第1章 统计学习及监督学习概论

内容提要

- 统计学习
- 统计学习的分类
- 统计学习方法三要素

0 机器学习

人工智能

- 人工智能(artificial intelligence，AI)，是让机器表现出类似人类智能行为的学科与技术总称，目标是使计算机能感知环境、推理决策、学习适应并执行复杂任务。
  - 通常是指通过普通计算机程序来呈现人类智能的技术
  - 研究这样的智能系统否能够实现，以及如何实现。同时，通过医学、神经科学、机器人学及统计学等的进步，常态预测则认为人类的很多职业也逐渐被其取代
- 人工智能在早期教材中的定义是“智能主体(intelligent agent)的研究与设计”
  - 智能主体，指一个可以观察周遭环境并作出行动以达致目标的系统
  - 约翰·麦卡锡(1955年)“制造智能机器的科学与工程”
  - 安德烈亚斯·卡普兰(Andreas Kaplan)和迈克尔·海恩莱因(Michael Haenlein) “系统正确解释外部数据，从这些数据中学习，并利用这些知识通过灵活适应实现特定目标和任务的能力”
  - 人工智能可定义为模仿人类与人类思维相关的认知功能的机器或计算机，学习和解决问题

机器学习

- 机器学习（Machine Learning, ML）是人工智能的一个分支
  - 人工智能：“推理”，到“知识”，再到“学习”。【模拟人类智能】
  - 机器学习：以机器学习为手段解决人工智能中的问题

- 机器学习，研究如何通过数据和算法使计算机自动从经验中改进性能，主要关注“从数据中学习模型”以完成预测、分类、聚类等任务
  - 机器学习理论，主要是设计和分析一些让计算机可以自动“学习”的算法
  - 机器学习算法，从数据中分析获得规律，并利用规律对未知数据进行预测
    - 因为学习算法中涉及了大量的统计学理论，机器学习与推断统计学联系尤为密切，也被称为统计学习理论
  - 算法设计方面，机器学习理论关注可以实现的，行之有效的学习算法
    - 很多推论问题属于无程序可循难度，所以部分的机器学习研究开发容易处理的近似算法

机器学习应用

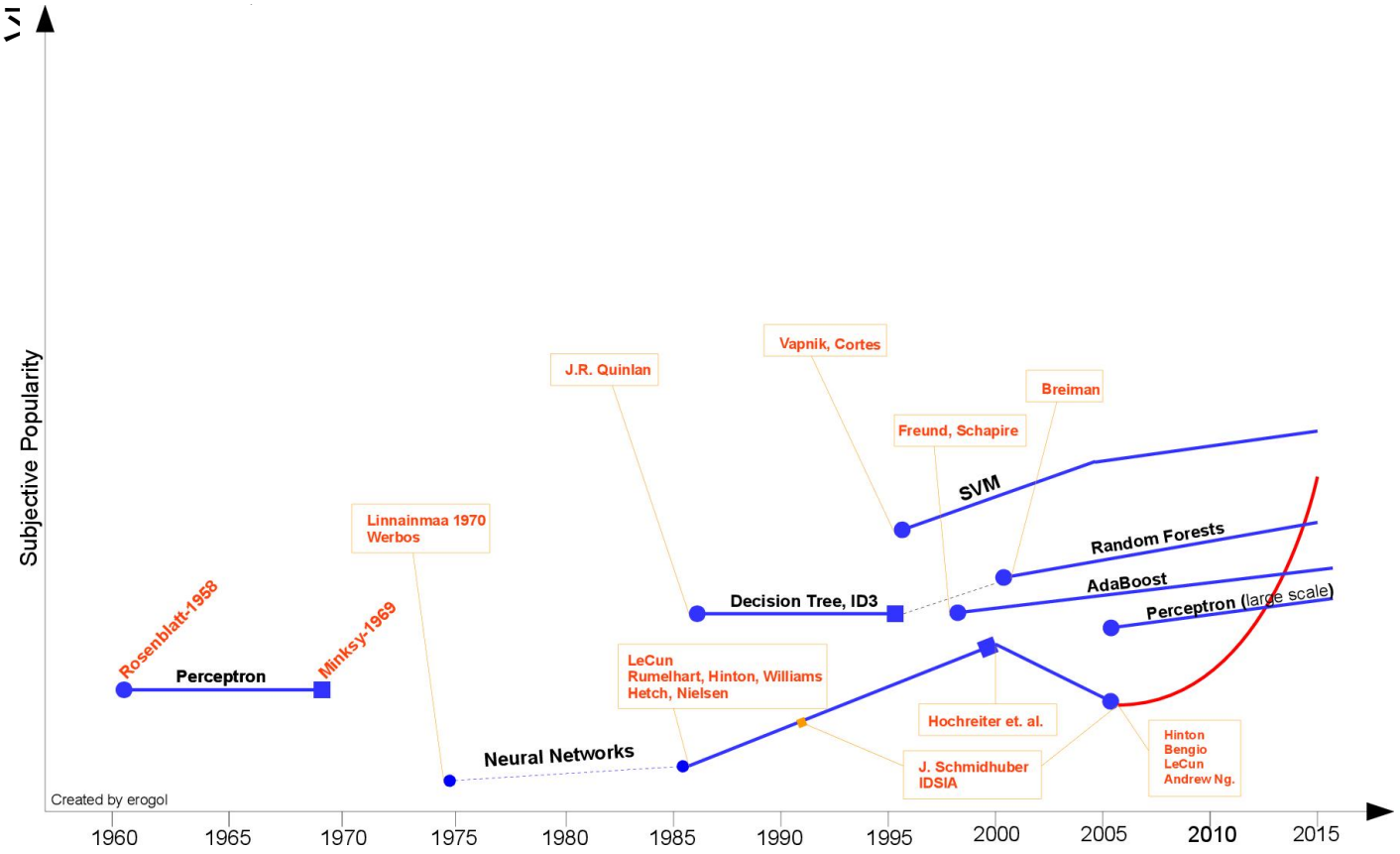
- 机器学习已广泛应用于
  - 数据挖掘
  - 计算机视觉
  - 自然语言处理
  - 生物特征识别
  - 搜索引擎
  - 医学诊断
  - 检测信用卡欺诈
  - 证券市场分析
  - DNA序列测序
  - 语音和手写识别
  - 战略游戏
  - 机器人
  - …

人工智能的诞生(1943年~1956年)

- Warren McCulloch和Walter Pitts在1943年发表了人工智能领域的开篇之作，提出了人工神经网络模型
- 1955 年夏天，几位计算机科学与信息学、人工智能先驱 John McCarthy、Marvin Minsky、Nathaniel Rochester、Claude Elwood Shannon 在达特茅斯发起了一个夏季科研项目提议，首次提出了“Artificial Intelligence”的概念

