第1章　绪论

刘家锋

哈尔滨工业大学

第1章 绪论

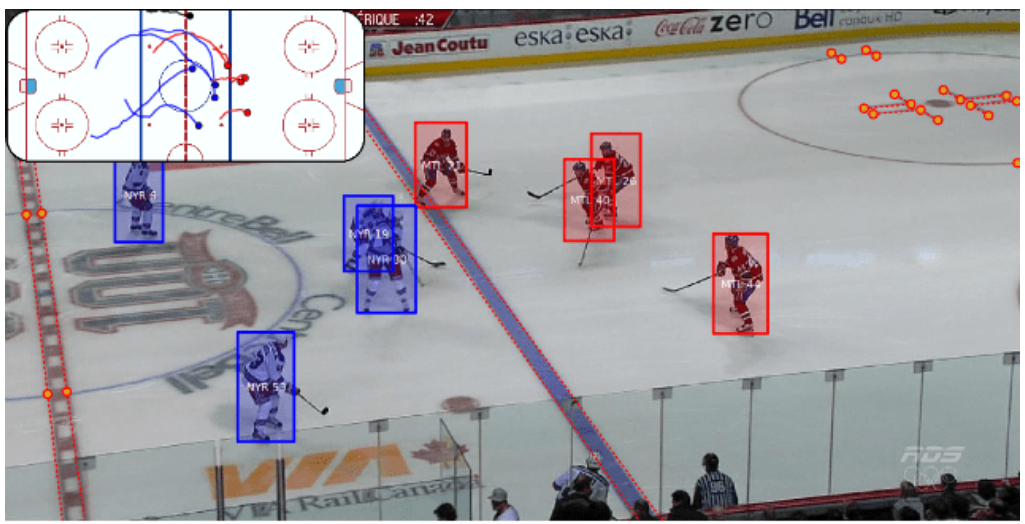
1.1 什么是机器学习？

机器学习

- **机器学习与人工智能**
  - 机器学习是人工智能的一个分支，并且可以说是最重要的一个组成部分；
  - 机器学习模拟的是人类智能中的归纳推理能力：
- **机器学习的定义**
  - 机器学习是智能体利用经验改善自身性能的计算方法；
  - 机器学习的目的是自动地从数据中发现内在的规律、模式，并能够对未来的数据做出预测；

机器学习与其它

- **机器学习与图像处理、计算机视觉**
  - 图像处理：输入和输出均为图像，主要关注图像本身而非图像内容；
  - 计算机视觉：从二维图像感知三维世界，关注图像内的物体以及场景之间的关系；
  - 机器学习：视觉是一类数据来源，更关注图像的内容和内在的规律；



机器学习与其它

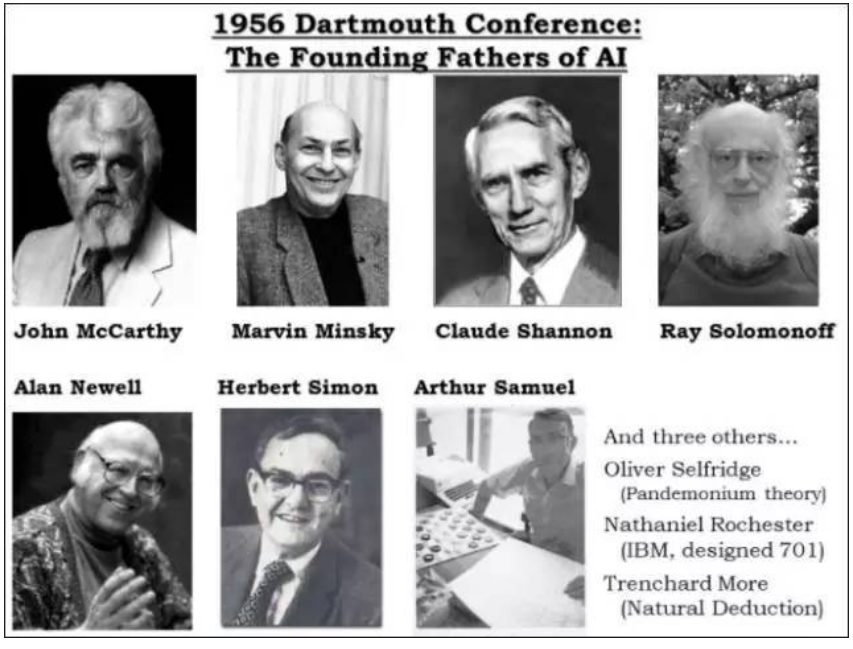
- **机器学习与自然语言处理**
  - 自然语言处理研究人与计算机之间用自然语言进行有效通信的各种理论和方法；
  - 人类的多种智能都与语言有着密切的关系，自然语言处理致力于让计算机能够理解人类的语言和文字；
  - 自然语言处理会使用大量机器学习的方法，实现对语言、文字的理解和处理；

