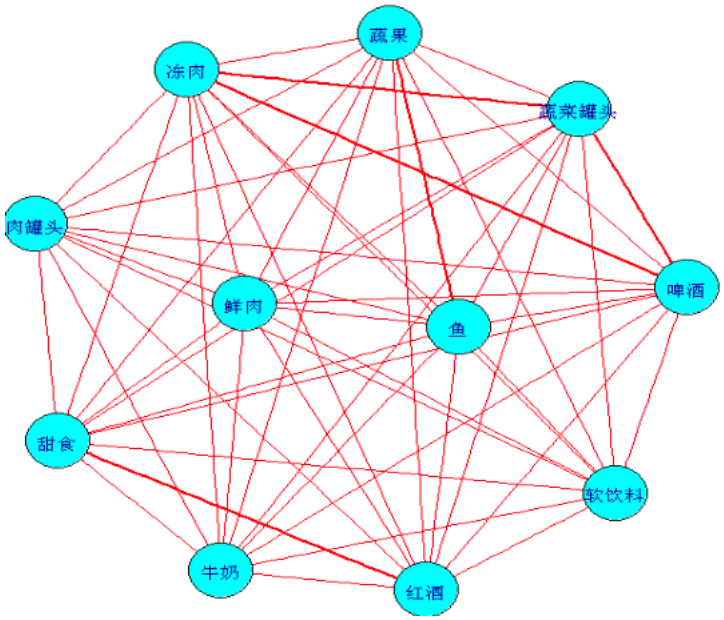
其他任务：

- **关联分析，**目的是找到事物之间的联系规律，发现它们之间的关联性。

对一段时间内某超市的购物小票数据集，每张小票记录了哪个人在哪个时间买了哪些商品以及数量等。

Q1: 购买蔬果的人中，同时购买鱼的可能性大，还是同时购买鲜肉的可能性大？ 空间关联性

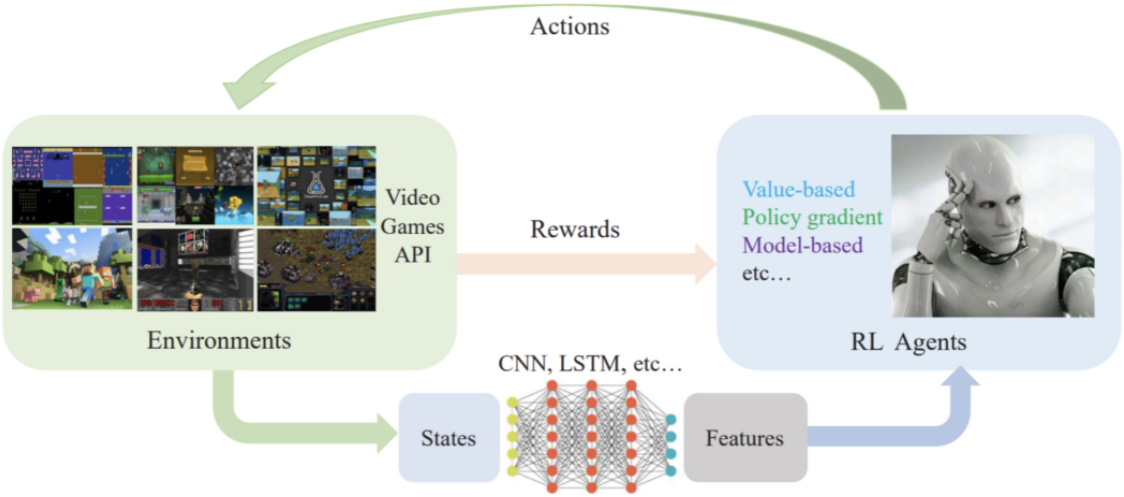
Q2: 购买甜食的人，未来一个月内购买红酒的可能性有多大？ 时间关联性



机器学习的任务

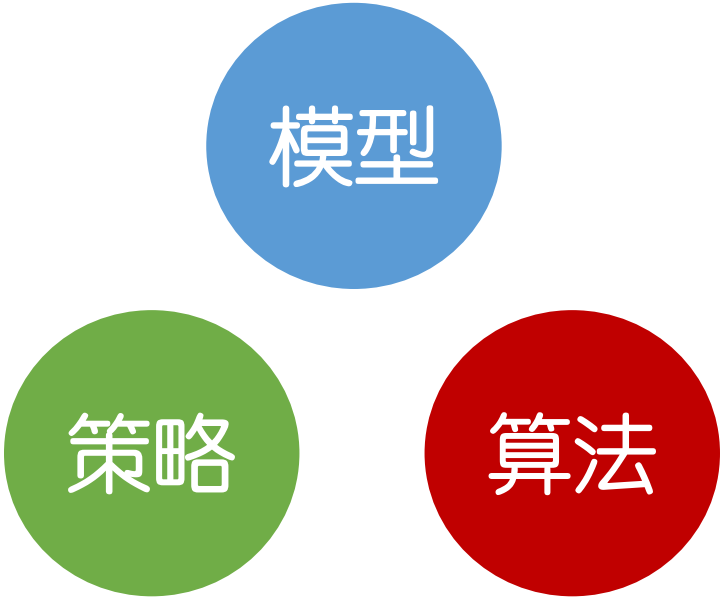
其他任务：

- **强化学习，**目的是寻找更好的决策的过程。



机器学习三要素

机器学习三要素



机器学习的任务

监督学习三要素：

假设决策函数是输入变量的线性函数，那么模型额假设空间就是所有这些线性函数的集合。

- **模型**

- 监督学习的模型是所要学习的条件概率分布或决策函数；
- 模型的假设空间包含所有可能的条件概率分布或决策函数；
- 假设空间的模型一般有无穷多个。

决策函数的集合：

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

非概率模型

条件概率分布的集合：

$$\mathcal{F} = \{P|P(Y|X)\} = \{P|P_\theta(Y|X), \theta \in \boldsymbol{R}^n\}$$

概率模型

机器学习的任务

监督学习三要素：

- **策略**

- 策略考虑按照什么样的准则学习或选择最优模型；
- 损失函数用来度量模型一次预测的好坏；
- 风险函数用来度量平均意义下模型预测的好坏。

损失函数：度量预测值$f(X)$和真实值Y的非负实值函数，记为$L(Y, f(X))$

风险函数：模型输入输出(X, Y)遵循联合分布$P(X, Y)$，所以损失函数的期望

$$R_{exp}(f) = E_P[L(Y, f(X)] = \int_{X \times Y} L\big(y, f(x)\big) P(x, y) dxdy$$

常用的损失函数有0-1损失函数、平方损失行数、绝对损失函数、对数损失函数、对数似然函数等。

