

Lecture 1: 机器学习基础概念

基本术语

概念: 基本术语 - 关于训练数据

- 数据集: 训练, 测试;
 - 测试数据和训练数据应该是分开、不相等的;
 - 测试数据一开始就应该分开、设计好的;
- 示例 (instance), 样例 (example)
 - 示例: 是有结果的;
 - 样例: 无结果的
- 样本 (sample)
- 属性 (attribute), 特征 (feature); 属性值;
 - 属性: 也称之为特征;
- 属性空间, 样本空间, 输入空间;
 - 把每一个属性看作坐标上的一个轴;
 - 属性构成的所有样本构成属性空间;
- 特征向量 (feature vector) 标记空间, 输出空间

概念: 基本术语 - 关于模型

- 模型:
 - 包含了一个规律, 形成了一个假设 (假设就是模型学到的东西);
- 假设:
 - 对一个问题可以形成很多个假设;
- 真相:
 - 而正确的答案就是真相
 - 比如在 $f(x) = y$
 - 这里的 $f(x)$ 就是假设;
 - 这里的 y 就是真相;
- 学习器:
 - 学到的模型, 但是是从过程的角度来描述的;

概念: 基本术语 - 关于输出

- 分类:
 - 如果做的事离散的输出 (只有 `TRUE` 或 `FALSE`), 则是分类;
- 回归:
 - 如果是做连续的输出 (比如0到1之间), 则是回归;
- 二分类:
 - 好、坏;
- 多分类:

- 可以拆分为若干个二分类；
- 正类：
 - 正类、反类通常是可交换的；
 - 因此正负类大多数时候是数据量差不多的；
- 反类：

机器学习分类

监督学习

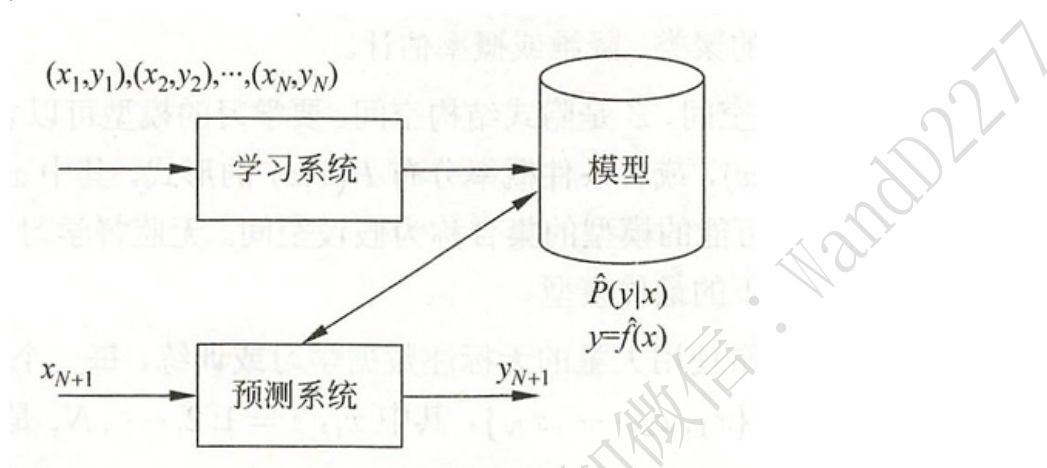
概念：监督学习

- 概念：
 - 有导师的学习；
 - 有样例的结果部分；
 - 拿到的数据里面有期望的结果，可以判断好坏；
- 描述：
 - 监督学习 (supervised learning) 是指从标注数据中学习预测模型的机器学习问题。标注数据表示输入输出的对应关系，预测模型对给定的输入产生相应的输出；
 - 监督学习的本质是学习输入到输出的映射的统计规律；
- 概念：输入与输出
 - 在监督学习中，将输入与输出看作是定义在输入 (特征) 空间与输出空间上的随机变量的取值；
 - 输入输出变量用大写字母表示，习惯上输入变量写作 X ，输出变量写作 Y 。输入输出变量的取值用小写字母表示，输入变量的取值写作 x ，输出变量的取值写作 y ；
 - 变量可以是标量或向量，都用相同类型字母表示；
- 概念：回归问题和分类问题
 - 输入变量 X 和输出变量 Y 有不同的类型，可以是连续的，也可以是离散的；
 - 人们根据输入输出变量的不同类型，对预测任务给予不同的名称；
 - 回归：
 - 输入变量与输出变量均为连续变量的预测问题，称为回归问题；
 - 分类：
 - 输出变量为有限个离散变量的预测问题称为分类问题；
 - 输入变量与输出变量均为变量序列的预测问题称为标注问题；
- 概念：假设空间
 - 监督学习的目的在于学习一个由输入到输出的映射，这一映射由模型来表示；
 - 换句话说，学习的目的就在于找到最好的这样的模型。模型属于由输入空间到输出空间的映射的集合，这个集合就是假设空间 (hypothesis space)；
 - 假设空间的确定意味着学习的范围的确定；

