机器学习导论

第二讲 基本术语和模型评估

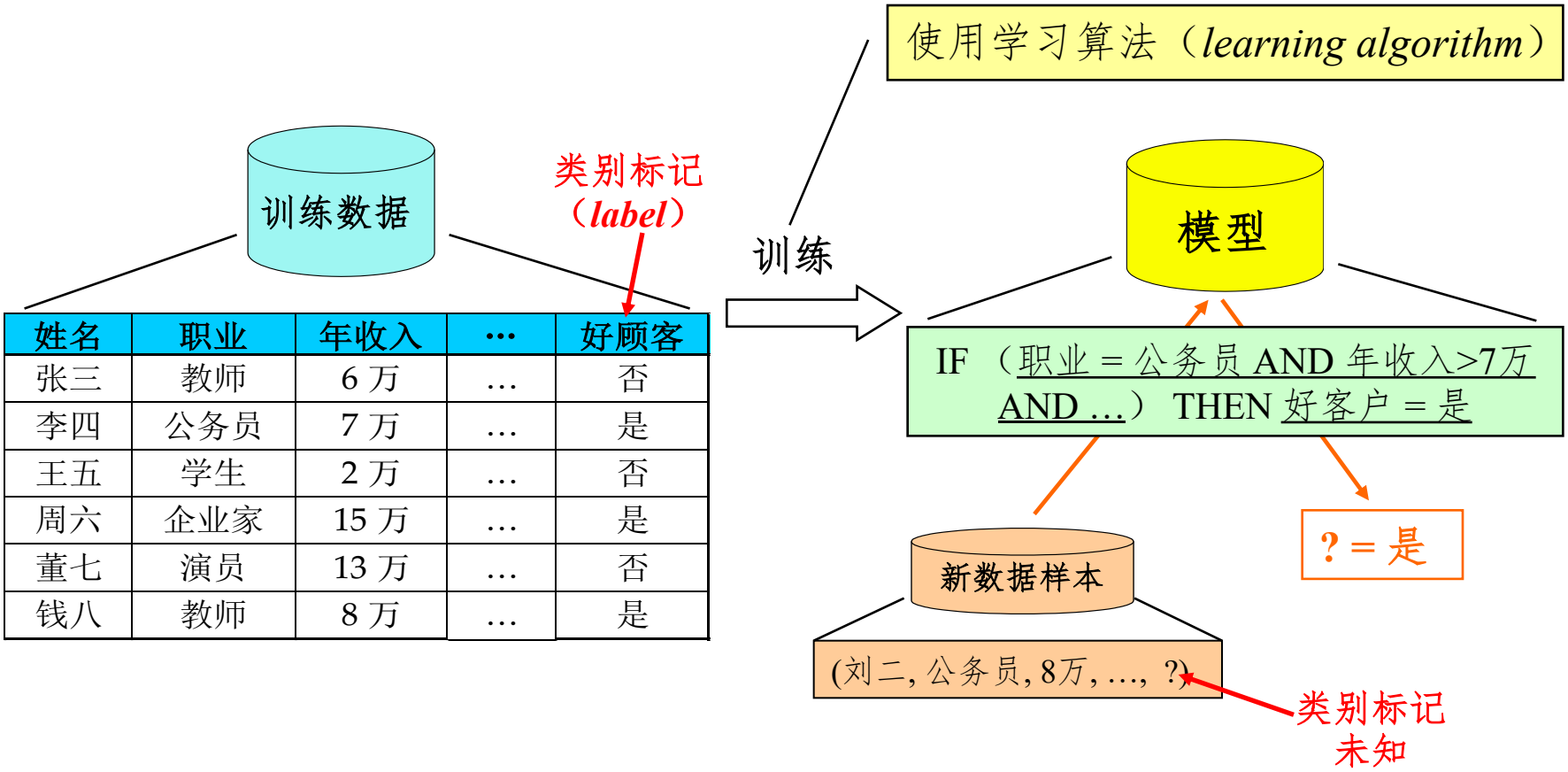
提纲

- 机器学习的基本术语和概念学习
- 机器学习的模型评估和选择

机器学习基本术语和概念学习

机器学习像人类学习怎么学?

典型的机器学习过程



姓名	职业	年收入	…	好顾客
张三	教师	6 万	…	否
李四	公务员	7 万	…	是
王五	学生	2 万	…	否
周六	企业家	15 万	…	是
董七	演员	13 万	…	否
钱八	教师	8 万	…	是

基本术语-数据

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	沉闷	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否
1	青绿	蜷缩	沉闷	?

机器学习是一项面向未来（输出）的技术

基本术语-数据

学习过程就是得到输入到输出的预测模型

- 术语扩展
  - 特征，又称为属性
  - 属性值：特征的离散取值，或者，连续取值；
  - 样本维度（Dimensionality）：特征个数
  - 特征张成的空间：属性空间/特征空间/输入空间
  - 标记张成的空间：标记空间/输出空间
  - 示例（Instance）/样本（Sample）：一个对象的输入（比如，一个西瓜的描述）示例不含标记
  - 样例（Example）：示例+标记
  - 训练集 = 一组训练样例
  - 测试集 = 一组测试样例

基本术语-任务

- 预测任务：**根据标记的取值情况**
  - 分类任务：标记为离散值
    - 二分类：例如（好瓜，坏瓜）（正类，反类）（+1，-1）
    - 多分类：例如（冬瓜，南瓜，西瓜）
  - 回归任务：标记为连续值
    - 例如，瓜的成熟度
  - 聚类任务：标记为空值，对示例进行自动分组
    - 例如，本地瓜，外地瓜

基本术语-任务

- 预测任务：根据标记的完整情况
  - （有）监督学习：所有示例都有标记
    - 分类、回归
  - 无监督学习：所有示例都没有标记
    - 聚类
  - 半监督学习：少量示例有标记，大量示例没标记
  - 噪音标记学习：标记有，但是不完全准确
  - …

基本术语-目标

机器学习技术的根本目标就是

泛化能力!

也可以理解为：应对未来未见测试样本的能力

但是未来是不知道的，一般依靠历史数据来逼近模型的泛化能力

历史和未来来自相同分布
（I.I.D. 假设）

概念学习

最理想的机器学习技术是学习到 **概念（人类学习，可理解的）**

勤奋+天赋 => 成功

学好数理化 => 走遍天下都不怕

三角形内角和180 => 四边形不是三角形

。。。

但是现实中很困难，因此很多时候机器学习采用的是黑盒模型

通过一窥概念学习的过程，可以体会最早机器学习范式的风采和机器学习存在的技术难点

概念学习-假设空间

最理想的机器学习技术是学习到 **概念（人类学习，可理解的）**

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	沉闷	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否