感知机到深度学习（1960-2012）

- 机器学习



深度学习时期（ 2012 -2020）

- 大数据（Big Data）： 互联网的发展催生了海量的数据集，如 ImageNet。没有足够的数据，深度神经网络无法得到充分训练，容易过拟合。
- 硬件算力（Hardware）： 英伟达（NVIDIA）等公司生产的图形处理器（GPU）被发现非常适合神经网络所需的大规模并行计算。AlexNet 就是使用两块 GTX 580 GPU 进行训练的，极大地缩短了训练时间，使得训练更深、更复杂的网络成为可能。
- 算法优化（Algorithms）： 一系列关键的算法改进被应用，例如：
  - ReLU 激活函数： 解决了传统Sigmoid函数在深层网络中容易出现的梯度消失问题。
  - Dropout： 一种有效的正则化技术，防止模型过拟合。
  - 多层卷积神经网络（CNN） 的成熟设计。
- 结论： 2012年的AlexNet 是深度学习从学术界的“圈内热”走向工业界和公众视野的“奇点”。此后，深度学习在图像识别、语音识别、自然语言处理等领域取得了一系列突破性进展，开启了至今仍在持续的AI浪潮。

大模型（技术引爆点 2020 年，大众引爆点 2022 年底）

- 大模型是深度学习发展到一定阶段的产物，其核心思想是：当模型的参数量和训练数据量跨越某个“临界点”后，会涌现出之前小模型所不具备的、令人惊叹的通用能力（如零样本/少样本学习、逻辑推理等）。
- 奠基性架构：Transformer (2017年)
  - 论文： Google 发表的著名论文 《Attention Is All You Need》，提出了完全基于自注意力机制的 Transformer 模型。
  - 意义： 它彻底取代了之前在NLP领域占主导地位的RNN/LSTM架构，其高度并行的特性完美契合了GPU集群，为构建前所未有的超大规模模型铺平了道路。
- 范式确立：BERT 和 GPT (2018-2019年)
  - BERT (Google, 2018)： 证明了通过在海量无标签文本上进行“预训练”，模型可以学到丰富的语言知识，然后在特定任务上进行“微调”即可取得极佳效果。
  - GPT-2 (OpenAI, 2019)： 拥有15亿参数，其生成的文本已经非常连贯流畅，让人们第一次看到了大规模语言模型的巨大潜力。当时OpenAI因担心其被滥用，甚至未立即放出完整模型。
- 技术引爆点：GPT-3 (2020年)
  - 模型： OpenAI 发布了拥有 1750亿 参数的 GPT-3。
  - 颠覆性成果： GPT-3 展示了惊人的“涌现能力”（Emergent Abilities）。它不需要为特定任务进行微调，仅通过少量示例（Few-shot）甚至零示例（Zero-shot）的提示（Prompting），就能完成翻译、写代码、写诗、做数学题等多种任务。这标志着“模型即服务”、通过Prompt与模型交互的全新AI范式诞生。
- 大众引爆点：ChatGPT (2022年11月30日)
  - 产品： OpenAI 发布了基于 GPT-3.5 优化的对话式AI——ChatGPT。

机器学习与其它学科关系

- 人工智能
- 数据挖掘
- 统计学习

机器学习与人工智能

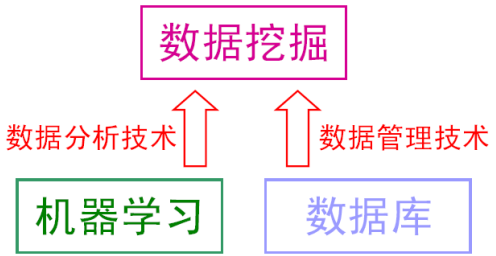
- 包含关系
  - 机器学习是实现人工智能的一种主要方法和工具。可以认为：人工智能 ⊃ 机器学习。
- 侧重点不同
  - AI更宽泛，关注“智能系统”的整体能力与架构
  - ML更聚焦于“从数据中学习模型”的方法和算法。
- 并非等同
  - 有些AI方法并不依赖机器学习（如基于规则的专家系统、符号推理）
  - 现代AI大量依赖机器学习（尤其是深度学习）来处理感知与理解任务

数据挖掘

- 数据挖掘(data mining)，一跨学科计算机科学分支。用人工智能、机器学习、统计学和数据库的交叉方法，在相对较大型的数据集中发现模式的计算过程
- 数据挖掘过程的总体目标，从一个数据集中提取信息，并将其转换成可理解的结构，以进一步使用
  - 除原始分析步骤，还涉及数据库和数据管理、数据预处理、模型与推断、兴趣度度量、复杂度考虑，及发现结构、可视化、在线更新等后处理
- “数据库知识发现”(Knowledge-Discovery in Databases, KDD)的分析步骤，本质上属机器学习范畴

机器学习和数据挖掘的关系

- 机器学习是数据挖掘的重要工具
  - 数据挖掘试图从海量数据中找出有用的知识
- 数据挖掘不仅仅要研究、拓展、应用一些机器学习方法
  - 还涉及许多非机器学习技术问题：数据仓储、大规模数据、数据噪音等实际问题
- 机器学习涉及面更宽
  - 用在数据挖掘上的方法通常只是“从数据学习”
  - 某些机器学习的子领域与数据挖掘关系不大，如增强学习与自动控制等
- 数据挖掘可以视为机器学习和数据库的交叉
  - 利用机器学习界提供的技术来分析海量数据
  - 利用数据库界提供的技术来管理海量数据



1 统计学习

统计学习的特点

- 统计学习(statistical learning)：关于计算机基于数据构建概率统计模型，并运用模型对数据进行分析与预测的一门学科
  - 统计学习也称为统计机器学习(statistical machine learning)
- 统计学习的主要特点
  - 统计学习以计算机系统为（硬件）平台
  - 统计学习以数据为研究对象，是数据驱动的学科
  - 统计学习的目的，是对数据进行分析与预测
  - 统计学习是概率论、统计学、信息论、计算理论、最优化理论及计算机科学等多个领域的交叉学科，并且在发展中逐步形成独自的理论体系与方法论

统计学习的对象：数据

- 统计学习的对象是数据(data)
  - 从数据出发，提取数据的特征，抽象出数据的模型，发现数据中的知识，又回到对数据的分析与预测中去
- 统计学习关于数据的基本假设
  - 同类数据具有统计规律，这是统计学习的前提

