机器学习简介

数据建模与分析

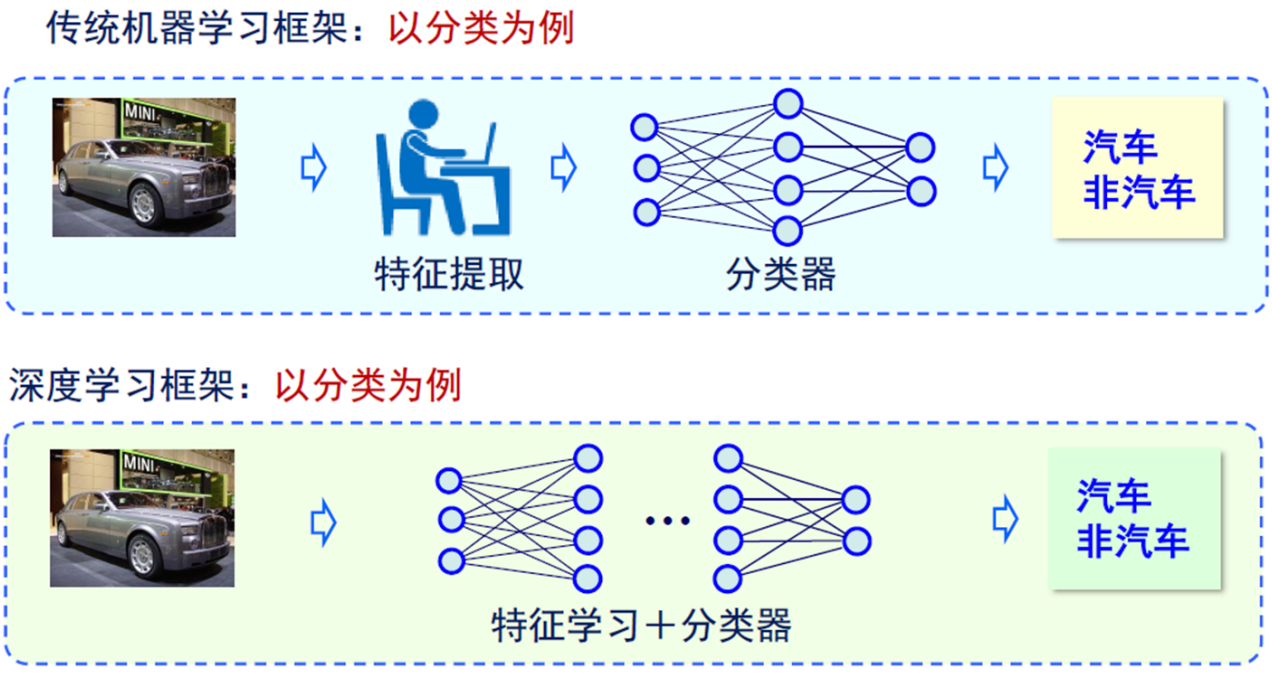
机器学习vs数据挖掘vs模式识别：

- 机器学习：侧重于学习模型的构建，强调从数据中学习，并最大化某个学习目标
- 数据挖掘：如何从无监督数据中发现未知的知识，强调其商业应用
- 模式识别：强调描述、解释和可视化一个特定的模式

大数据(4V)：

- 数据量大(Volume)
- 数据种类多样(Variety)
- 数据价值巨大(Value)（价值巨大但价值密度低
- 高实时性(Velocity)

深度学习是以深度神经网络为假设空间的机器学习方法。深度学习与经典机器学习方法的核心区别在于不同的假设空间，因而对深度学习的理论解释及其运行机理的探索可转向于研究深度神经网络相较传统浅层假设空间的优越性。