(色泽=?)∧(根蒂=?)∧(敲声=?)↔好瓜

假设空间大小：3*4*4+1=49

概念学习-假设空间

枚举所有可能的假设

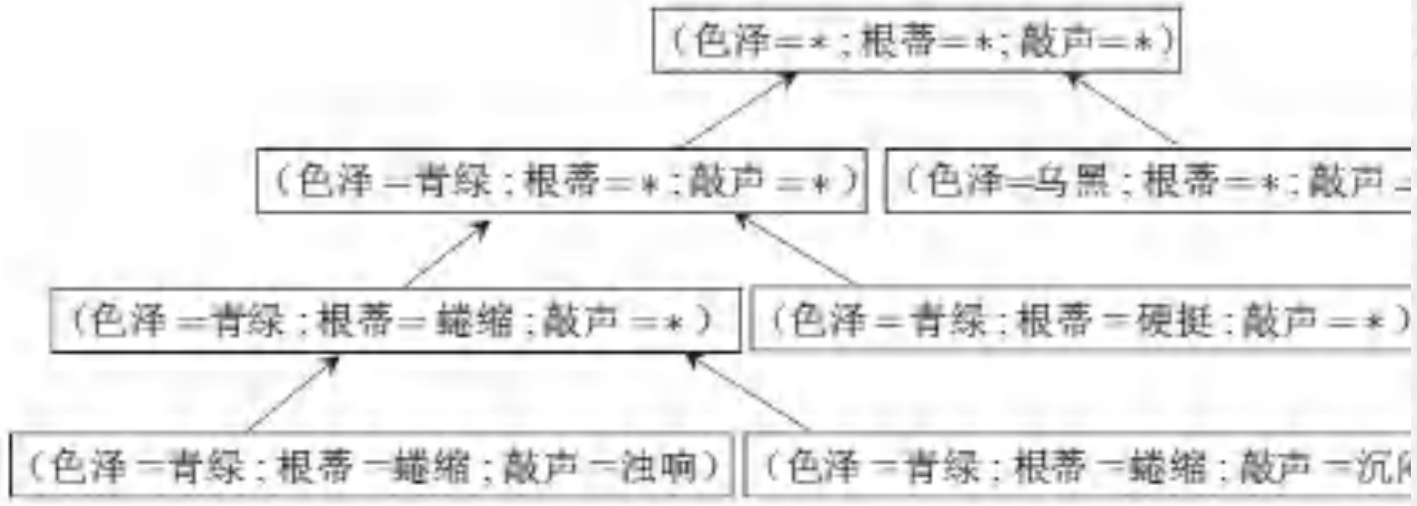
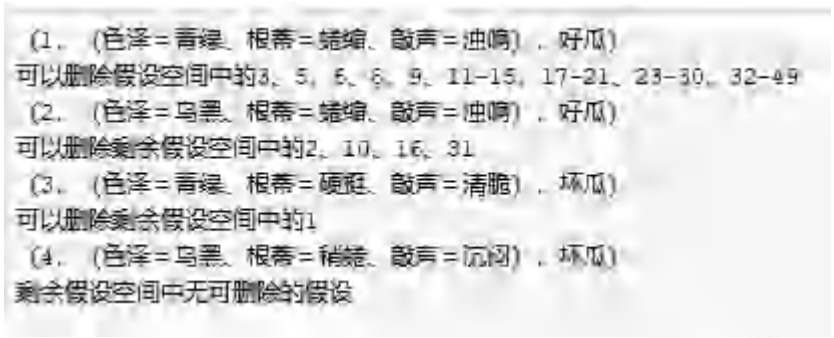
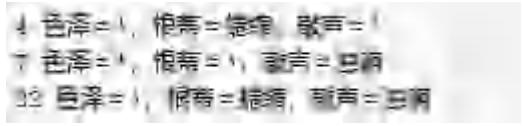


图 1.1 西瓜问题的假设空间

概念学习-版本空间

- 假设空间的子集：跟训练集一致的“假设集合”