- 数据具有统计规律性，可以用概率统计方法来加以处理
  - 如，可以用随机变量描述数据中的特征，用概率分布描述数据的统计规律

- 在统计学习过程中，以变量或变量组表示数据。数据分为由连续变量和离散变量表示的类型

统计学习的目的：对数据进行分析与预测

- 统计学习用于对数据进行预测与分析，特别是对未知新数据进行分析与预测
  - 对数据的预测可以使计算机更加智能化，或者说使计算机的某些性能得到提高
  - 对数据的分析可以让人们获取新的知识，给人们带来新的发现

- 对数据的分析与预测是通过构建概率统计模型实现的

- 统计学习总的目标就是考虑学习什么样的模型和如何学习模型
  - 模型能对数据进行准确的分析与预测
  - 尽可能提高学习效率

统计学习的方法

- 统计学习的方法：基于数据，构建统计模型，从而对数据进行预测与分析
- 统计学习
  - 监督学习(supervised learning)
  - 非监督学习(unsupervised learning)
  - 半监督学习(semisupervisedlearning)
  - 强化学习(reinforcement learning)
- 实现统计学习方法的步骤：
  - 得到一个有限的训练数据集合
  - 确定包含所有可能的模型的假设空间，即学习模型的集合
  - 确定模型选择的准则，即学习的策略
  - 实现求解最优模型的算法，即学习的算法
  - 通过学习方法选择最优模型
  - 利用学习的最优模型对新数据进行预测或分析

统计学习的研究

- 统计学习研究
  - 统计学习方法(statistical learning method)
    - 旨在开发新的学习方法【如何做，具体过程】
  - 统计学习理论(statistical learning theory)
    - 在于探求统计学习方法的有效性与效率，以及统计学习的基本理论问题
    - 【为什么这样做可行以及性能等，偏理论】
  - 统计学习应用(application of statistical learning)
    - 主要考虑将统计学习方法应用到实际问题中去，解决实际问题
    - 【应用】

统计学习的重要性

- 统计学习学科在科学技术中的重要性
  - 统计学习是处理海量数据的有效方法
    - 现实中的数据规模大，常具有不确定性，统计学习往往是处理这类数据最强有力的工具
  - 统计学习是计算机智能化的有效手段
    - 智能化是计算机发展的必然趋势，也是计算机技术研究与开发的主要目标。利用统计学习模仿人类智能的方法，虽有一定的局限性，但仍然是实现这一目标的最有效手段
  - 统计学习是计算机科学发展的一个重要组成部分
    - 计算机科学由三维组成：系统、计算、信息。统计学习主要属于信息这一维，并在其中起着核心作用

统计学习和机器学习的差异

- 研究方法差异
  - 统计学研究形式化和推导
  - 机器学习更容忍一些新方法
- 维度差异
  - 统计学强调低维空间问题的统计推导(confidence intervals, hypothesis tests, optimal estimators)【统计学模型中推导】
  - 机器学习强调高维预测问题(偏应用)【不一定有统计模型】

统计学习和机器学习(专业术语)

统计学	机器学习
Estimation (估计)	Learning(学习)
Classifier(分类器)	Model(模型)
Data point（数据)	Example/Instance(样本/实例)
Regression (回归)	Supervised Learning(监督学习，连续变量输出)
Classification(分类)	Supervised Learning(监督学习，离散变量输出)
Feature (特征)	Feature(特征，本质的属性)
Response(输出变量，响应变量)	Label(标签，类别)
Prediction(预测)	
Inference(推断，寻找关联)	

统计学习 - 基本假设

- 统计学习的对象
  - 数据：计算机及互联网上的各种数字、文字、图像、视频、音频数据以及它们的组合
- 统计学习的假设
  - 数据的基本假设是同类数据具有一定的统计规律性
- 统计学习的目的
  - 用于对数据(特别是未知数据)进行预测和分析

2 统计学习的分类

统计学习的分类

- 基本分类
  - Supervised learning
  - Unsupervised learning
  - Semi-supervised learning
  - Reinforcement learning
- 按模型分类
- 按算法分类
- 按技巧分类

基本分类

- Supervised learning
- Unsupervised learning
- Semi-supervised learning
- Reinforcement learning

特征

- “特征”
  - 通常指数据中的某些属性或特点，这些特点可以用来描述数据的特征
  - 界定数据自身
  - 区别于其它

监督学习(Supervised Learning)

- 监督学习指从标注数据中学习预测模型的机器学习问题
  - 标注数据(对输入做标记/标签)：表示输入输出的对应关系(数值/类别)
  - 预测模型：对给定的输入产生相应的输出
  - 监督学习的本质：学习输入到输出的映射的统计规律

- 输入空间(input space)：输入所有可能取值的集合
- 输出空间(output space) ：输出所有可能取值的集合
- 实例(Instance)：每一个具体的输入
  - 用特征向量(feature vector)方式表示
- 特征空间(feature space): 所有特征向量存在(可能取值)的空间

监督学习 – 输入空间、特征空间和输出空间

- 输入变量X，实例x的特征向量$x = \left(x^{(1)}, x^{(2)}, \cdots, x^{(i)}, \cdots, x^{(n)}\right)^{\mathrm{T}}$
  - 多输入变量中的第i个：$x_i = \left(x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)}\right)^{\mathrm{T}}$ 【列向量】
- 输出变量Y

- 训练(training data)集$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$
  - 样本/样本点：输入与输出对
- 监督学习：从训练数据(training data)集合中学习模型，对测试数据(test data)预测

- 输入与输出对又称为样本(sample) 或样本点

监督学习 – 分类

- 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
  - 分类问题：输出变量为有限个离散变量
  - 回归问题：输入输出均为连续变量
  - 标注问题：输入输出变量均为变量序列

监督学习 – 联合概率分布

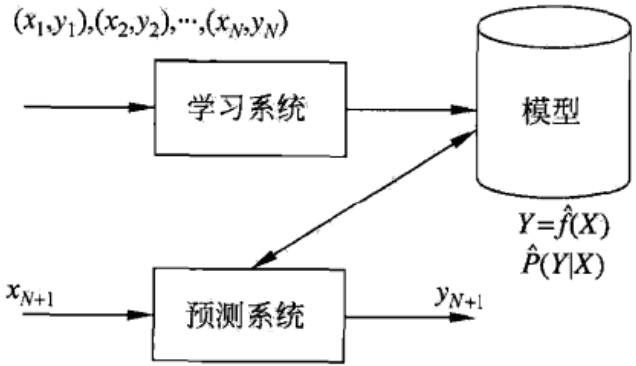
- 统计学习假设数据存在一定的统计规律
- 联合概率分布
  - 联合概率分布$P(X,Y)$，分布函数或分布密度函数，假定这个联合概率分布存在
  - 对于学习系统来说，联合概率分布未知
    - 训练数据和测试数据被看作是依联合概率分布$P(X,Y)$独立同分布产生
  - 学习就是求解联合概率分布或者其推导形式