机器学习的任务

监督学习三要素：

- 策略
- 实际学习时，给定训练数据集，模型$f(X)$关于训练数据集的平均损失，称为经验风险或经验损失。

$$R_{emp}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i))$$

根据大数定律，当样本N足够大时，经验风险趋于期望风险。然而实际学习中，训练样本数量有限，甚至很小，因此需要对经验风险进行矫正。

机器学习的任务

监督学习三要素：

样本数量越大，ERM的效果越好。然而样本不足时，容易出现过拟合现象。

- **策略**
- 经验风险矫正的方法有两种：
  - 经验风险最小化（empirical risk minimization，ERM）

$$\min_{f \in \mathcal{F}} R_{emp}(f)$$

  - 结构风险最小化（structural risk minimization，SRM）

结构风险最小化的目的是防止过拟合，主要思想是在经验风险上加上表示模型复杂度的正则化项或惩罚项。模型越复杂，惩罚项越大。

$$\min_{f \in \mathcal{F}} R_{emp}(f) + \lambda J(f)$$

机器学习的任务

监督学习三要素：

样本数量越大，ERM的效果越好。然而样本不足时，容易出现过拟合现象。

- **算法**
- 确定了模型和策略，在给定训练数据集上，使用具体算法求解最优模型。
- 对ERM或SRM来说，其实就是优化问题。
  - 存在解析解：直接给出；
  - 解析解难以获得：数值计算。

机器学习方法的分类

机器学习方法的分类

从建模的角度：

- **参数化方法：**在一套具体的模型族中，每一个具体的模型都可以用一个具体的参数向量来唯一确定。因此确定了参数向量也就确定了模型。

以分类为例： $y = f(\boldsymbol{x}; \boldsymbol{\Theta})$

优化目标： $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \frac{1}{|D|} \sum_{(x,y)\in D} L(y, f_{\boldsymbol{\theta}}(\boldsymbol{x}))$