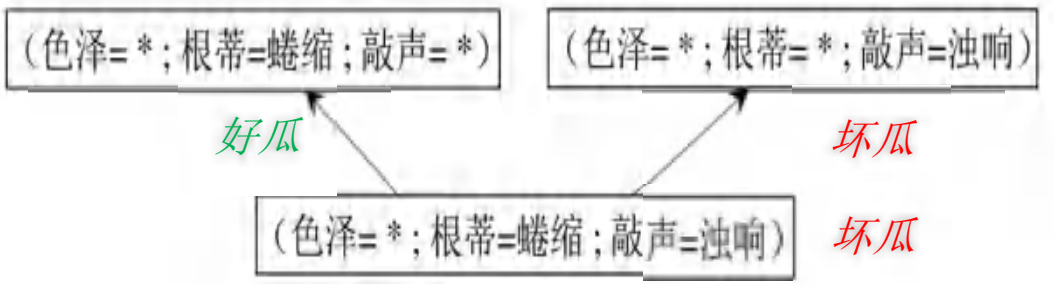


在所有假设中，有三个假设跟训练集数据情况一致

概念学习-归纳偏好

假设空间中有三个与训练集一致的假设，但他们对(色泽=青绿；敲声=沉闷)的瓜

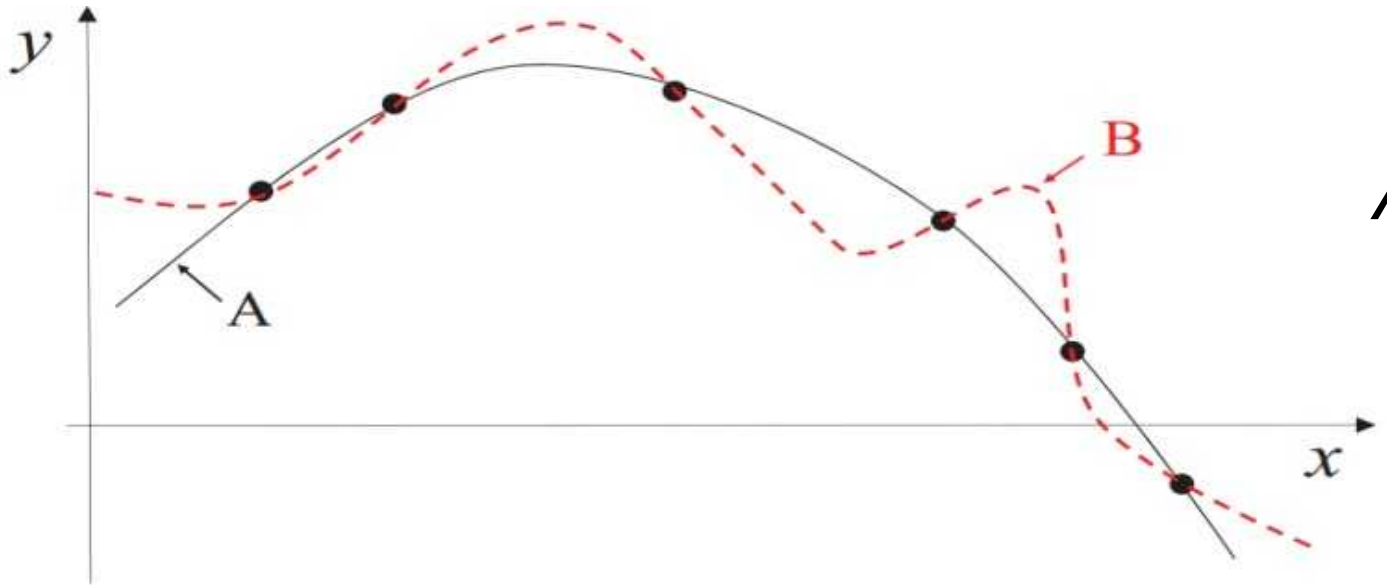
会预测出不同的结果：



选取哪个假设作为学习模型？

概念学习-归纳偏好

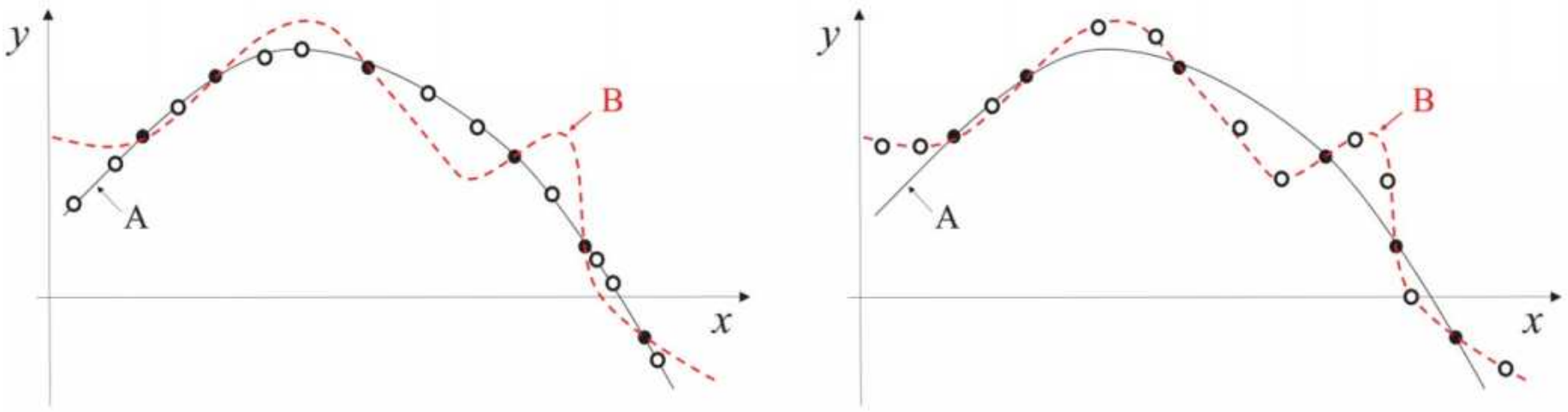
学习过程中对某种类型假设的偏好称作归纳偏好



A or B?

存在多条曲线与有限样本训练集一致

概念学习-No Free Lunch



(a) A 优于 B　　(b) B 优于 A

没有免费的午餐. (黑点: 训练样本; 白点: 测试样本)

因为未来数据是不知道的，总有一种未来数据的分布让你失败

No Free Lunch（没有免费午餐）

一个算法ξ_a如果在某些问题上比另一个算法ξ_b好，必然存在另一些问题，ξ_b比ξ_a好, 也即没有免费的午餐定理。

为简单起见，假设样本空间$\mathcal{X}$和假设空间$\mathcal{H}$都是离散的. 令$P(h|X, \mathfrak{L}_a)$代表算法$\mathfrak{L}_a$基于训练数据X产生假设h的概率，再令f代表我们希望学习的真实目标函数. $\mathfrak{L}_a$的“训练集外误差”，即$\mathfrak{L}_a$在训练集之外的所有样本上的误差为

$$E_{ote}(\mathfrak{L}_a|X, f) = \sum_{h} \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x})\, \mathbb{I}\left(h\left(\boldsymbol{x}\right) \neq f\left(\boldsymbol{x}\right)\right) P\left(h \mid X, \mathfrak{L}_a\right)$$

其中$\mathbb{I}(\cdot)$是指示函数，若$\cdot$为真则取值1，否则取值0.

No Free Lunch（没有免费午餐）

考虑二分类问题，且真实目标函数可以是任何函数 $\mathcal{X} \mapsto \{0,1\}$，函数空间为 $\{0,1\}^{|\mathcal{X}|}$，对所有可能的 f 按均匀分布对误差求和，有

$$
\begin{aligned}
\sum_f E_{ote}(\mathfrak{L}_a|X,f) &= \sum_f \sum_h \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x})\,\mathbb{I}(h(\boldsymbol{x})\neq f(\boldsymbol{x}))\,P(h\mid X,\mathfrak{L}_a) \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h\mid X,\mathfrak{L}_a) \sum_f \mathbb{I}(h(\boldsymbol{x})\neq f(\boldsymbol{x})) \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h\mid X,\mathfrak{L}_a) \frac{1}{2} 2^{|\mathcal{X}|} \\
&= \frac{1}{2} 2^{|\mathcal{X}|} \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h\mid X,\mathfrak{L}_a) \\
&= 2^{|\mathcal{X}|-1} \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \cdot 1 \,.
\end{aligned}
$$

总误差与学习算法无关！

实际问题中，并非所有问题出现的可能性都相同

脱离具体问题，空谈“什么学习算法更好”毫无意义

小结

- 机器学习的基本术语
  - 数据集、特征、标记
  - 分类、回归、聚类
  - 监督、无监督、半监督学习
  - 泛化能力
- 了解概念学习
  - 版本空间
  - 假设空间
  - 归纳偏执
  - 没有免费午餐定理

术语总结

机器学习术语	疾病诊断例子
• 数据集，特征，标记	✓ 某疾病患者人群
• 假设空间	✓ 所有可能的药
• 版本空间	✓ 能治好的药
• 归纳偏执	✓ 偏执：中药西药；副作用大小；费用高低?
• 没有免费午餐	✓ 没有特效药，万能药

“好”算法来自于对数据的好假设，好偏执

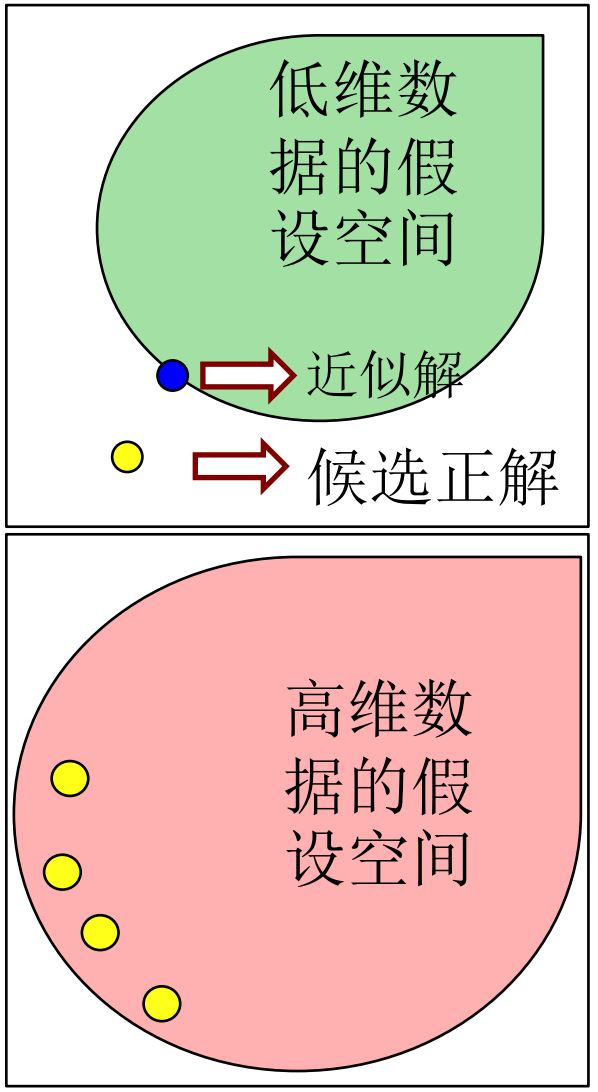
“大胆假设，小心求证”