监督学习 – 假设空间

- 统计学习假设数据存在一定的统计规律
- 监督学习
  - 学习目的：学习由输入到输出的映射
    - 假设空间(hypothesis space)：由输入空间到输出空间的映射的集合
  - 从假设空间中选出最优的那个映射/模型
- 模型表示的类别
  - 概率模型
    - 条件概率分布 $P(Y|X)$
    - 非决定的，如分类（属于某一类的可能性）
  - 决策函数
    - $Y = f(X)$
    - 决定性的，如回归

监督学习 – 问题形式化

- 问题形式化：学习与预测

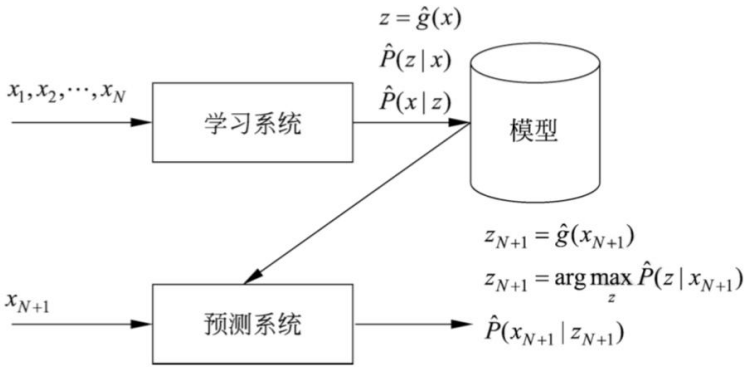


$$y_{N+1} = \arg\max_{y_{N+1}} \hat{P}(y_{N+1} \mid x_{N+1})$$

$$y_{N+1} = \hat{f}(x_{N+1})$$

无监督学习

- 从无标注数据中学习预测模型的机器学习问题
  - 无监督学习的本质是学习数据中的统计规律或潜在结构【侧重于自身特征和结构的建模】
- 无监督学习
  - 输入空间$\mathcal{X}$，隐式结构空间$\mathcal{Z}$，假设空间
- 无监督学习旨在从假设空间中选出在给定评价标准下的最优模型
- 训练集：$U = \{x_1, x_2, \cdots, x_N\}, x_i, i = 1,2, \cdots, N$
- 模型
  - 函数$z = g(x)$
  - 条件概率分布$P(z \mid x), P(x \mid z)$
- 学习得到的模型
  - 函数$z = \hat{g}(x)$
  - 条件概率分布$\hat{P}(z \mid x)$或者$\hat{P}(x \mid z)$



强化学习

- 强化学习（reinforcement learning) 是指智能系统在与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
- 强化学习的目标就是在所有可能的策略中选出价值函数最大的策略

半监督学习

- 半监督学习(semi-supervised learning)：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
  - 少量标注数据，大量未标注数据
  - 利用未标注数据的信息，辅助标注数据，进行监督学习
  - 优点：较低成本

按模型的种类

- 分类准则：统计规律的模型类型

- 概率与非概率
- 线性与非现性
- 参数化与非参数化

概率模型与非概率模型

- 监督学习
  - 概率模型(probabilistic model)，不确定
    - 如条件概率$P(\mathrm{y} \mid x)$
  - 非概率模型(non-probabilistic model)，确定性模型(deterministic model)
    - 如直接的映射关系$\mathrm{y} = f(x)$
- 在无监督学习
  - 概率模型，取条件概率分布形式$P(z \mid x)$,或$P(x \mid z)$
  - 非概率模型，取函数形式$z = g(x)$, 其中x是输入， z是输出
- 在监督学习中，概率模型是生成模型，非概率模型是判别模型
- 概率模型和非概率模型的区别不在于输入与输出之间的映射关系， 而在于模型的内在结构。 概率模型通常可以表示为联合概率（或者条件概率）分布的形式

概率模型与非概率模型

- 概率模型
  - 决策树、朴素贝叶斯、隐马尔可夫模型、条件随机场、概率潜在语义分析、潜在狄利克雷分配、高斯混合模型
- 非概率模型
  - 感知机、支持向量机、K 近邻、AdaBoost 、K 均值、潜在语义分析，以及神经网络
- 逻辑斯谛回归既可看作是概率模型，又可看作是非概率模型

概率模型中的生成模型与判别模型

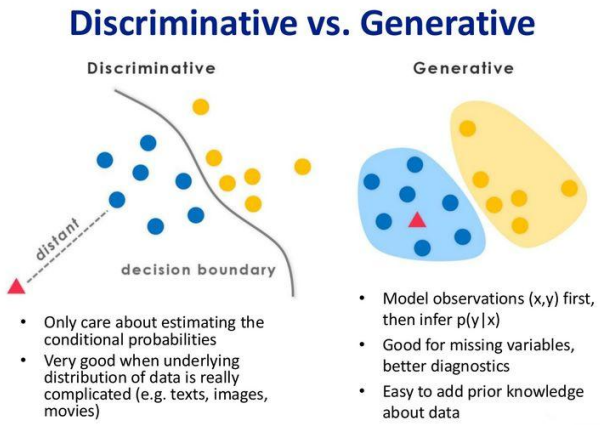
- 判别模型(关心条件概率分布)
  - 对$P(\mathrm{y} \mid x)$建模
    - 计算不同类别之间的最优分类面(decision boundary)
    - 着重不同类别数据之间的差异

- 生成模型(关心联合概率分布)
  - 对$P(x, \mathrm{y})$进行建模
    - 再根据贝叶斯公式计算$P(\mathrm{y} \mid x)$
  - 可用于存在隐藏参数的数据建模
  - 便于对模型添加先验知识
  - 可以用来生成新的数据

	sample 1	sample 2	sample 3	sample 4
x	0	0	1	1
y	0	0	0	1

	y = 0	y = 1
x = 0	1	0
x = 1	1/2	1/2

	y = 0	y = 1
x = 0	1/2	0
x = 1	1/4	1/4



概率图模型(probabilistic graphical model)

- 概率图模型
  - 联合概率分布由有向图或者无向图表示
  - 联合概率分布可以根据图的结构分解为因子乘积的形式
- 贝叶斯网络、 马尔可夫随机场、 条件随机场是概率图模型

线性与非现性

- 线性
  - 数乘和加法的组合
- 线性模型与非线性模型
  - 如果决策函数是线性函数，则称模型是线性模型，否则称模型是非线性模型

- 线性模型
  - 感知机、线性支持向量机、K近邻、K均值、潜在语义分析
- 非线性模型
  - 核函数支持向量机、AdaBoost、神经网络

- 深度学习(deep learning)实际是复杂神经网络的学习，是复杂的非线性模型

参数化模型与非参数化模型

- 参数化模型
  - 模型参数的维度固定，模型可以由有限维参数完全刻画
  - 【有结构】，有固定结构，求参数
- 非参数化模型
  - 模型参数的维度不固定或者说无穷大，随着训练数据量的增加而不断增大
  - 【无确定性结构】，通过算法计算得到