机器学习与其它

- **机器学习与模式识别**
  - 两者的研究内容和方法有很大的相关性；
  - 模式识别起源于控制和信号处理领域，研究的是对数据的分类问题；
  - 机器学习起源于计算机、人工智能领域，既研究分类问题，也研究回归问题；
- **机器学习与数据挖掘**
  - 机器学习 + 数据库 $\longrightarrow$ 数据挖掘
  - 机器学习提供数据分析技术，数据库提供数据管理技术；

人工智能的发展历程

- **人工智能的起源**
  - 1956年，美国达特茅斯学院；
  - 达特茅斯会议标志着人工智能这一学科的诞生；

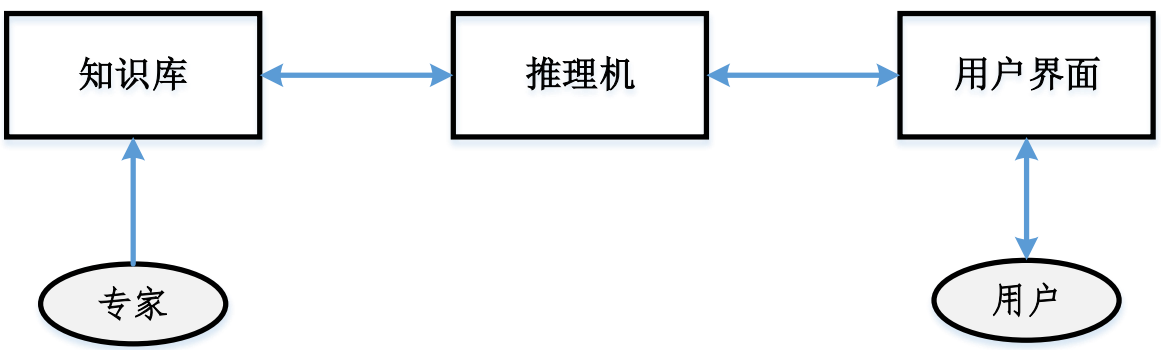


人工智能的发展历程

- **第一阶段：推理期(1956-1960年代)**
  - 希望用数学的方式模型化人类的智能，成功地让计算机具有了逻辑推理的能力；
  - 主要成就是西蒙与纽厄尔的自动定理证明系统，希望能够实现通用的问题求解器；
  - 这一时期的研究者发现，逻辑推理并不是人类智能的全部；

人工智能的发展历程

- **第二阶段：知识期(1970年代-1980年代)**
  - 认为知识是智能的基础，希望计算机能够掌握和利用更多的知识进行推理；
  - 主要成就是专家系统，费根鲍姆等人的DENDRAL系统；
  - 这一时期的研究者发现，总结人类的所有知识“教”给计算机几乎是不可能的；



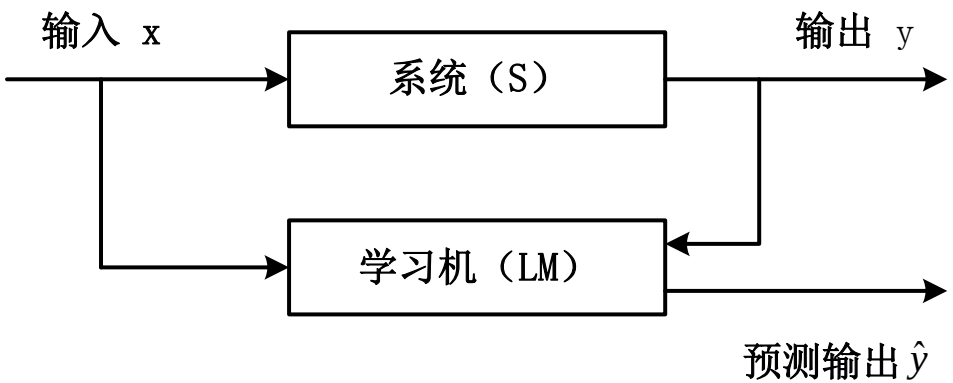
人工智能的发展历程

- **第三阶段：学习期(1990年代至今)**
  - 希望计算机能够自己“学会”知识，用于解决问题；
  - 信息技术的发展和互联网的出现，人类发现自己淹没在数据的海洋中；
  - 迫切需要对数据的自动分析，挖掘数据背后的规律和知识；