思考题

- 样本维度越高越好？还是越低越好？

- 假设空间大
  - 版本空间大，偏执越重要

- 假设空间小
  - 版本空间小，偏执越不需要



机器学习技术：“大胆假设”和“小心求证”的折衷

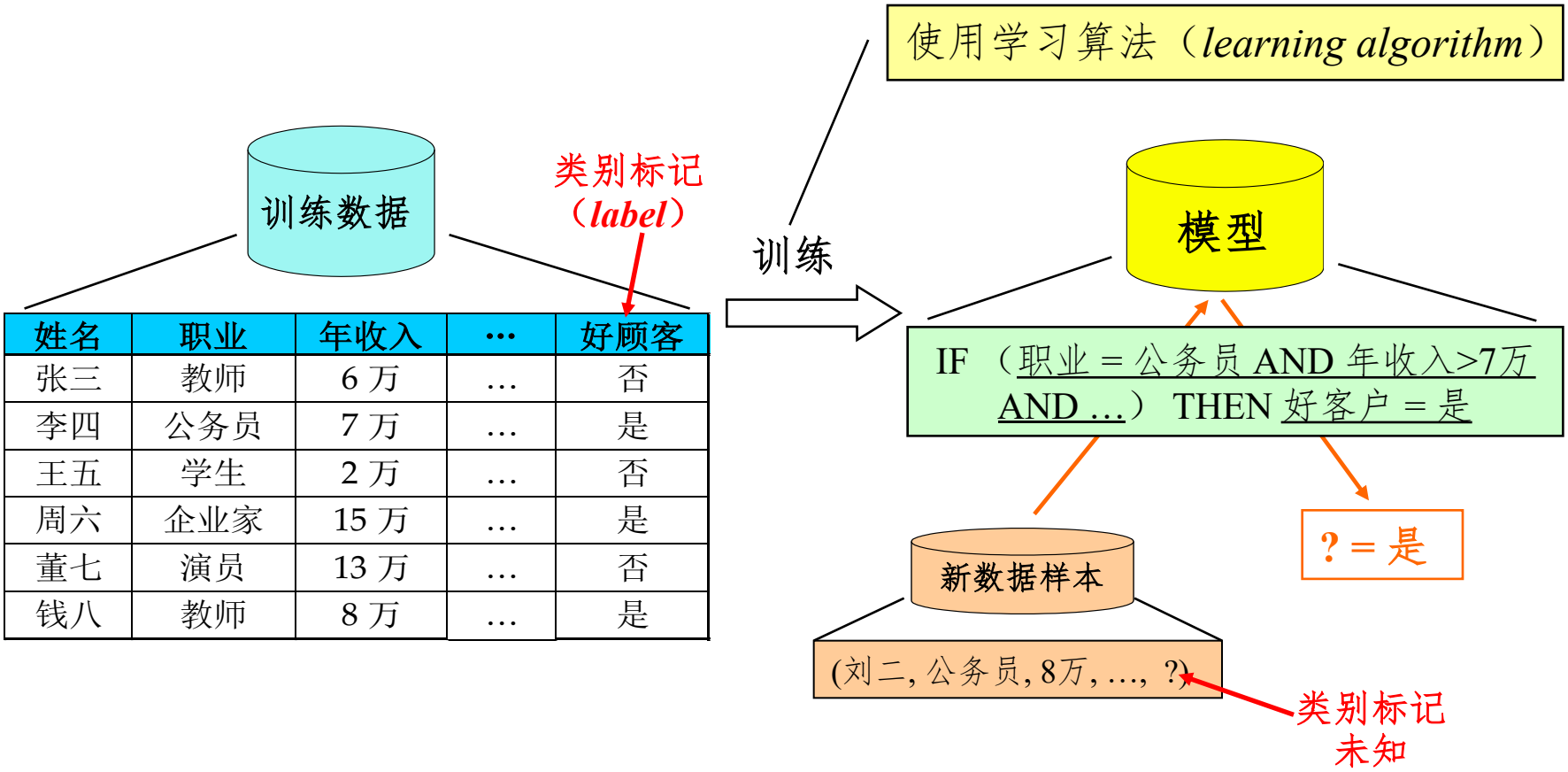
模型评估与选择

怎么选机器学习模型?

提纲

- 经验误差与过拟合
- 评估方法
- 性能度量
- 比较检验
- 偏差与方差

典型的机器学习过程



姓名	职业	年收入	…	好顾客
张三	教师	6 万	…	否
李四	公务员	7 万	…	是
王五	学生	2 万	…	否
周六	企业家	15 万	…	是
董七	演员	13 万	…	否
钱八	教师	8 万	…	是

原理导图

经验性能E（历史表现） ≈ 泛化性能E*（未来表现得不到）

iid假设

模型选择

经验性能E的定义 → ✧ 经验误差与过拟合

经验性能E的估计 → ✧ 评估方法

经验性能E的指标多样 → ✧ 性能度量

如何评价说模型A比模型B好? → ✧ 比较检验

模型评价的理论基础 → ✧ 偏差与方差

让历史尽量少偏离未来

经验误差与过拟合

- 模型选择的目标是泛化误差：模型在未来未见样本的错误率
  - 不可行!

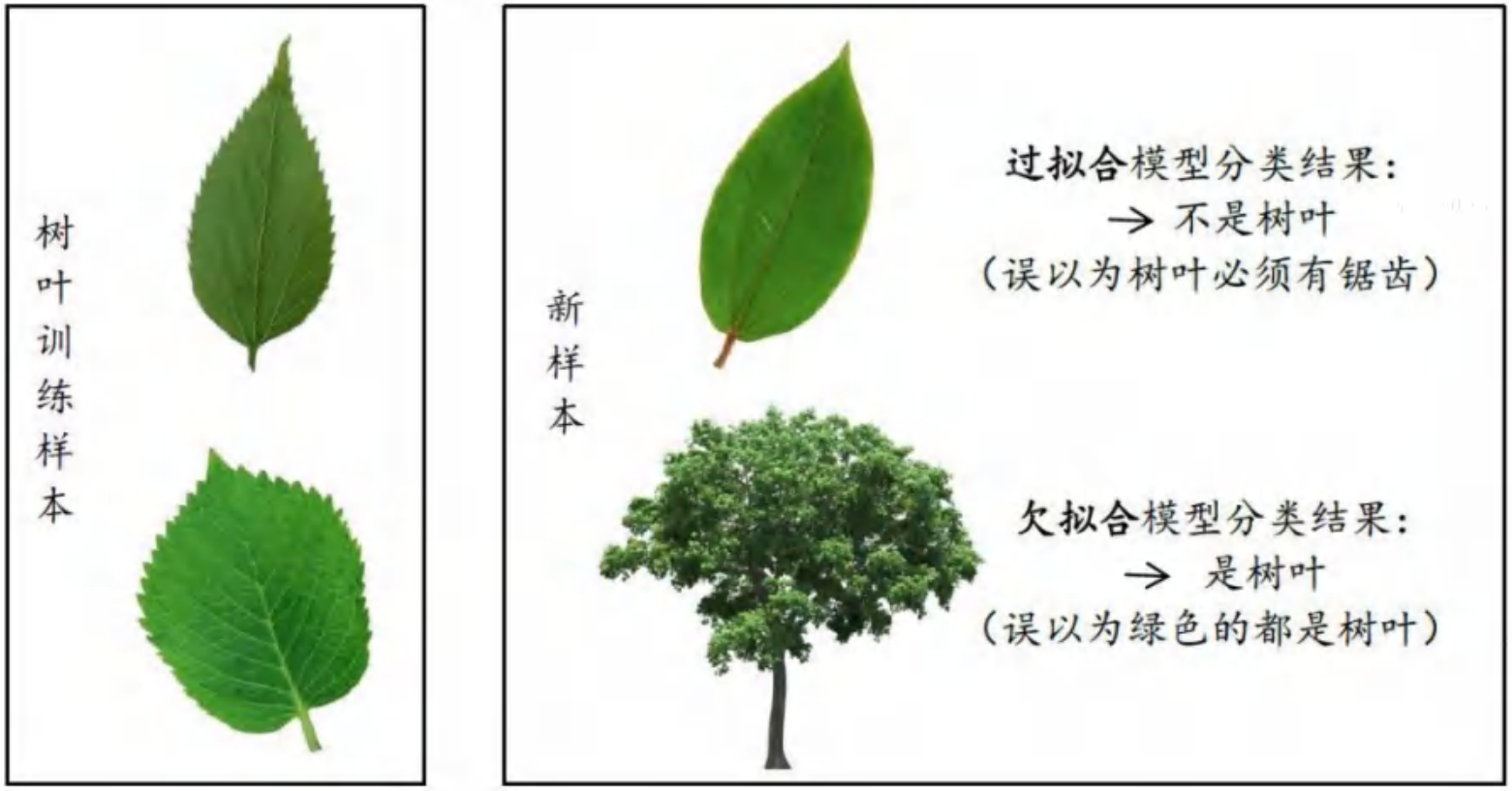
- 错误率error rate&误差error:
  - 错误率：错分样本的占比 $E = a/m$
  - 误差：样本真实输出与预测输出之间的差异
    - 训练(经验)误差：训练集的误差
    - 测试误差：测试集的误差