- 参数化模型适合问题简单的情况
  - 感知机、朴素贝叶斯、逻辑斯谛回归、K 均值、高斯混合模型
- 现实中问题，非参数化模型更有效
  - 决策树、支持向量机、AdaBoost 、K 近邻、潜在语义分析、概率潜在语义分析、潜在狄利克雷分配

按技巧分类

- 贝叶斯学习
- 核方法

贝叶斯学习(Bayesian learning)

- 贝叶斯学习, 贝叶斯推理(Bayesian inference) ,是统计学、机器学习中重要的方法
  - 在概率模型的学习和推理中，利用贝叶斯定理，计算在给定数据条件下模型的条件概率，即后验概率，并应用这个原理进行
    - 模型估算（学习），数据预测（推理）
  - 将模型、未观测要素及其参数用变量表示，使用模型的先验分布是贝叶斯学习的特点
- 贝叶斯定理
  - 全概公式(加法规则)
    - $P(x) = \sum_y P(x, y)$
  - 乘法规则
    - $P(x, y) = P(x)P(y \mid x)$，联合概率=边沿概率*条件概率
  - 贝叶斯定理
    - $P(\theta \mid D) = \frac{P(\theta)P(D|\theta)}{P(D)}$，根据数据估算参数，转换成先验概率和条件概率的计算

先验概率和后验概率

- 先验概率
  - 指基于以往经验和现有知识对其发生概率的判断【以往的知识】
  - 先验概率反映了我们对一个事件发生概率的原始判断，不依赖于当前实验或观察数据
- 后验概率
  - 指在考虑了某一具体事件的新证据后，对该事件发生概率的重新评估
  - 后验概率在观察到新信息后，它反映了在考虑了新证据之后对事件可能性的评估
- $P(\theta \mid D) = \frac{P(\theta)P(D|\theta)}{P(D)}$
  - D表示数据；模型符合某个分布， θ为参数
    - θ是需要计算的变量，$P(\theta)$：先验概率，（数据空间中）代表数据的参数θ的概率分布
  - $P(\theta \mid D)$，根据观测值D，更新θ的估计
  - $P(D \mid \theta)$：似然概率，给定一个参数θ，得到一个具体数据D的概率$P(D \mid \theta)$
  - 根据观测估算模型的问题，转换为，基于先验概率和似然概率的计算

贝叶斯学习(Bayesian learning) – 模型估算

- 记号
  - D表示数据，随机变量θ表示模型参数
  - 先验概率 $P(\theta)$【统计出来的，出现参数为θ的可能性】
  - 条件概率$P(D \mid \theta)$【描述了因果联系。在参数为θ的情况下，出现(观测到)样本D的可能性】
  - 后验概率$P(\theta \mid D)$【在观测到样本D的情况下，(推断/更新)参数为θ的可能性】
- 以模型学习为例
  - 给定观察样本D前提下，每一个参数θ一种可能性$P(\theta \mid D)$
  - 其中最有可能的就是模型参数
    - $\hat{\theta} = \arg\max_{\theta} P(\theta \mid D)$
  - 后验概率$P(\theta \mid D)$通过贝叶斯公式计算
    - $P(\theta \mid D) = \frac{P(\theta)P(D|\theta)}{P(D)}$
  - 即，后验概率 = (似然度 * 先验概率)/标准化常量

参数的极大似然 – 模型估计

- 通过似然函数来考察不同的模型(其参数为θ)下出现(观测到)样本D的可能性
- 似然函数$P(D \mid \theta)$
  - 参数为θ的情况下，出现/观察到样本为D的可能性
- 极大似然估计
  - 模型θ估计准则：出现/观察到样本D的各种θ可能性中， 似然函数$P(D \mid \theta)$最大的那个θ

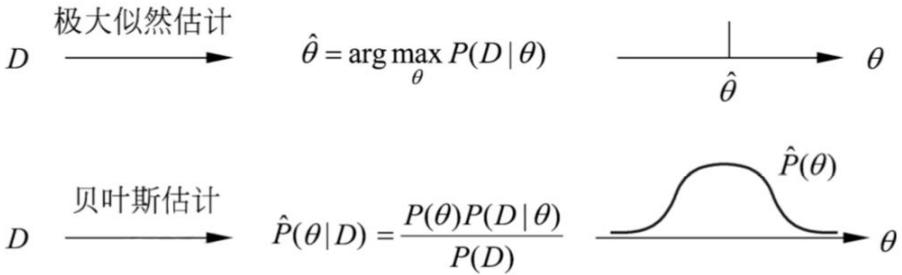
$$\hat{\theta} = \arg\max_{\theta} P(D \mid \theta)$$

  - 【解释】对于生成样本D，每一个参数θ，都有一个可能性（概率$P(D \mid \theta)$），最好的估计参数对应着最大的可能性

- 比较：贝叶斯学习之最大后验概率
  - $\hat{\theta} = \arg\max_{\theta} P(\theta \mid D)$

贝叶斯学习与极大似然估计

- 假设先验分布是均匀分布，贝叶斯学习与极大似然估计
  - 贝叶斯学习，表达式连续，曲线比较光滑
  - 极大似然估计，没有连续表达式
- 极大似然把模型参数当作固定但未知的常数，通过数据寻找最能解释观测的那一个参数值；贝叶斯把参数视为随机变量，结合先验与数据得到参数的概率分布，从而以分布形式表达不确定性并进行推断。



核方法(Kernel method)

- 核方法(Kernel method)
  - 使用核函数表示和学习非线性模型，将线性模型学习方法扩展到非线性模型的学习
- 有一些线性模型的学习方法基于相似度计算，更具体地，向量内积计算。核方法可以把它们扩展到非线性模型的学习，使其应用范围更广泛
  - 定义（ 可以隐式定义 ）从输入空间（低维空间）到特征空间（ 高维空间） 的映射， 在特征空间中进行内积计算