1.2 机器学习方法

机器学习的一般过程

- **学习的过程**
  - 机器学习是归纳推理，从特殊的示例归纳出一般的知识；
  - 系统S为学习对象，输入$\mathbf{x}$时输出y；
  - 学习机LM从一系列的示例中归纳出系统S的规律；
  - 使得输入$\mathbf{x}$时，LM的输出$\hat{y}$能够尽量准确地预测S的输出y；



基本术语

- **输入数据$\mathbf{x}$**
  - 称为示例(instance)或样本(sample)，表示为向量的形式；
  - $\mathbf{x}$的元素称为属性或特征，反映对象某方面的表现或性质；
  - 例如为了挑西瓜，可以观察3方面属性：色泽，根蒂，敲声
- **输出数据y**
  - 希望预测的结果；
  - 结果为离散值称为分类问题，例如：$y \in \{\text{好瓜}, \text{坏瓜}\}$
  - 结果为连续值称为回归问题，例如：$y \in [0, 1]$表示成熟度
- **学习的模型LM**
  - 总结归纳出根据输入$\mathbf{x}$预测输出y的潜在规律；

基本术语

- **训练集**
  - 对系统S的一系列观测，每一次观测是一个输入-输出对：

$$D = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_m, y_m)\}$$

  - 训练集是归纳、学习的依据，例如西瓜的训练数据表示为：

$\mathbf{x}$	色泽	根蒂	敲声	y
$\mathbf{x}_1$	青绿	蜷缩	浊响	好瓜
$\mathbf{x}_2$	乌黑	蜷缩	浊响	好瓜
$\mathbf{x}_3$	青绿	硬挺	清脆	坏瓜
$\mathbf{x}_4$	乌黑	稍蜷	沉闷	坏瓜

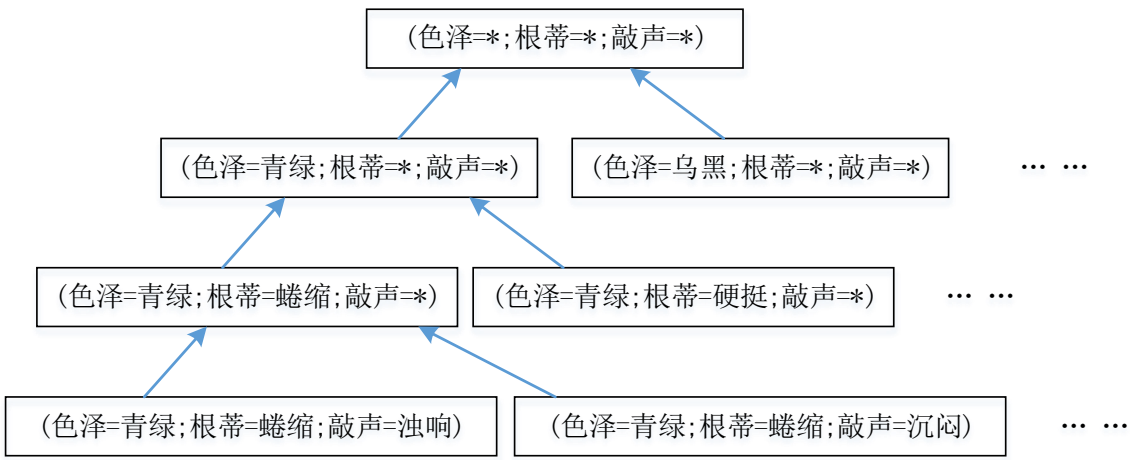
- **测试数据**
  - 有了一个新的观测$\mathbf{x} = (\text{青绿}, \text{蜷缩}, \text{沉闷})$，如何预测输出$\hat{y}$?

基本术语

- **假设和假设空间**
  - 形如下式的布尔表达式可以称为一个假设：

$$(\text{色泽}=?) \wedge (\text{根蒂}=?) \wedge (\text{敲响}=?) \rightarrow \text{好瓜}$$