经验误差与过拟合

- 过拟合：学习器把训练样本学习的“太好”，将训练样本本身的特点；当做所有样本的一般性质，导致泛化性能下降
  - 解决办法
    - 优化目标加正则项
    - early stopping

- 欠拟合：对训练样本的一般性质尚未学好
  - 解决方法
    - 决策树:拓展分支
    - 神经网络:增加训练轮数

经验误差与过拟合



过拟合、欠拟合的直观类比

过拟合和欠拟合无法完全避免，只能不断折衷

原理导图

经验性能E（历史表现） ≈ 泛化性能E*（未来表现得不到） ⇐ 模型选择

iid假设

经验性能E的定义 → ✧ 经验误差与过拟合

经验性能E的估计 → ✧ 评估方法

经验性能E的指标多样 → ✧ 性能度量

如何评价说模型A比模型B好? → ✧ 比较检验

模型评价困难的理论基础 → ✧ 偏差与方差

让历史尽量少偏离未来

评估方法

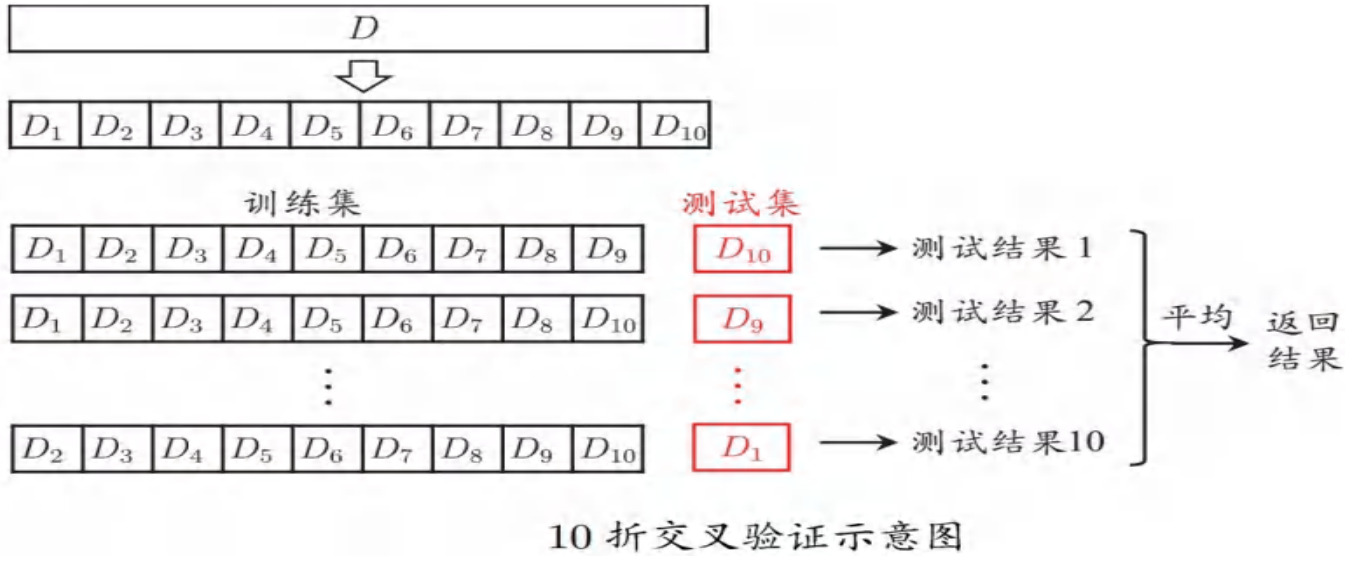
可是，我们只有一个包含 m 个样例的数据集 $D=\{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\}$，既要训练，又要测试，怎样才能做到呢？答案是：通过对 D 进行适当的处理，从中产生出训练集 S 和测试集 T．下面介绍几种常见的做法．

- 留出法（hold-out）：
  - 直接将数据集划分为两个互斥集合——训练和测试集
  - 训练/测试集划分要尽可能保持数据分布的一致性
  - 分层采样（stratified sampling）：保持类别比例一致
  - 一般若干次随机划分、重复实验取平均值
  - 训练/测试样本比例通常为2:1~4:1，效果还不错

评估方法

- 交叉验证法（cross-validation）：

  将数据集分层采样划分为k个大小相似的互斥子集，每次用k-1个子集的并集作为训练集，余下的子集作为测试集，最终返回k个测试结果的均值，k最常用的取值是10.



10 折交叉验证示意图

评估方法

与留出法类似，将数据集D划分为k个子集同样存在多种划分方式，为了减小因样本划分不同而引入的差别，k折交叉验证通常随机使用不同的划分重复p次，最终的评估结果是这p次k折交叉验证结果的均值。例如，常见的“10次10折交叉验证”

假设数据集D包含m个样本，若令k=m，则得到留一法(LOO)：

- 不受随机样本划分方式的影响
- 结果往往比较准确
- 当数据集比较大时，计算开销难以忍受，实际中不采用

评估方法

- 自助法：

“自助法”(bootstrapping)是一个比较好的解决方案，它直接以自助采样法(bootstrap sampling)为基础 [Efron and Tibshirani, 1993]. 给定包含 m 个样本的数据集 D，我们对它进行采样产生数据集 D'：每次随机从 D 中挑选一个样本，将其拷贝放入 D'，然后再将该样本放回初始数据集 D 中，使得该样本在下次采样时仍有可能被采到；这个过程重复执行 m 次后，我们就得到了包含 m 个样本的数据集 D'，这就是自助采样的结果. 显然，D 中有一部分样本会在 D' 中多次出现，而另一部分样本不出现. 可以做一个简单的估计，样本在 m 次采样中始终不被采到的概率是 $\left(1-\frac{1}{m}\right)^m$，取极限得到