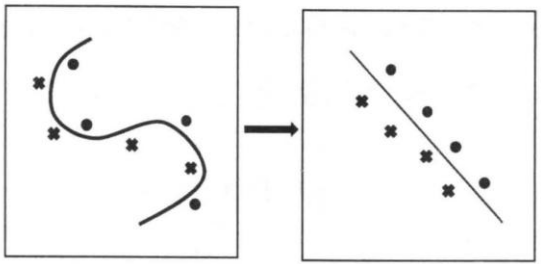


图 1.7　输入空间到特征空间的映射

按算法分类

- 在线学习
- 批量学习

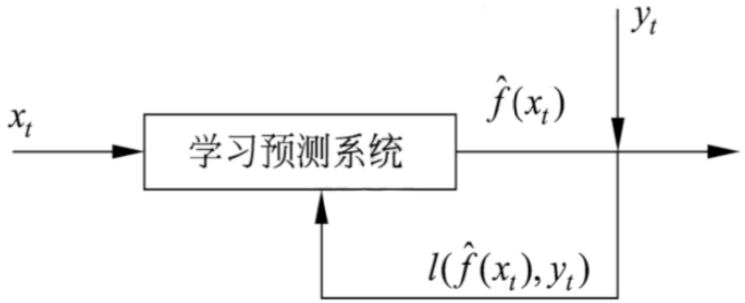
在线学习(online learning)

- 在线学习(online learning)
  - 每次接受一个样本，进行预测，之后学习模型，并不断重复该操作
- 适合场景
  - 数据依次达到无法存储，系统需要及时做出处理；
  - 数据规模很大，不可能一次处理所有数据
  - 数据的模式随时间动态变化，需要算法快速适应新的模式(不满足独立同分布假设)

- 在线学习可以是监督学习，也可以是无监督学习
- 强化学习本身就拥有在线学习的特点
- 利用随机梯度下降的感知机学习算法为在线学习算法

批量学习(batch learning)

- 批量学习(batch learning)
  - 一次接受所有数据，学习模型，之后进行预测



- 在线学习通常比批量学习更难，很难学到预测准确率更高的模型，因为每次模型更新中，可利用的数据有限

3 统计学习三要素

统计学习三要素

- 方法 = 模型 + 策略 + 算法
  - 模型：假定数据符合XX规律，对应模型结构（参数未知）
  - 策略：该模型下学习采用XX准则。对于给的的输入，模型输出和参考答案之间满足的约束
  - 算法：如何找到最优的参数
- 以监督学习为例

模型

- 模型【数据规律的表达】：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，$\mathcal{F} = \{f \mid Y = f(X)\}$

- X和Y是定义在输入空间$\mathcal{X}$ 输出空间$\mathcal{Y}$上的变量
- 如果假设空间定义为决策函数的集合：
  - $\mathcal{F}$为定义在参数空间上的函数族：$\mathcal{F} = \{f \mid Y = f_\theta(X), \theta \in \mathbf{R}^n\}$
  - θ取值于n维欧氏空间$\mathbf{R}^n$，称为参数空间(parameter space)
  - 由决策函数表示的模型，非概率模型
- 如果假设空间可以定义成条件概率的集合：
  - $\mathcal{F}$为定义在参数空间的条件概率分布族：$\mathcal{F} = \{P \mid P_\theta(Y \mid X), \theta \in \mathbf{R}^n\}$
  - 由条件概率表示的模型，概率模型

策略

- 策略：按照什么样的准则来学习或选择最优的模型
  - 统计学习的目标在于从假设空间中选取最优模型

- 引入损失函数来定义模型的选择标准
  - 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

损失函数与风险函数

- 损失函数(loss function)/代价函数(cost function)
  - 度量预测错误的程度
  - $L(Y, f(X))$
- 风险函数(risk function)/期望损失(expected loss)
  - 平均意义下的损失
  - $R_{\exp}(f)$

常见损失函数

- 0-1 损失函数

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

- 平方损失函数(quadratic loss function)

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

- 绝对损失函数(absolute loss function)

$$L(Y, f(X)) = |Y - f(X)|$$

- 对数损失函数(logarithmic loss function) 或对数似然损失函数

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

风险函数/期望损失

- 风险函数(risk function)，期望损失(expected loss)：损失函数的期望
  - $R_{\exp}(f) = E_P\left[L\left(Y, f(X)\right)\right] = \int_{\mathcal{X}\times\mathcal{Y}} L\left(y, f(x)\right)P(x,y)\mathrm{d}x\mathrm{d}y$
- 学习目标：选择期望损失最小的模型
- 监督学习为病态问题
  - 计算$R_{\exp}(f)$需要知道$P(x,y)$，但如果$P(x,y)$已知，则可直接求出条件概率，不需要学习

- 【思路：通过已有数据来推测一般规律】
- 训练集$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$
- 经验风险(empirical risk)，经验损失(empirical loss)
  - 模型$f(X)$关于训练集的平局损失：$R_{\mathrm{emp}}(f) = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, f(x_i)\right)$
- 当样本容量N 趋于无穷时，经验风险$R_{\mathrm{emp}}(f)$趋近于期望风险$R_{\exp}(f)$

经验风险最小化与结构风险最小

- 经验风险最小化 (empirical risk minimization, ERM)
  - 经验风险最小的模型是最优的模型
  - 求解最优化问题：$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L\left(y_i, f(x_i)\right)$
  - 当样本容量足够大时，经验风险最小化能保证有很好的学习效果
  - 当样本容量很小时，效果未必很好，会产生“过拟合over-fitting”现象

经验风险最小化与结构风险最小

- 结构风险最小化(structure risk minimization)
  - 为防止过拟合，等价于正则化(regularization)
  - 加入正则化项(regularizer)，或罚项(penalty term)

$$R_{\text{srm}}(f) = \frac{1}{N}\sum_{i=1}^{N} L\big(y_i, f(x_i)\big) + \lambda J(f)$$

  - 其中$J(f)$为模型的复杂度，是定义在假设空间$\mathcal{F}$上的泛函。模型f越复杂，复杂度$J(f)$就越大；反之，模型f越简单，复杂度$J(f)$就越小
    - 【结构指的是模型的形式，模型的形式/结构越复杂，可靠性越低】

策略

- 求最优模型就是求解最优化问题

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L\left(y_i, f(x_i)\right) + \lambda J(f)$$

算法

- 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
  - 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
- 算法分类
  - 如果最优化问题有显式的解析式
  - 但通常解析式不存在，需要数值计算的方法
- 如何保证找到全局最优解，并使求解的过程非常高效，为统计学习的重要问题

4 模型评估与模型选择

训练误差与测试误差

- 统计学习的目的
  - 使学到的模型$Y = \hat{f}(X)$对已知数据(训练集)及未知数据(测试集及真实数据)都有很好的预测能力
- 训练误差，模型关于训练数据集的平均损失

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

- 测试误差，模型关于测试数据集的平均损失

$$e_{\text{test}} = \frac{1}{N'}\sum_{i=1}^{N'} L\left(y_i, \hat{f}(x_i)\right)$$