  - 所有的假设组成了假设空间，机器学习就是在假设空间中搜索与训练集“匹配”的假设；



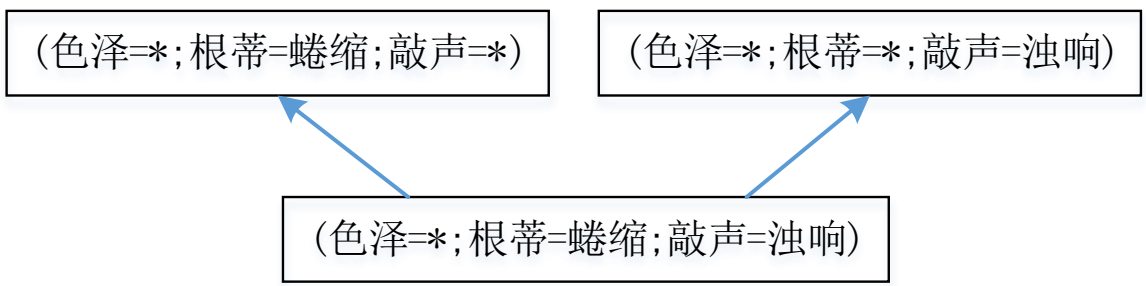
每个属性3种取值，假设空间大小$4 \times 4 \times 4 + 1 = 65$

机器学习需要解决的问题

- **假设的模型化**
  - 离散属性离散输出问题，可以用布尔表达式来模型化假设；
  - 连续属性或连续输出问题，需要其它数学模型来描述假设；
- **假设空间的搜索**
  - 假设空间的规模可能非常大，甚至会包含无穷多个假设；
  - 需要有效的方法来搜索假设空间，不能找到最优假设的时候，可以接受次优解；
- **模型的选择**
  - 假设空间中可能存在多个与训练集“匹配”的解，应该选择哪一个作为学习结果？
  - 模型的选择常常与具体问题相关，与人的先验知识相关；

版本空间与归纳偏好

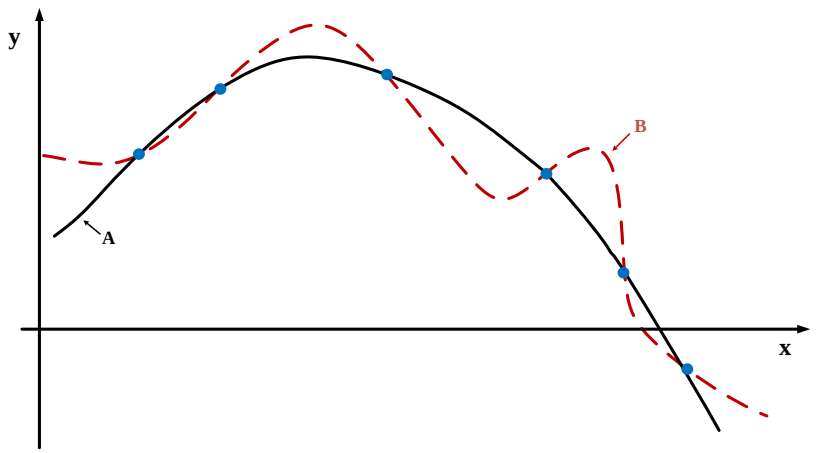
- **版本空间(version space)**
  - 与训练集一致的所有假设的集合称为版本空间;
  - 版本空间中的所有假设，都可以正确地预测训练集的示例;
  - 但对新的未见示例会产生不同的输出结果;



选择哪个假设预测(青绿;蜷缩;沉闷)?

模型选择

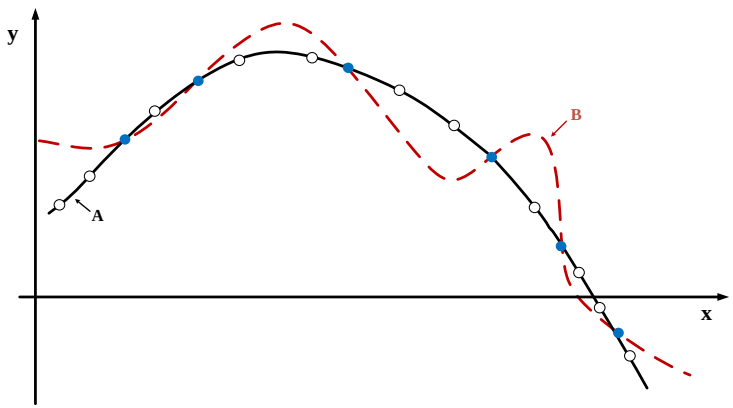
- **归纳偏好需要与问题匹配**
  - 回归问题中，同样一组训练数据可以用不同的函数拟合；
  - 奥卡姆剃刀(Ocam's razor)原理：若有多个假设与观察一致，则选最简单的那个