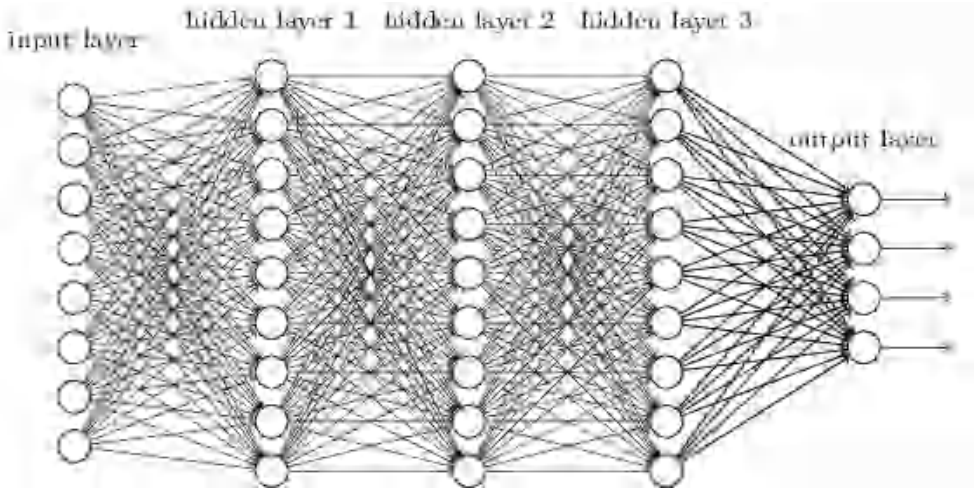
$$\lim_{m\mapsto\infty}\left(1-\frac{1}{m}\right)^m \mapsto \frac{1}{e}\approx 0.368\,, \tag{2.1}$$

自助法: 数据集小，难以划分训练/测试集很有用；产生多个训练集，对集成学习有用；数据量足够，一般采用留出法和交叉验证

评估方法

- 调参数：

大多数学习算法都有些参数(parameter)需要设定，参数配置不同，学得模型的性能往往有显著差别. 因此，在进行模型评估与选择时，除了要对适用学习算法进行选择，还需对算法参数进行设定，这就是通常所说的“参数调节”或简称“调参”(parameter tuning).



深度网络
上亿参数

解决办法：验证集 （在验证集上确定参数，付诸于最终模型和测试）

原理导图

经验性能E（历史表现） ≈ 泛化性能E*（未来表现得不到）

iid假设

模型选择

经验性能E的定义 → ✧ 经验误差与过拟合

经验性能E的估计 → ✧ 评估方法

经验性能E的指标多样 → ✧ 性能度量

如何评价说模型A比模型B好? → ✧ 比较检验

模型评价的理论基础 → ✧ 偏差与方差

让历史尽量少偏离未来

性能度量

性能度量是衡量模型泛化能力的评价标准，反映了任务需求；

使用不同的性能度量往往会导致不同的评判结果

在预测任务中，给定样例集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$，其中 y_i 是示例 $\boldsymbol{x}_i$ 的真实标记. 要评估学习器 f 的性能，就要把学习器预测结果 $f(\boldsymbol{x})$ 与真实标记 y 进行比较.

回归任务最常用的性能度量是“均方误差”：

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

性能度量

对于分类任务，**错误率和精度**是最常用的两种性能度量：

- 错误率：分错样本占样本总数的比例
- 精度：分对样本占样本总数的比率

分类错误率

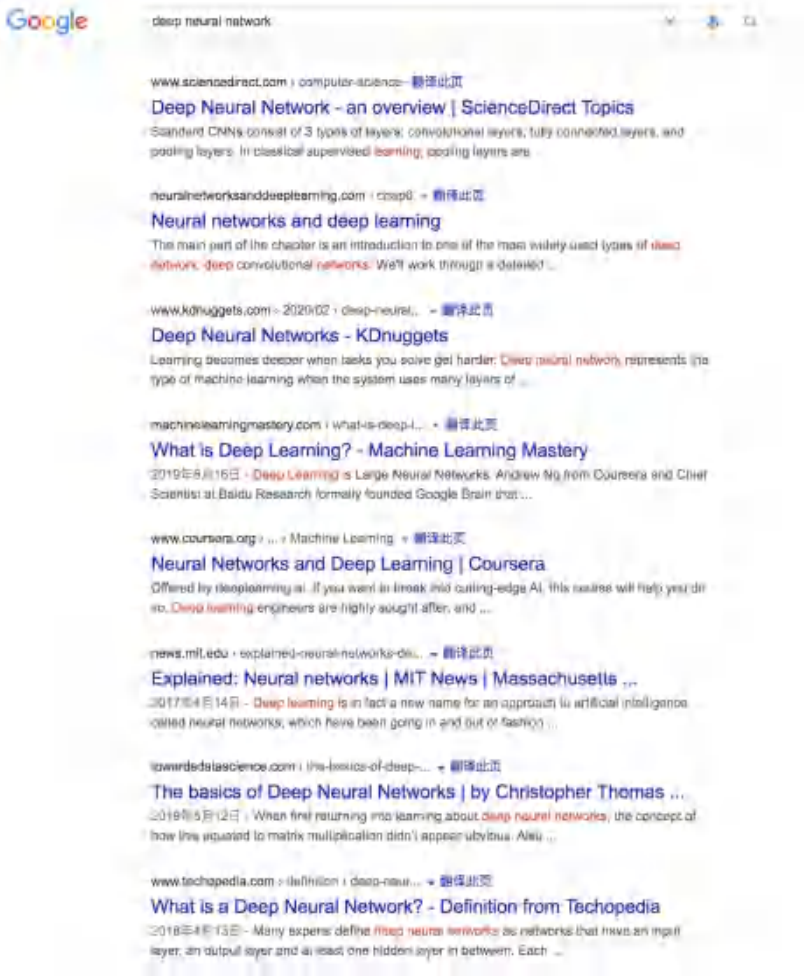
$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(\boldsymbol{x}_i)\neq y_i)$$

精度

$$\begin{aligned}\mathrm{acc}(f;D) &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(\boldsymbol{x}_i)=y_i)\\ &= 1-E(f;D)\,.\end{aligned}$$

性能度量

信息检索、Web搜索等场景中经常需要衡量正例被预测出来的比率或者预测出来的正例中正确的比率，此时查准率和查全率比错误率和精度更适合。



性能度量

信息检索、Web搜索等场景中经常需要衡量正例被预测出来的比率或者预测出来的正例中正确的比率，此时查准率和查全率比错误率和精度更适合。

统计真实标记和预测结果的组合可以得到“混淆矩阵”

分类结果混淆矩阵

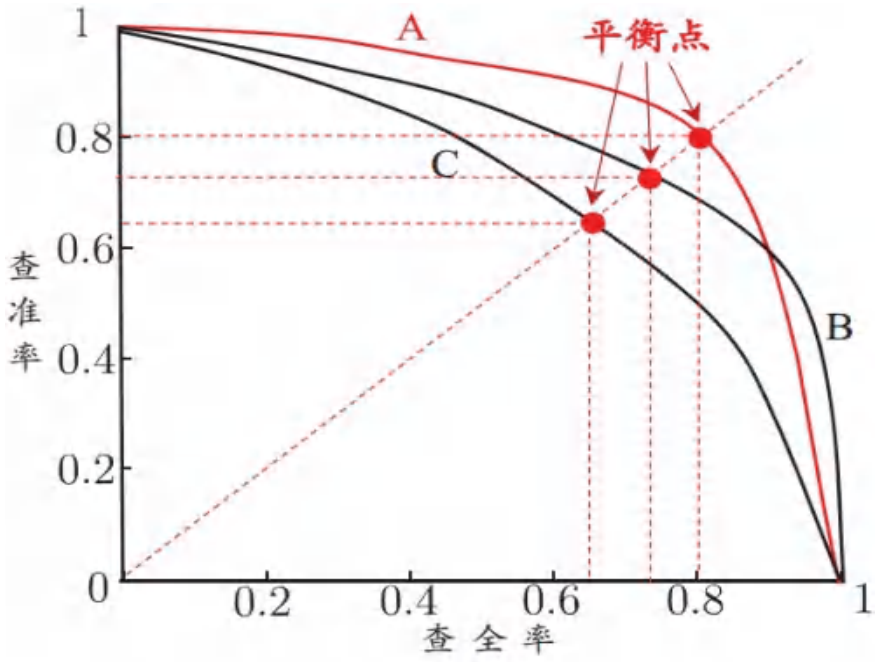
真实情况	预测结果	
	正例	反例
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