机器学习vs.深度学习

统计机器学习是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预测与分析的一门学科

- 特点：
  - 以数据驱动为主的学科

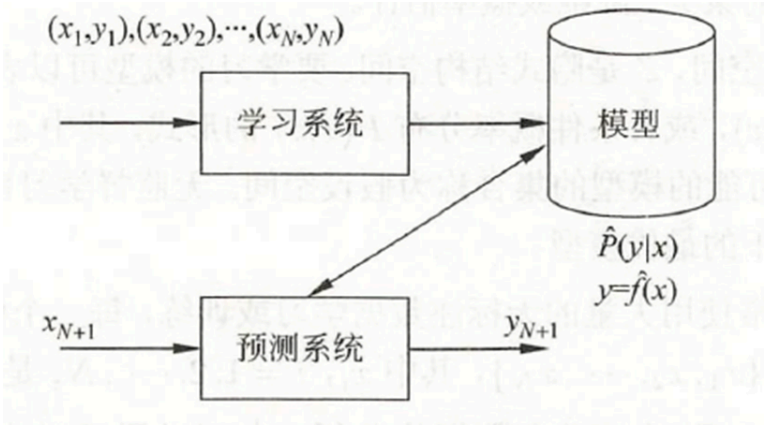
    - 中心任务是构建学习模型与方法，并由此对数据进行预测与分析
- 对象：
    - 数据：计算机及互联网上的各种数字、文字、图像、音视频数据以及它们的组合
    - 数据的基本假设是同类数据具有一定的统计规律性
- 目的：
    - 对数据进行预测和分析

监督学习：从标注数据中学习预测模型。标注数据表示输入与输出的对应关系，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与输出的映射的统计规律。

- 训练样数据集：$T=\{(x_1,y_1),(x_2,y_2),\dots,(x_N,y_N)\}$
- 输入空间：$\mathcal{X}\subset\mathbb{R}^n, x_i\in\mathcal{X}$
- 输出空间：$\mathcal{Y}, y_i\in\mathcal{Y}$
- 实例（instance）：$x_i=(x_i^{(1)},x_i^{(2)},\dots,x_i^{(n)})^T$
- 特征（feature）：x_i^n 表示第 n 个特征

假设输入与输出的随机变量 X 和 Y 遵循联合概率分布 $P(X,Y)$，对于学习系统来说，联合概率分布是未知的，通常假设训练数据和测试数据被看作是依联合概率分布 $P(X,Y)$ 独立同分布产生的。

监督学习目的是学习一个由输入到输出的映射，称为模型，模型的结合就是假设空间（hypothesis space）

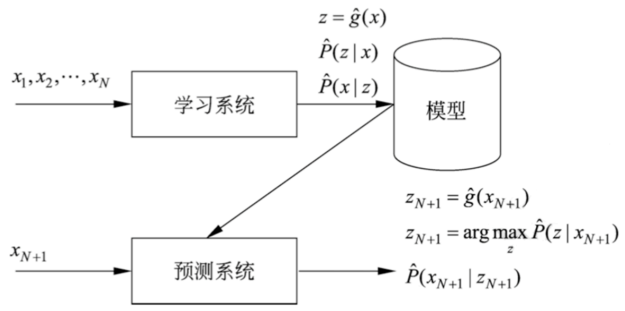


问题形式化

学习系统（学习算法）试图通过训练样本数据集中学习模型。学习系统通过不断尝试，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尽可能好的推广。

$$y_{N+1}=\arg\min_y \hat{P}(y|x_{N+1})$$

无监督学习：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、转换或者概率。无监督学习的本质是学习数据的统计规律或者潜在结构



半监督学习

半监督学习(semi supervised learning)：

- 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。

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。强化学习的马尔可夫决策过程是状态、奖励、动作序列上的随机过程，由五元组 (S, A, P, r, γ) 组成

- S 是有限状态(state)的集合
- A 是有限动作(action)的集合
- P 是状态转移概率(transition probability)函数：

$$P(s'|s,a) = P(s_{t+1} = s'|s_t = s, a_t = a)$$

- r 是奖励函数(reward function):$r(s,a) = E(r_{t+1}|s_t = s, a_t = a)$
- γ 是衰减系数(discount factor):$\gamma \in [0,1]$