- 损失函数是0-1 损失时
  - 测试误差就变成了常见的测试数据集上的误差error rate：$e_{\text{test}} = \frac{1}{N'}\sum_{i=1}^{N'} I\left(y_i \neq \hat{f}(x_i)\right)$
  - 测试数据集的准确率(accuracy)：$r_{\text{test}} = \frac{1}{N'}\sum_{i=1}^{N'} I\left(y_i = \hat{f}(x_i)\right)$
  - $r_{\text{test}} + e_{\text{test}} = 1$，其中$I$是指示函数(indicator function)
- 通常将学习方法对未知数据的预测能力称为泛化能力(generalization ability)

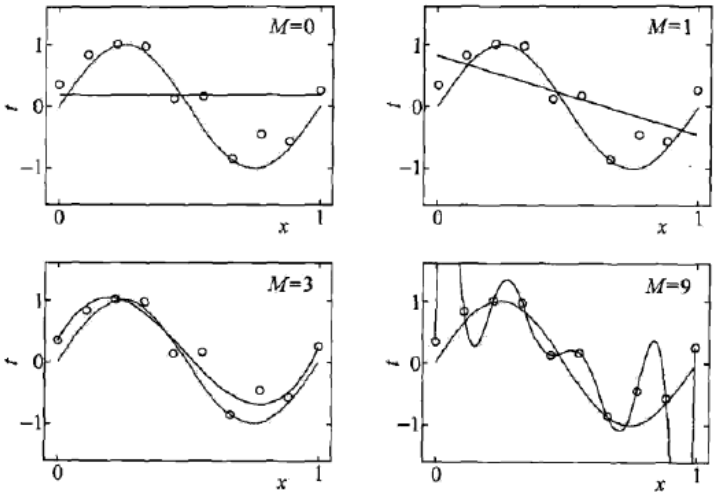
过拟合与模型选择

- 当假设空间含有不同复杂度(如，不同的参数个数)的模型时，面临模型选择(model selection) 问题
- 选择合适模型
  - 如果在假设空间中存在“真”模型，那么所选择的模型应该逼近真模型
  - 参数个数相同，数值接近
- 过拟合，指学习时选择的模型所包含的参数过多，比“真”模型复杂度更高【凑】
  - 对已知数据预测得很好，但对未知数据预测得很差
  - 经验风险最小

  - 下一页…例子

过拟合与模型选择

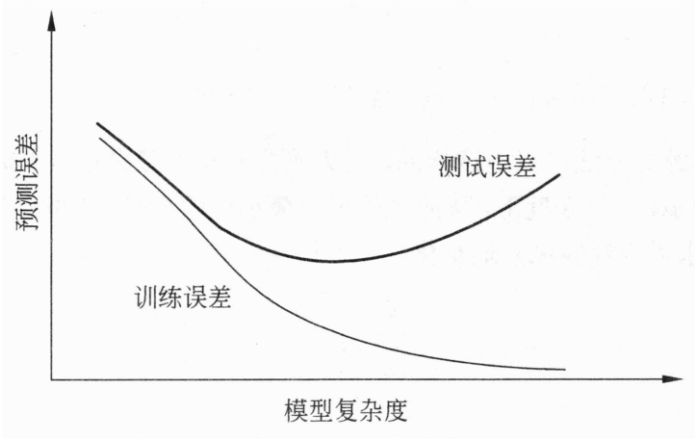
- 假设给定训练数据集$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$，在M次多项式函数中选择一个对已知数据以及未知数据都有很好预测能力的函数
  - M：函数的阶数，模型的复杂度
  - M次多项式：$f_M(x, w) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$
  - 经验风险最小化：$L(w) = \frac{1}{2} \sum_{i=1}^{N} (f(x_i, w) - y_i)^2$



训练误差和测试误差与模型复杂度的关系

- 当模型的复杂度增大时
  - 训练误差会逐渐减小并趋向于0
  - 测试误差会先减小，后又增大

- 当选择的模型复杂度过大时，过拟合现象可能发生
  - 训练误差变小，但测试误差变大



5 正则化与交叉验证

正则化

- 模型选择的典型方法是正则化(regularization)
  - 结构风险最小化策略的实现
  - 在经验风险上加一个正则化项(regularizer) 或罚项(penalty term)
  - 正则化项一般是模型复杂度的单调递增函数，模型越复杂，正则化值就越大
- 正则化形式：

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L\left(y_i, f(x_i)\right) + \lambda J(f)$$

  - 正则化项$J(f)$的形式：L_2范数，L_1范数
- 奥卡姆剃刀(Occam ‘s razor) 原理，应用于模型选择时
  - 在所有可能选择的模型中，能够很好地解释已知数据，并且十分简单才是最好的模型
  - 为一个第一性的科学假设原理
  - 精简表达：如无必要，勿增实体

交叉验证

- 交叉验证(cross validation)：一种常用的模型选择方法
- 如果给定的样本数据充足，随机地将数据集切分成三部分，选择对验证集有最小预测误差的模型
  - 训练集 training set，用于训练模型
  - 验证集 validation set，用于模型选择
  - 测试集 test set，用于最终对学习方法的评估

- 数据不充足时，采用交叉验证选择模型：重复地使用数据
  - 简单交叉验证
    - 随机地将已给数据分为两部分， 一部分作为训练集，另一部分作为测试集
  - S折交叉验证
    - 首先随机地将已给数据切分为S 个互不相交、大小相同的子集；然后利用S-1 个子集的数据训练模型，利用余下的子集测试模型；将这一过程对可能的S 种选择重复进行
  - 留一交叉验证

6 泛化能力

泛化误差

- 学习方法的泛化能力(generalization ability)
  - 学习到的模型对未知数据的预测能力

- 泛化能力评估
  - 通过测试误差来评价学习方法的泛化能力。依赖于测试数据集

- 泛化误差 generalization error

$$R_{\exp(\hat{f})} = E_P\left[L\left(Y, \hat{f}(X)\right)\right] = \int_{\mathcal{X}\times\mathcal{Y}} L\left(y, \hat{f}(x)\right) P(x,y)\mathrm{d}x\mathrm{d}y$$

泛化误差上界

- 泛化误差上界(generalization error bound)
  - 通过比较两种学习方法的泛化误差上界的大小来比较它们的优劣

- 泛化误差上界通常
  - 它是样本容量的函数，当样本容量增加时，泛化上界趋于0
  - 它是假设空间容量(capacity) 的函数，假设空间容量越大，模型就越难学，泛化误差上界就越大

7 生成模型与判别模型

生成模型与判别模型

- 概率模型（Probabilistic Models）
  - 概率模型，也称为生成模型（Generative Models）。通过学习数据的联合分布来建立模型
  - 不仅学习输入数据X到输出标签Y的映射关系，而且还试图理解数据背后的生成过程。
- 判别模型（Discriminative Models）
  - 学习从输入空间到输出空间的映射关系，即条件概率分布P(Y|X)，而不试图理解数据背后的生成过程
  - 决策函数$Y = f(X)$，判别方法