查准率 $P = \frac{TP}{TP + FP}$

查全率 $R = \frac{TP}{TP + FN}$

性能度量

根据学习器的预测结果按正例可能性大小对样例进行排序，并逐个把样本作为正例进行预测，则可以得到查准率-查全率曲线，简称“P-R曲线”



P-R曲线与平衡点示意图

平衡点是曲线上“查准率=查全率”时的取值，可用来用于度量P-R曲线有交叉的分类器性能高低

性能度量

比P-R曲线平衡点更用常用的是F1度量：

$$F1 = \frac{2 \times P \times R}{P + R} = \frac{2 \times TP}{样例总数 + TP - TN}$$

比F1更一般的形式 ，

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

$\beta = 1$： 标准F1

$\beta > 1$：偏重查全率(逃犯信息检索)

$\beta < 1$：偏重查准率(商品推荐系统)

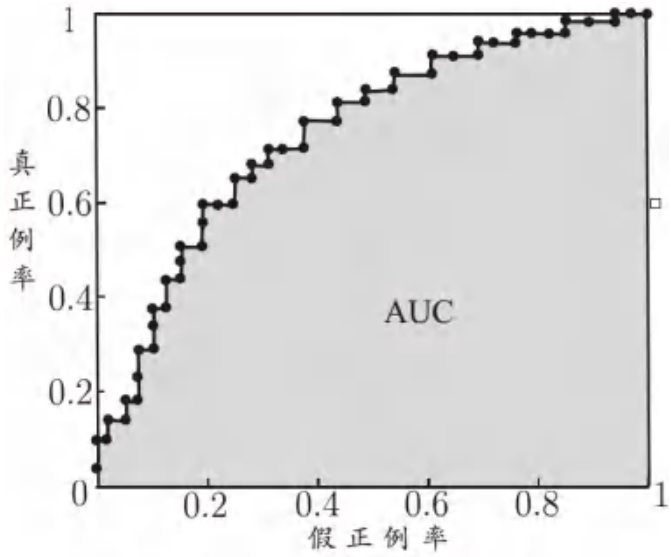
性能度量

类似P-R曲线，根据学习器的预测结果对样例排序，并逐个作为正例进行预测，以“假正例率”为横轴，“真正例率”为纵轴可得到ROC曲线，全称“受试者工作特征”.

ROC图的绘制：给定[illegible]个正例和 m^- 个负例，根据学习器预测结果对样例进行排序，将分类阈值设为每个样例的预测值，当前标记点坐标为 (x,y)，当前若为真正例，则对应标记点的坐标为 [illegible]；当前若为假正例，则对应标记点的坐标为 $(x+\frac{1}{m^-},y)$ ，然后用线段连接相邻点.

性能度量

若某个学习器的ROC曲线被另一个学习器的曲线“包住”，则后者性能优于前者；否则如果曲线交叉，可以根据ROC曲线下面积大小进行比较，也即AUC值.



基于有限样例绘制的 ROC 曲线与 AUC

假设ROC曲线由$\{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\}$的点按序连接而形成，则：AUC可估算为：

$$\mathrm{AUC}=\frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1}-x_i)\cdot(y_i+y_{i+1})$$

AUC衡量了样本预测的排序质量。

$$\mathrm{AUC}=1-\ell_{rank}.$$

代价敏感错误率

现实任务中不同类型的错误所造成的后果可能不同。为了权衡不同类型错误所造成的不同损失，为错误赋予“非均等代价”。

健康

小

→

错分代价不同

←

大

癌症患者

代价敏感错误率

现实任务中不同类型的错误所造成的后果可能不同。为了权衡不同类型错误所造成的不同损失，为错误赋予“非均等代价”。

以二分类为例，可根据领域知识设定“代价矩阵”，如下表所示，其中 $cost_{ij}$ 表示将第i类样本预测为第j类样本的代价。损失程度越大，$cost_{01}$ 与 $cost_{10}$ 值的差别越大。

在非均等代价下，不再最小化错误次数，而是最小化“总体代价”，则“代价敏感”错误率相应的为：

$$E(f;D;cost)=\frac{1}{m}\left(\sum_{\boldsymbol{x}_i\in D^+}\mathbb{I}(f(\boldsymbol{x}_i)\neq y_i)\times cost_{01}+\sum_{\boldsymbol{x}_i\in D^-}\mathbb{I}(f(\boldsymbol{x}_i)\neq y_i)\times cost_{10}\right)$$

原理导图

经验性能E（历史表现） ≈ 泛化性能E*（未来表现得不到）

iid假设

模型选择

经验性能E的定义 → ✧ 经验误差与过拟合

经验性能E的估计 → ✧ 评估方法

经验性能E的指标多样 → ✧ 性能度量

如何评价说模型A比模型B好? → ✧ 比较检验

模型评价困难的理论基础 → ✧ 偏差与方差

让历史尽量少偏离未来

性能评估

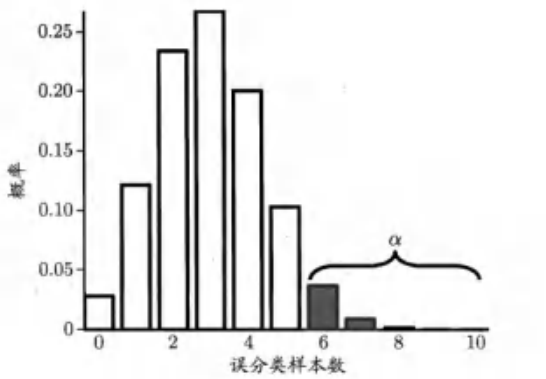
- 关于性能比较：
  - 测试性能并不等于泛化性能
  - 测试性能随着测试集的变化而变化
  - 很多机器学习算法本身有一定的随机性

直接选取相应评估方法在相应度量下比大小的方法不可取！

假设检验为学习器性能比较提供了重要依据，基于其结果我们可以推断出若在测试集上观察到学习器A比B好，则A的泛化性能是否在统计意义上优于B，以及这个结论的把握有多大。

二项检验

记泛化错误率为ϵ，测试错误率为$\hat{\epsilon}$，假定测试样本从样本总体分布中独立采样而来，我们可以使用“二项检验”对 $\epsilon \leq \epsilon_0$ 进行假设检验。