概念：监督学习的过程

- 概念：
 - 监督学习利用训练数据集学习一个模型，再用模型对测试样本集进行预测；

- 由于在这个过程中需要标注的训练数据集，而标注的训练数据集往往是人工给出的，所以称为监督学习；
- 监督学习分为学习和预测两个过程，由学习系统与预测系统完成；
- 概念：学习系统
 - 学习系统 (也就是学习算法) 试图通过 **训练数据集** 中的样本 (x_i, y_i) 带来的信息学习模型；
 - 具体地说，对于输入 x_i ，一个具体的模型 $y = f(x)$ 可以产生一个输出 $f(x_i)$ ，而训练数据集中对应的输出是 y_i ；
 - 如果这个模型有很好的预测能力，训练样本输出 y_i 和模型输出 $f(x_i)$ 之间的差就应该足够小。学习系统通过不断地尝试，选取最好的模型，以便对训练数据集有足够好的预测，同时对未知的测试数据集的预测也有尽可能好的推广；
- 图示：



无监督学习

概念：无监督学习

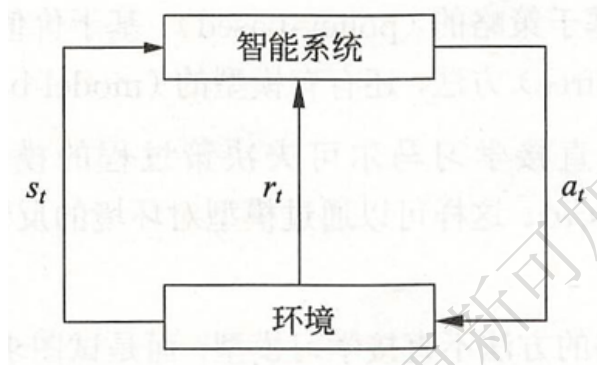
- 概念：无监督的学习：
 - 样例的结果部分为空；
 - 拿到数据后，连要得出什么期望的结果都没有；
 - 这时通常是将数据分成很多堆，进行聚类、未来再进行进一步处理；
 - 无监督学习 (unsupervised learning) 是指从**无标注数据中学习预测模型**的机器学习问题。无标注数据是自然得到的数据，预测模型表示数据的类别、转换或概率；
 - 无监督学习的本质是学习**数据中的统计规律或潜在结构**；
 - 应用：聚类、密度估计
- 概念：未见样本
 - 新的、未见的数据；
- 概念：独立同分布
 - 每个样本摘下来、都属于同一个分布，但是是独立的；
 - 因此才可以用每个样本的频率、去逼近它的概率；
 - 概念：泛化
 - 可以理解为推广；
 - 模型的泛化性能；

- 泛化的性能可以指导我们是否要做这个事情，如果一个模型泛化性能太差，则不需要做这个模型；

强化学习

概念：强化学习

- 概念：
 - 强化学习 (reinforcement learning) 是指智能系统在环境的连续互动中学习最优行为策略的机器学习问题；
 - 假设智能系统与环境的互动基于马尔可夫决策过程 (Markov decision process)，智能系统能观测到的是与环境互动得到的数据序列；
 - 强化学习的本质是学习最优的序贯决策；
 - 马尔可夫决策过程具有马尔可夫性，下一个状态只依赖于前一个状态与动作；
- 图示：
 - 智能系统的目标不是短期奖励的最大化，而是长期累积奖励的最大化；
 - 强化学习过程中，系统不断地试错 (试验和错误)，以达到学习最优策略的目的；