在每一步 t，智能系统从环境中观测到一个状态 s_t 与一个奖励 r_t 采取一个动作 a_t 环境根据智能系统选择的动作，决定下一步 $t+1$ 的状态 s_t 与奖励 r_t

统计机器学习三要素：模型、策略、算法

模型：

- 决策函数的集合：$\mathcal{F} = \{f|Y = f(X)\}$
- 参数空间：$\mathcal{F} = \{f|Y = f_\theta(X), \theta \in \mathbb{R}^n\}$
- 条件概率的集合：$\mathcal{F} = \{P|P(Y|X)\}$
- 参数空间：$\mathcal{F} = \{P|P_\theta(Y|X), \theta \in \mathbb{R}^n\}$

损失函数：一次预测的好坏 $L(Y, f(X))$ 或者 $L(Y, P(Y|X))$，损失函数值越小，模型越好

策略：经验风险最小化与结构风险最小化

经验风险最小化：

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

- 当样本容量很大时，经验风险最小化能够保证有较好的学习效果。
- 最大似然估计是一个很好的经验风险最小化实例，其所使用的模型是一个条件概率分布函数。
- 当样本容量很小时，经验风险最小化学习效果未必好，可能产生过拟合。

结构风险最小化

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

为了防止过拟合提出的策略，等价于正则化。正则化项 $\ell(f)$ 描述模型的复杂度，是定义在假设空间上的泛函。正则化参数 $\lambda\geq 0$ 用于权衡经验风险和模型复杂度

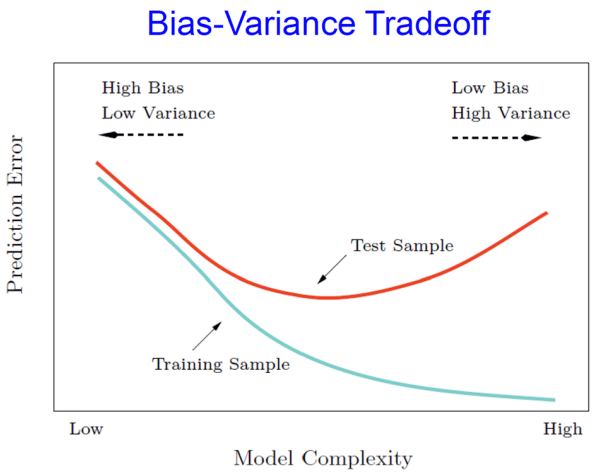
算法：求最优模型就是求解最优化问题

$$\min\min_{f\in\mathcal{F}}\frac{1}{N}\sum_{i=1}^{N}L(y_i,f(x_i))$$

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

过拟合：是指过于紧密或精确地匹配特定数据集，以致于无法良好地拟合其他数据或预测未来的观察结果的现象。过拟合模型指的是相较有限的数据而言，参数过多或者结构过于复杂的统计模型。发生过拟合时，模型的偏差小而方差大。

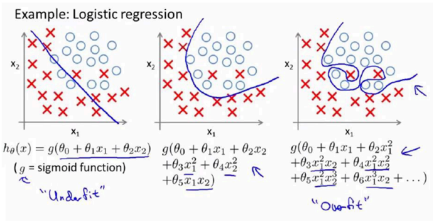
欠拟合：它是指相较于数据而言，模型参数过少或者模型结构过于简单，以至于无法捕捉到数据中的规律的现象。发生欠拟合时，模型的偏差大而方差小。