参数化方法模型的参数量不会随数据集大小而变化。因此在计算过程中，模型占用计算机的资源（内存或显存）是固定的。

常见的参数化方法有线性回归、逻辑斯蒂回归、神经网络等。

机器学习方法的分类

从建模的角度：

- **非参数化方法：**非参数化模型并非由一个具体的参数向量决定，其训练的算法也不是更新模型的参数，而是由具体的计算规则直接在模型空间中寻找模型实例。

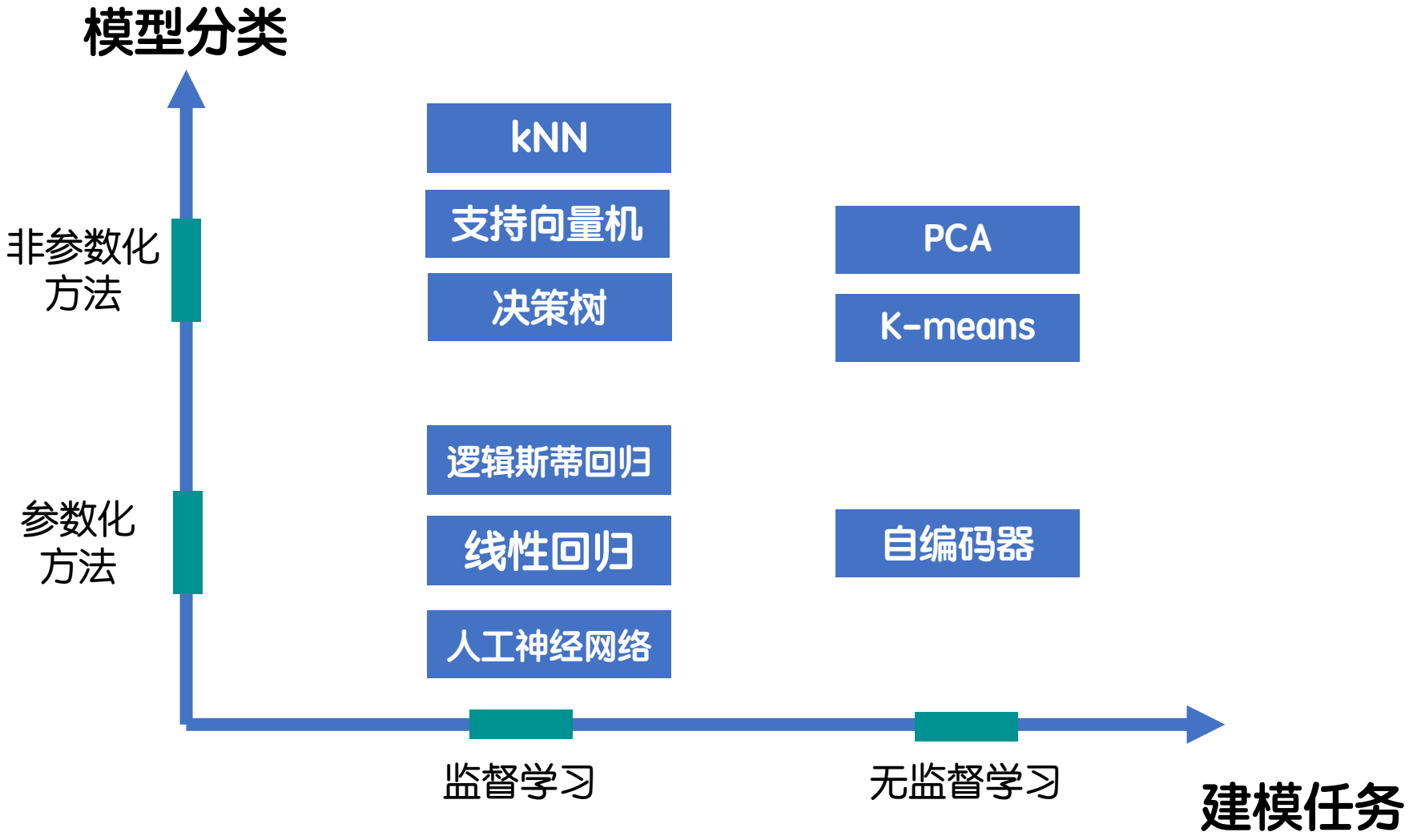
以聚类为例： $P(z|\boldsymbol{x})$

非参数化模型和参数并非一一对应，因此数据量不同（或数据不同）会导致模型中具体使用的参数量也不同。

常见的非参数化方法有kNN，支持向量机，树模型等。

课程体系

课程体系


模型分类
非参数化
方法
参数化
方法
kNN
支持向量机
决策树
PCA
K-means
逻辑斯蒂回归
线性回归
自编码器
人工神经网络
监督学习
无监督学习
建模任务