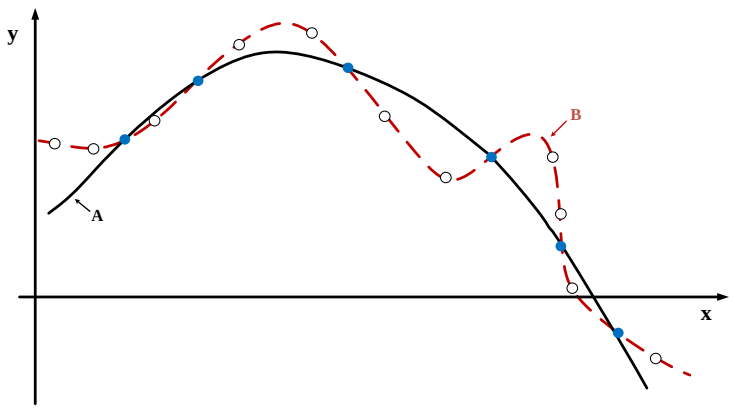
A曲线和B曲线哪一个更好？

模型选择

- **归纳偏好需要与问题匹配**
  - 没有免费的午餐定理(No Free Lunch Theorem)：一个算法$\mathfrak{L}_a$若在某些问题上比另一个算法$\mathfrak{L}_b$好，必存在另一些问题，$\mathfrak{L}_b$比$\mathfrak{L}_a$好



A优于B



B优于A

蓝点:训练样本，白点:测试样本

1.3 模型评估与选择

学习结果的评估

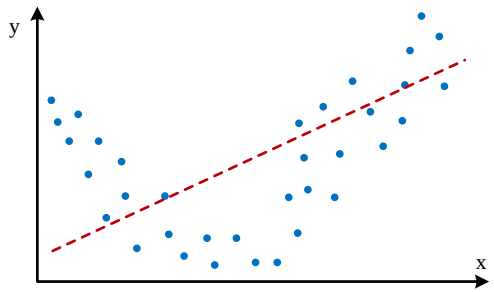
- **评估的目的**
  - 检验结果是否满足设计的要求；
  - 在多个假设中选择一个最优的；
  - 学习算法的参数调整；
- **评估的依据**
  - 经验误差：在训练集上的预测误差，也称为“训练误差”；
  - 泛化误差：在“未来”的测试样本上的预测误差；
  - 泛化误差的大小是评估学习结果的最终指标，但“未来”的样本在学习时并没有，泛化误差是无法直接计算的；
  - 是否可以用经验误差代替泛化误差？

学习结果的评估

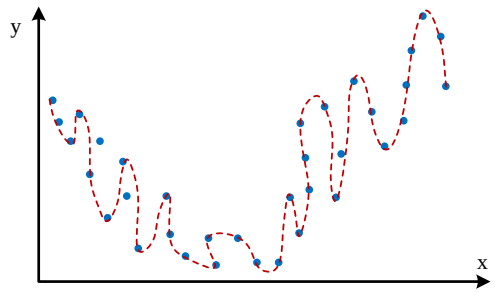
- **独立同分布假设**
  - 训练误差在一定程度上是可以代替泛化误差的，因为在机器学习中一般假设训练数据与未来的测试数据是独立同分布的(i.i.d, independent and identically distributed)；
  - 独立同分布假设：每一个训练数据和未来的测试数据都来自于同一个样本空间，服从同一个未知的分布；
  - 一般来说，训练样本越多，关于未知分布的信息越多，机器学习越有可能获得具有强泛化能力的模型；
- **经验误差与泛化误差**
  - 训练样本是有限的，经验误差不能绝对地代替泛化误差；
  - 片面地优化训练误差，有可能导致“过拟合”的出现；
  - 模型只能预测训练数据，无法很好地预测未来的测试数据；

学习结果的评估

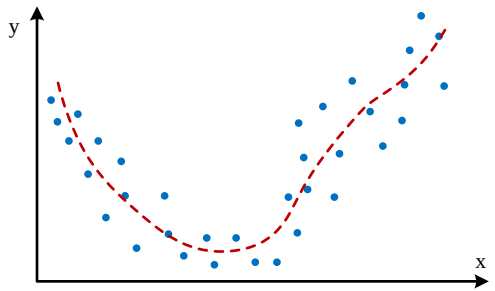
- **欠拟合与过拟合**
  - 欠拟合：学习的太差，训练样本的一般性质尚未学好；
  - 过拟合：能力过强，学到了训练样本中的特有特性；