- 学习识别猫的图片
  - 生成模型（艺术家）：它的目标是学习猫的“本质”是什么样的。它会观察成千上万张猫的图片，学习猫的毛发、眼睛、胡须、轮廓等所有特征的分布和组合方式。学成之后，它不仅能判断一张新图片是不是猫，甚至还能自己画出一只全新的、看起来很逼真的猫。它学习的是“猫长什么样”。
  - 判别模型（裁判）：它的目标是学习如何区分“猫”和“非猫”。它会观察猫和非猫（比如狗、桌子）的图片，并专注于寻找它们之间的差异和分界线。学成之后，它能高效地判断一张新图片是猫还是非猫，但它自己完全画不出一只猫。

生成模型与判别模型

- 生成方法(Generative approach)
  - 由数据学习联合概率分布$P(X,Y)$, 然后贝叶斯公式求出条件概率分布$P(Y|X)$作为预测的模型
    - 朴素贝叶斯法和隐马尔科夫模型
- 判别方法(Discriminative approach)
  - 直接学习决策函数$Y = f(X)$或条件概率分布$P(Y|X)$作为预测的模型
  - 判别方法关心：对给定的输入X, 应该预测什么样的输出Y
    - K近邻法、感知机、决策树、logistic回归模型、最大熵模型、支持向量机、提升方法和条件随机场

生成模型与判别模型

- 学习目标不同
  - 概率模型学习数据的整体分布，包括输入和输出之间的关系以及数据生成过程；判别模型直接学习从输入到输出的映射关系。
- 应用场景不同
  - 概率模型适用于需要理解数据生成过程或需要生成新数据实例的场景，如文本生成、语音识别等；判别模型适用于分类和回归任务，关注于准确预测输出。
- 性能和效率
  - 判别模型通常在预测任务上更为高效和准确，因为它们专注于直接学习输入到输出的关系；而概率模型在处理有关数据生成过程的复杂问题时更有优势。
- 其它
  - 当存在隐变量时，仍可以用生成方法学习，而判别方法不能用

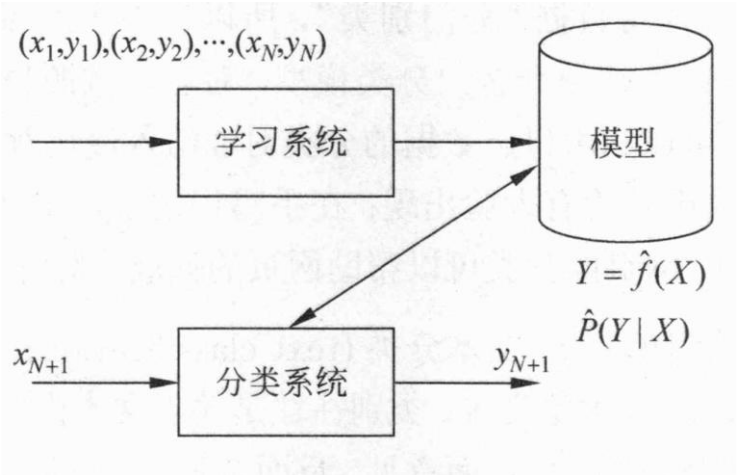
8 监督学习应用

监督学习

- 分类
- 标注
- 回归

分类问题

- 在监督学习中，当输出变量Y 取有限个离散值时，预测问题便成为分类问题
- 分类器(classifier)：监督学习从数据中学习到的一个分类模型或分类决策函数
- 分类(classification)：分类器对新的输入进行输出的预测



分类问题

- 评价分类器性能的指标一般是分类准确率(accuracy),
  - 对于给定的测试数据集，分类器正确分类的样本数与总样本数之比
- 二分类评价指标，将正类/负类预测正确/不正确的情况
  - TP true positive
  - FN false negative
  - FP false positive
  - TN true negative
  - 形式：【预测对/错+预测成正/负】

		预测类别		
		Yes	No	总计
实际类别	Yes	TP	FN	P（实际为Yes）
	No	FP	TN	N（实际为No）
	总计	P'（被分为Yes）	N'（被分为No）	P+N

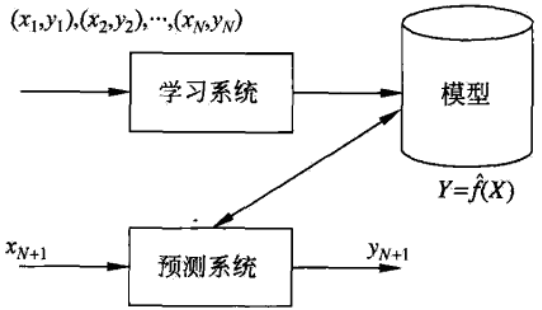
- 精确率(查准率)：正类预测的预测TP + FP的正确率(正确的正类预测TP)
  - $P = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$
- 召回率(查全率)：真实的正类TP + FN中，预测正确TP的比率
  - $R = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$
- F_1：精确率和召回率的调和均值 $\frac{2}{F_1} = \frac{1}{P} + \frac{1}{R}$

回归问题

- 回归模型
  - 表示从输入变量到输出变量之间映射的函数
  - 回归问题的学习等价于函数拟合
- 学习和预测两个阶段

- 回归学习最常用的损失函数
  - 平方损失函数，在此情况下，回归问题可以由 著名的最小二乘法(least squares)求解

- 股价预测



标注问题

- 标注(tagging)问题是分类问题的一个推广， 标注问题又是更复杂的结构预测问题的简单形式
  - 输入：观测序列， 输出：标记序列或状态序列
- 标注问题分为学习和标注两个过程
- 训练集： $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$
  - 观测序列： $x_i=\left(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(n)}\right)^{\mathrm{T}}$
  - 对应输出标记序列： $\mathrm{y}_i=\left(y_i^{(1)},y_i^{(2)},\cdots,y_i^{(n)}\right)^{\mathrm{T}}$
- 模型为条件概率分布： $P\left(Y^{(1)},Y^{(2)},\cdots,Y^{(n)}\mid X^{(1)},X^{(2)},\cdots,X^{(n)}\right)$
  - $X^{(i)}(i=1,2,\cdots,n)$ 取值为所有可能的观测， $Y^{(i)}(i=1,2,\cdots,n)$取值为所有可能的标记
- 标注常用的统计学习方法有：隐马尔可夫模型、条件随机场
- 标注问题在信息抽取、自然语言处理等领域被广泛应用，是这些领域的基本问题