半监督学习与主动学习

概念：半监督学习与主动学习

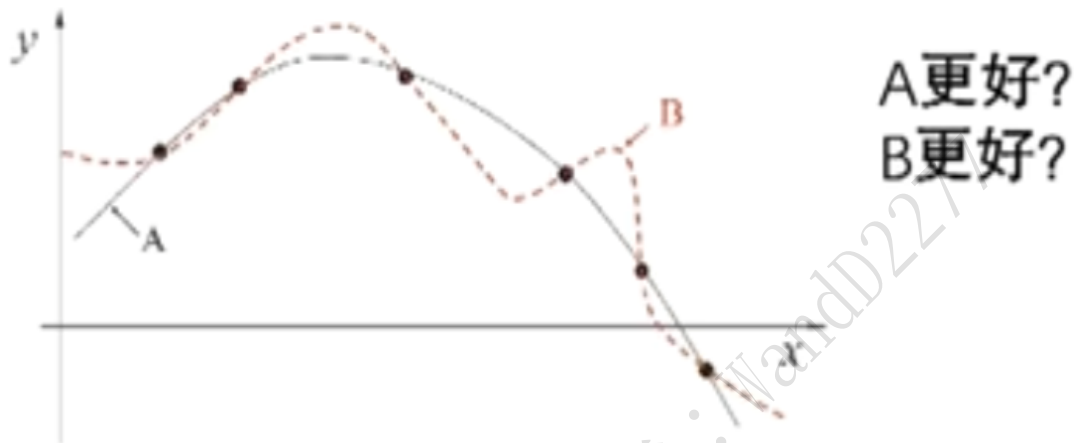
- 概念：半监督学习
 - 半监督学习 (semi-supervised learning) 是指利用标注数据和未标注数据学习预测模型的机器学习问题；
 - 通常有少量标注数据、大量未标注数据，因为标注数据的构建往往需要人工，成本较高，未标注数据的收集不需太多成本；
 - 半监督学习旨在利用未标注数据中的信息，辅助标注数据，进行监督学习，以较低的成本达到较好的学习效果；
- 概念：主动学习
 - 主动学习 (active learning) 是指机器不断主动给出实例让教师进行标注，然后利用标注数据学习预测模型的机器学习问题；
 - 通常的监督学习使用给定的标注数据，往往是随机得到的，可以看作是“被动学习”，主动学习的目标是找出对学习最有帮助的实例让教师标注，以较小的标注代价，达到较好的学习效果；

归纳偏好与 NFL

归纳偏好

概念：归纳偏好

- 英文：
 - Inductive Bias
- 概念：
 - 机器学习算法在学习过程中对某种类型假设的偏好；



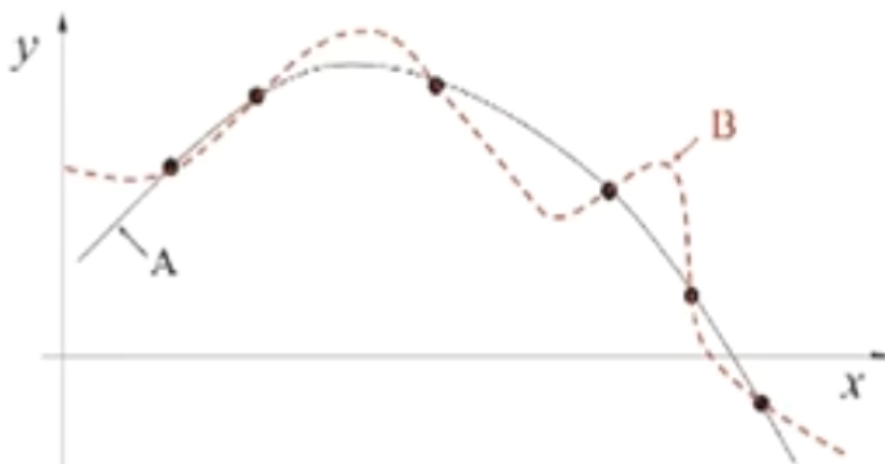
- 任何一个有效的机器学习算法必有其偏好；
- 学习算法的归纳偏好是否**与问题本身匹配**，大多数时候直接决定了算法能否取得好的性能；
- 原则：
 - 奥卡姆剃刀：若非必要，勿增实体
 - 如果出现很多假说都可以解释这个现象时，就选择最简单的假设来进行解释；
 - 如果多个模型都可以解释同一个现象，那就使用一个最简单的模型；
 - 因此如果 A 和 B 都可以描述同一个现象，应该选择 **A**；
- 概念：关于最简单
 - 比如 $y = ax^2 + bx + c$
 - 或者 $y = ax^3 + c$
 - 这两个模型、一个次方更小，一个变量更少，哪一个更简单、是一个不简单的问题；
 - 模型背后形成的假设、哪一个更简单，不简单进行判断；