欠拟合



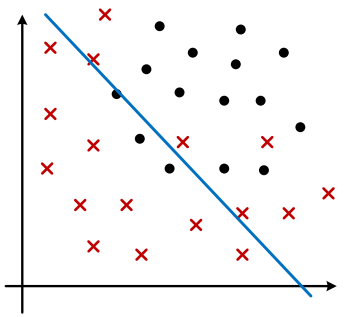
过拟合



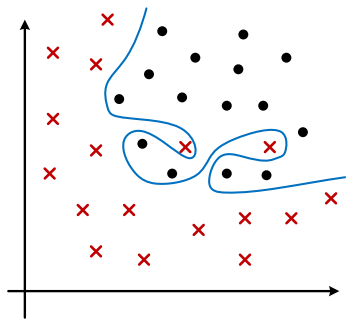
适合的

学习结果的评估

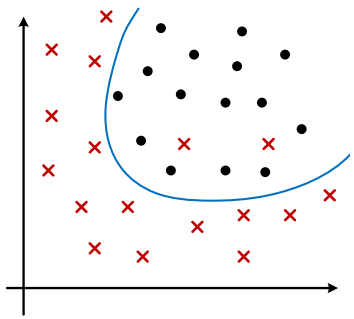
- **欠拟合与过拟合**
  - 分类问题中同样存在欠拟合和过拟合问题;



欠拟合



过拟合



适合的

评估的方法

- **训练集和测试集**
  - 最好的评估方法是使用训练集S学习模型，使用另外的测试集T评估模型的性能；
  - 数据集D既用于训练，又用于测试，需要适当的划分：

$$D = S \cup T, \qquad S \cap T = \Phi$$

- **常用的方法**
  - 留出法(hold-out)；
  - 交叉验证法(cross validation)
  - 自助法(bootstrap)

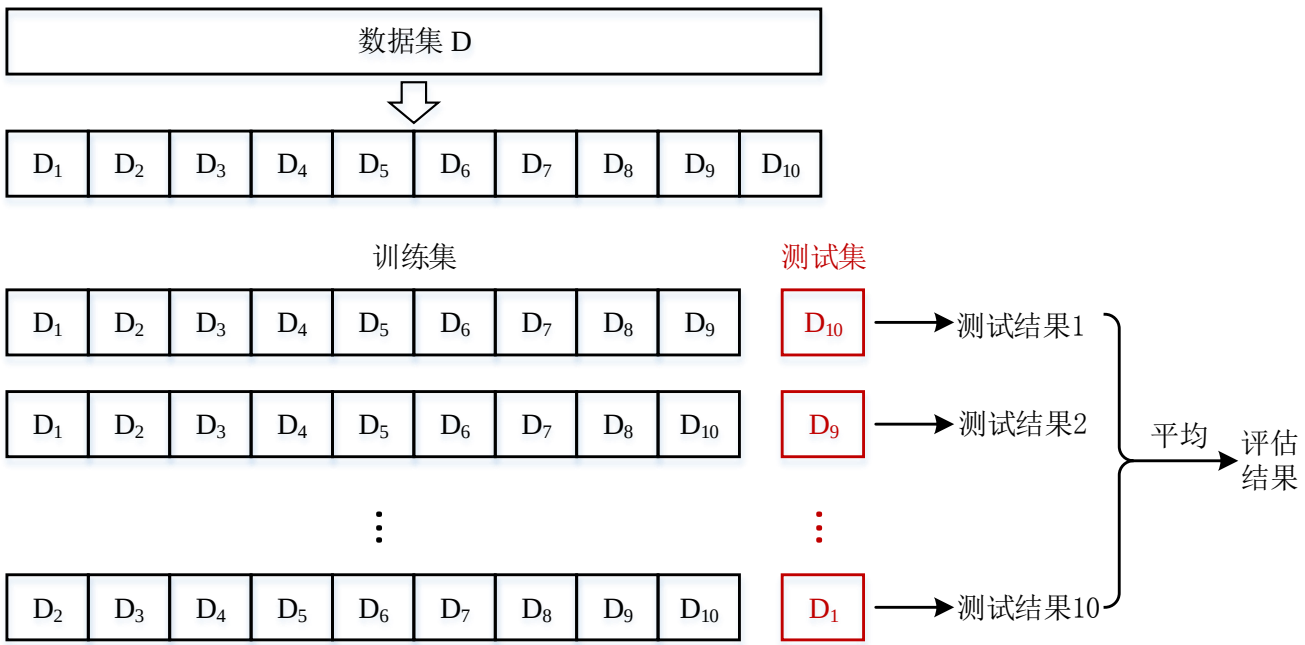
留出法

- **数据集的划分原则**
  - 保持数据分布一致性，例如分类问题中应该不同类别样本的比例是一致的；
  - 测试集不宜太大或太小，一般选择样本总数的20% ~ 30%；