T. Hastie, T. Tibshrani, J. Friedman (2001)
The Elements of Statistical Learning.

分类器（模型）复杂度对泛化性能的影响

- 训练数据不变的情况下，分类器越复杂，对训练数据拟合程度越高
- 过拟合情况下，泛化性能会下降



过拟合的例子，来自www.verydemo.com

泛化能力：指该方法学习得到的模型对未知数据的预测能力，是学习方法本质上重要的性质。

泛化误差：若学到的模型为 $\hat{f}$，则这个模型对未知数据的预测的误差即为泛化误差 (generalization error)，也即模型的期望风险：

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

Theorem

对二类分类问题，当假设空间是有限个函数的集合 $\mathcal{F} = \{f_1, f_2, \ldots, f_d\}$ 时，对任意一个函数 $f \in \mathcal{F}$，至少以概率 $1-\delta, 0 < \delta < 1$，以下不等式成立：

$$R(f) \leq \hat{R}(f) + \varepsilon(d, N, \delta)$$

其中

$$\varepsilon(d, N, \delta) = \sqrt{\frac{1}{2N}\left(\log d + \log\frac{1}{\delta}\right)}$$

泛化误差界与假设空间的容量有关，假设空间越复杂，误差界越大。

- 集中不等式：描述了一个随机变量是否集中在某个取值附近，例如大数定律说明了一系列独立同分布随机变量的平均值在概率上趋近于它们的数学期望。

Hoeffding 不等式

设 $X_1, X_2, \ldots, X_n$ 是随机独立变量，且 $X_i \in [a_i, b_i], i = 1, 2, \ldots, N$; $\bar{X}$ 是 $X_1, X_2, \ldots, X_N$ 的经验均值，即 $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ ，则对任意 $t > 0$ ，以下不等式成立：

$$\mathbb{P}[\bar{X} - E(\bar{X}) \geq t] \leq \exp\left(-\frac{2N^2t^2}{\sum_{i=1}^{N}(b_i - a_i)^2}\right)$$

$$\mathbb{P}[|\bar{X} - E(\bar{X})| \geq t] \leq 2\exp\left(-\frac{2N^2t^2}{\sum_{i=1}^{N}(b_i - a_i)^2}\right)$$

Bennet 不等式

Let $\{\xi_i\}_{i=1}^{m}$ be independent random variables on probability space Z with means $\{\mu_i\}$ and variances $\{\sigma_i^2\}$ .Set $\Sigma^2 := \sum_{i=1}^{m}\sigma_i^2$. If for each i there holds $|\xi_i - \mu_i| \leq M$ almost everywhere, then for every $\varepsilon > 0$ we have

$$\mathbb{P}\left\{\sum_{i=1}^{m}[\xi_i - \mu_i] > \varepsilon\right\} \leq \exp\left\{-\frac{\varepsilon}{M}\left[\left(1 + \frac{\Sigma^2}{M\varepsilon}\right)\log\left(1 + \frac{M\varepsilon}{\Sigma^2}\right) - 1\right]\right\}$$

进一步放缩可得 Bernstein 不等式

Bernstein 不等式

Let $\{\xi_i\}_{i=1}^m$ be independent random variables on probability space Z with means $\{\mu_i\}$ and variances $\{\sigma_i^2\}$ and satisfying $|\xi_i(z) - \mathbb{E}(\xi_i)| \geq M$ for each i and almost all $z \in Z$ .Set $\Sigma^2 := \sum_{i=1}^m \sigma_i^2$. Then for every $\varepsilon > 0$ we have

- Generalized Bennett

$$\mathbb{P}\left\{\sum_{i=1}^m [\xi_i - \mu_i] > \varepsilon\right\} \leq \exp\left\{-\frac{\varepsilon}{2M}\log\left(1 + \frac{M\varepsilon}{\Sigma^2}\right)\right\}$$

- Bernstein

$$\mathbb{P}\left\{\sum_{i=1}^m [\xi_i - \mu_i] > \varepsilon\right\} \leq \exp\left\{-\frac{\varepsilon^2}{2\left(\Sigma^2 + \frac{1}{3}M\varepsilon\right)}\right\}$$