NFL 定理

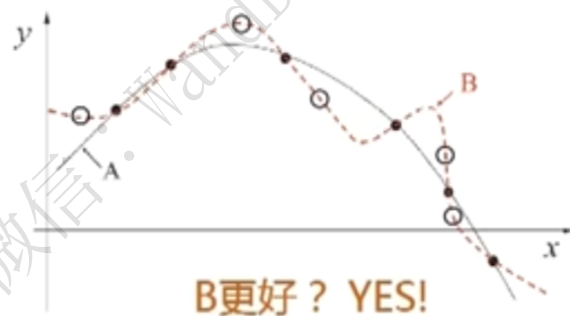
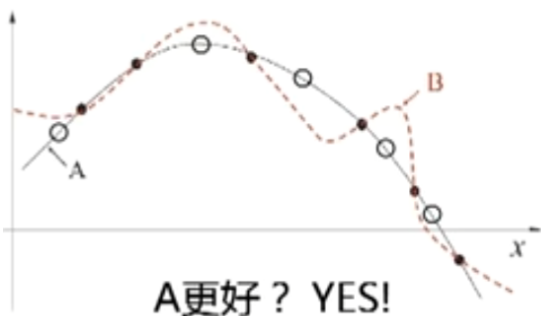
概念：NFL 定理

- 引入：

黑点：训练样本；白点：测试样本



- 对于若干数据点，A 和 B 两个曲线都完全的穿过去，哪一个更好？



• 概念：NFL 定理

- NFL 定理：一个算法若在某些问题上比另一个算法好，必存在另一些问题好；
- 所有“问题”出现的机会相同、或所有问题同等重要，实际情形并非如此；
- 我们通常只关注自己正在试图解决的问题脱离具体问题，空泛地谈论“什么学习算法更好”毫无意义；
- 具体问题、具体分析；

• 分析：

- 什么是机器学习里面的问题？给定一个输入域 X ，输出域 Y ， X 到 Y 是确定的，这时建立函数使得 $f: X \rightarrow Y$ ，这是个问题；
- 因此：比如同样是推荐问题，在腾讯推荐游戏、和在阿里巴巴推荐衣服，就是不一样的，因为他们的 X 和 Y 不一样，因此在数据分布上建模就会不一样；
- 因此如果不谈 X 、不谈 Y ，只是泛泛的讨论模型、 $F(x)$ ，是毫无意义的；

补充：关于机器学习算法的认识

- 机器学习并非“十大套路”“二十大招数”的简单堆积现实任务千变万化，以有限的“套路”应对无限的“问题”，焉有不败？
- 最优方案往往来自：按需设计、度身定制

统计学习方法三要素

概念：统计学习方法三要素

- 概念:

- $$\text{方法} = \text{模型} + \text{策略} + \text{算法}$$

- 概念: 模型

- 统计学习首要考虑的问题是学习什么样的模型;
 - 在监督学习过程中, 模型就是所要学习的条件概率分布或决策函数。模型的假设空间 (hypothesis space) 包含所有可能的条件概率分布或决策函数;
 - 例如, 假设决策函数是输入变量的线性函数, 那么模型的假设空间就是所有这些线性函数构成的函数集合。假设空间中的模型一般有无穷多个。
 - 假设空间用 F 表示。假设空间可以定义为决策函数的集合;

- 概念: 策略

- 损失函数度量模型一次预测的好坏, 风险函数度量平均意义下模型预测的好坏;
 - 常见的损失函数:

- (1) 0-1 损失函数

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

- (2) 平方损失函数

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

- (3) 对数损失函数

- $$L(Y, P(Y|X)) = -\log P(Y|X)$$

- 概念: 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集, 根据学习策略从假设空间中选择最优模型, 最后需要考虑用什么样的计算方法求解最优模型;
 - 这时, 统计学习问题归结为最优化问题, 统计学习的算法成为求解最优化问题的算法;
 - 如果最优化问题有显式的解析解, 这个最优化问题就比较简单。但通常解析解不存在, 这就需要用数值计算的方法求解;
 - 如何保证找到全局最优解, 并使求解的过程非常高效, 就成为一个重要问题。统计学习可以利用已有的最优化算法, 有时也需要开发独自的最优化算法;