- **评估过程**
  - 按比例将数据集D随机划分为S和T；
  - 训练集S学习模型，T测试模型的性能；
  - 重复划分和测试若干次，计算平均性能；

交叉验证法

- k-折交叉验证法
  - 数据集D随机划分为k个子集；
  - $k-1$个子集用于训练，1个用于测试；



自助法

- **自助法的过程**
  - 从数据集D中有放回地随机抽样m个样本，构成训练集S；
  - 从数据集D中有放回地随机抽样m个样本，构成测试集T；
  - 重复若干次，计算平均的评估结果；
- **自助法的优点**
  - 数据集D中约有30%的样本没有出现在训练集S中：

$$P(\mathbf{x} \notin S) \approx \lim_{m \to \infty} \left(1 - \frac{1}{m}\right)^m = \frac{1}{e} \approx 0.368, \quad \forall \mathbf{x} \in D$$

  - S和T的样本数与D相同，测试可以进行任意多次，评估结果更接近最终模型的性能；
  - 自助法更适用于样本数m较小时，样本数多时留出法和交叉验证更常用；

机器学习的“调参”

- **超参数**
  - 机器学习算法依据训练集学习模型的参数；
  - 学习算法本身也有一些参数需要人工设定，称为超参数；
  - 选择超参数的过程称为“调参”，常常会影响最终的性能；
- **调参过程**
  - 数据集D划分为：训练集+验证集+测试集；
  - 设置不同的超参数，使用训练集学习多个模型；
  - 使用验证集选择“最优的”超参数，“训练集+验证集”学习最终的模型；
  - 使用测试集评估学习的结果；

性能度量

- **模型性能的评价**
  - 性能度量(performance measure)是衡量模型泛化能力的评价标准，使用不同的性能度量往往会导致不同的评判结果；
  - 评价什么样的模型是“好”的，不仅取决于算法和数据，还取决于任务的需求；
- **回归任务的性能度量**
  - 回归任务常用平方误差度量模型$f(\mathbf{x})$在数据集D上的性能：

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\left(f(\mathbf{x}_i)-y_i\right)^2$$

性能度量

- **分类任务的性能度量**
  - 分类错误率：

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(f(\mathbf{x}_i)\neq y_i\right)$$

  - 分类精度：

$$acc(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(f(\mathbf{x}_i)=y_i\right)=1-E(f;D)$$

其中，$\mathbb{I}(\alpha)$为指示函数

$$\mathbb{I}(\alpha)=\begin{cases}1, & \alpha=true\\ 0, & \alpha=false\end{cases}$$

性能度量

- **检索任务的性能度量**
  - 检索任务是一个两分类问题，相关的内容为正例，无关的内容为反例；
  - 分类的结果可以用混淆矩阵表示：
    - 真正例(TP)：真实的正例被分类为正例的数量；
    - 假反例(FN)：真实的正例被分类为反例的数量；
    - 假正例(FP)：真实的反例被分类为正例的数量；
    - 真反例(TN)：真实的反例被分类为反例的数量；

真实情况	预测结果	
	正例	反例
正例	TP(真正例)	FN(假反例)
反例	FP(假正例)	TN(真反例)

性能度量

- **检索任务的性能度量**
  - 查准率：检索出来的结果中，真实正例所占的比例

$$P = \frac{TP}{TP + FP}$$

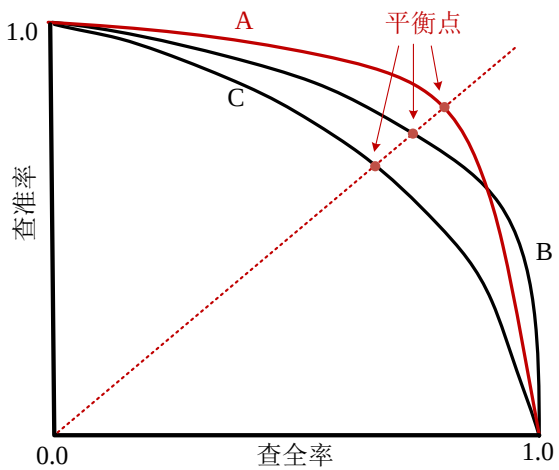
  - 查全率：所有真实正例中被检索出来的比例，也称为召回率

$$R = \frac{TP}{TP + FN}$$

  - 查全率和查准率是相互矛盾的，一般与检出的总数有关：
    - 检索结果数量多，查全率高，查准率低；
    - 检索结果数量少，查全率低，查准率高；

性能度量

- **检索任务的性能度量**
  - P-R曲线：设置不同的检索结果数量，将相应的查全率和查准率的关系画成曲线，曲线下的面积可以度量不同分类器的性能
  - 平衡点：P-R曲线上“查全率=查准率”的点，值越大的分类器性能越好



性能度量

- **检索任务的性能度量**
  - F_1度量：综合考虑查全率和查准率，取两者的几何平均

$$F_1 = \frac{2 \times P \times R}{P + R} = \frac{2 \times TP}{m + TP - TN}$$

  - F_β度量：通过参数β调节对查全率和查准率的关心程度

$$F_\beta = \frac{(1 + \beta^2) \times P \times R}{(\beta^2 \times P) \times R}$$

$\beta > 1$更关心查全率，$\beta < 1$更关心查准率；

偏差与方差

- **机器学习的结果是概率近似正确的**
  - 机器学习的目的是要能够很好地预测“未知”的测试样本，但依赖的只能是“已知”的训练样本；
  - 训练样本集D是对学习对象样本总体的一次随机抽样，依据不同的训练集学习算法会得到不同的学习结果；
  - 很多学习算法本身也是有随机性的；
  - 学习结果可以看作是一个随机事件，从它的统计特性可以进一步了解机器学习泛化误差的组成；

偏差与方差

- **泛化误差的数学期望**
  - 令$f(\mathbf{x};D)$是学习算法依据训练集D学习的模型，对输入$\mathbf{x}$的预测输出；y是$\mathbf{x}$的真实标记，y_D是训练集D中$\mathbf{x}$的标记；
  - 训练集D是随机的，可以证明，学习到的模型在$\mathbf{x}$上(回归)泛化误差对D的数学期望：

$$
\begin{aligned}
E(f;D) &= \mathbb{E}_D\left[(f(\mathbf{x};D)-y_D)^2\right] \\
&= \mathbb{E}_D\left[(f(\mathbf{x};D)-\bar{f}(\mathbf{x}))^2\right] + \mathbb{E}_D\left[(\bar{f}(\mathbf{x})-y)^2\right] + \mathbb{E}_D\left[(y_D-y)^2\right] \\
&= var(\mathbf{x}) + bias^2(\mathbf{x}) + \epsilon^2
\end{aligned}
$$

偏差与方差

- **泛化误差的组成**
  - $\bar{f}(\mathbf{x}) = \mathbb{E}_D[f(\mathbf{x}; D)]$是学习算法依据不同的训练集$D$得到的模型，对$\mathbf{x}$预测输出的数学期望；
  - 噪声项：与学习算法无关，是对于标记y的测量误差

$$\epsilon^2 = \mathbb{E}_D\left[(y_D - y)^2\right]$$

  - 偏差项：学习算法预测输出的期望与真实标记之间的偏差，体现了学习算法的能力

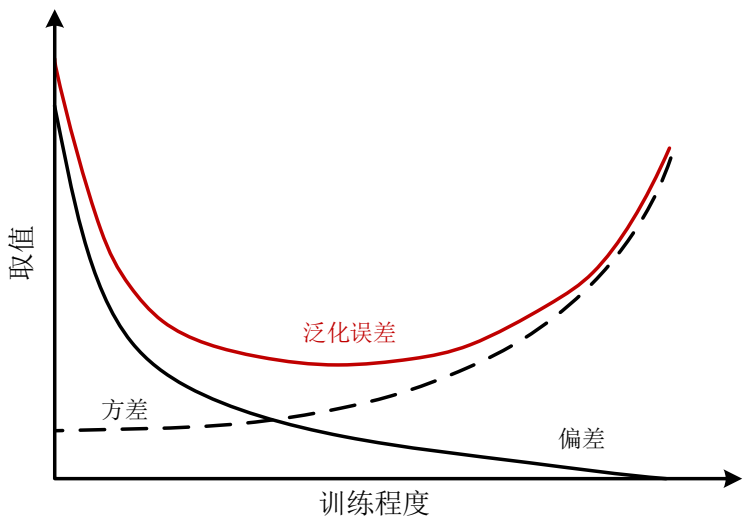
$$bias^2(\mathbf{x}) = \mathbb{E}_D\left[(\bar{f}(\mathbf{x}) - y)^2\right]$$

  - 方差项：学习算法使用不同训练集学习，预测结果之间的方差，体现了学习算法的稳定性

$$var(\mathbf{x}) = \mathbb{E}_D\left[(f(\mathbf{x}; D) - \bar{f}(\mathbf{x}))^2\right]$$

偏差与方差

- **偏差与方差的矛盾**
  - 训练不足时，学习器拟合能力不强，偏差大而方差小；
  - 训练充足后，学习器的拟合能力很强，方差大而偏差小；
  - 学习算法需要综合考虑偏差和方差，才能保证泛化能力；



关于课程

课程教材

- **课程教材**
  - 周志华,机器学习,清华大学出版社,2016

课程QQ群

群名称：机器学习-2020本
群　号：1147919939