假设 $\epsilon \leq \epsilon_0$，若测试错误率小于

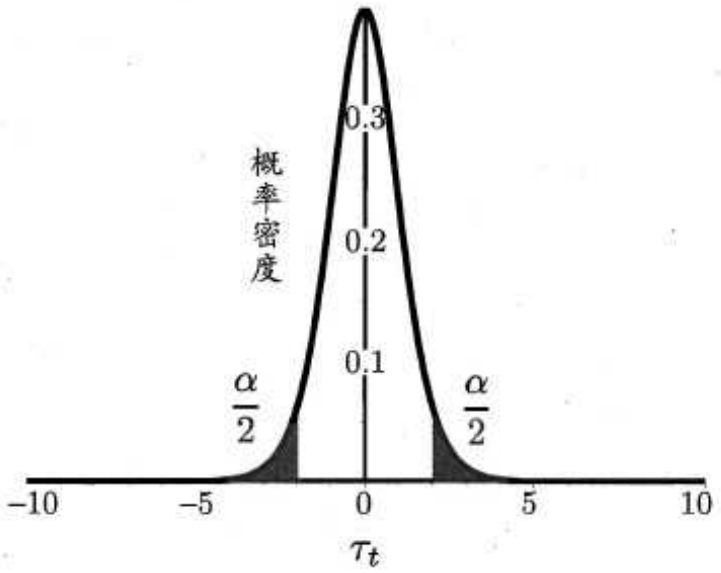
$$\bar{\epsilon} = \max \epsilon \quad \text{s.t.} \quad \sum_{i=\epsilon_0 \times m+1}^{m} \binom{m}{i} \epsilon^i (1-\epsilon)^{m-i} < \alpha$$

则在α 的显著度下，假设不能被拒绝，也即能 [illegible] 的置信度认为，模型的泛化错误率不大于 [illegible].

T-检验

对应的，面对多次重复留出法或者交叉验证法进行多次训练/测试时可使用“t检验”

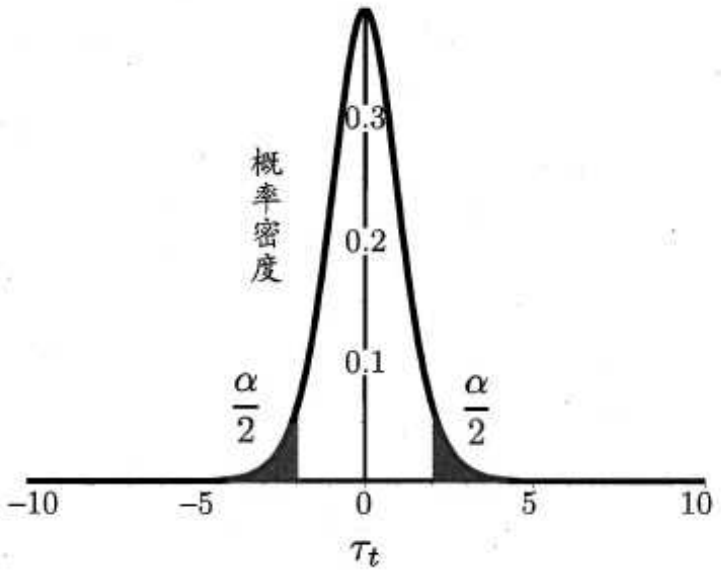
假定得到了k个测试错误率，$\hat{\epsilon}_1, \hat{\epsilon}_2, \ldots, \hat{\epsilon}_k$，假设$\epsilon = \epsilon_0$
对于显著度α，若[illegible]位于临界范围$|\mu - \epsilon_0|$内，则假设不能被拒绝，即可认为泛化错误率$\epsilon = \epsilon_0$，其置信度为[illegible].



交叉验证 T-检验

现实任务中，更多时候需要对不同学习器的性能进行比较

对两个学习器A和B,若k折交叉验证得到的测试错误率分别为$\epsilon_1^A, \ldots, \epsilon_k^A$ 和 $\epsilon_1^B, \ldots, \epsilon_k^B$，可用k折交叉验证“成对t检验”进行比较检验。若两个学习器的性能相同，则他们使用相同的训练/测试集得到的测试错误率应相同，即 $\epsilon_i^A = \epsilon_i^B$.



原理导图

经验性能E（历史表现） ≈ 泛化性能E*（未来表现得不到）

iid假设

模型选择

经验性能E的定义 → ✧ 经验误差与过拟合

经验性能E的估计 → ✧ 评估方法

经验性能E的指标多样 → ✧ 性能度量

如何评价说模型A比模型B好? → ✧ 比较检验

模型评价的理论基础 → ✧ 偏差与方差

让历史尽量少偏离未来

偏差与方差

通过多次实验可以估计学习算法的泛化性能。了解“为什么”泛化性能是这样的呢？ “偏差-方差分解”可以用来帮助解释泛化性能。

对测试样本x，令 y_D 为 x 在数据集中的标记，y 为 x 的真实标记，$f(x;D)$ 为训练集 D 上学得模型 f 在 x 上的预测输出。以回归任务为例：学习算法的期望预期为：

$$\bar{f}(\boldsymbol{x}) = \mathbb{E}_D[f(\boldsymbol{x}; D)]$$

使用样本数目相同的不同训练集产生的方差为

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

噪声为

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

偏差与方差

期望输出与真实标记的差别称为偏差，即 $bias^2(\boldsymbol{x}) = \left(\bar{f}(\boldsymbol{x}) - y\right)^2$

为便与讨论，假定噪声期望为0，也即 $\mathbb{E}_D[y_D - y] = 0$，对泛化误差分解

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

偏差与方差

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

又由假设中噪声期望为0，可得

$$
E(f; D) = \mathbb{E}_D\left[\left(f\left(\boldsymbol{x}; D\right) - \bar{f}\left(\boldsymbol{x}\right)\right)^2\right] + \left(\bar{f}\left(\boldsymbol{x}\right) - y\right)^2 + \mathbb{E}_D\left[\left(y_D - y\right)^2\right]
$$

于是：

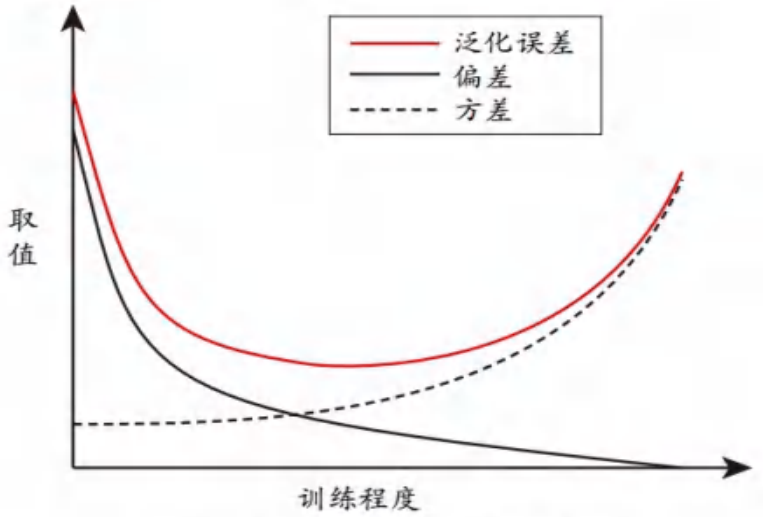
$$
E(f; D) = bias^2\left(\boldsymbol{x}\right) + var\left(\boldsymbol{x}\right) + \varepsilon^2
$$

也即泛化误差可分解为方差、偏差与噪声之和。

偏差与方差

一般来说，偏差与方差是有冲突的，称为偏差-方差窘境。如右图所示，假如我们能控制算法的训练程度：

- 在训练不足时，学习器拟合能力不强，训练数据的扰动不足以使学习器的拟合能力产生显著变化，此时偏差主导泛化错误率；
- 随着训练程度加深，学习器拟合能力逐渐增强，方差逐渐主导泛化错误率；
- 训练充足后，学习器的拟合能力非常强，训练数据的轻微扰动都会导致学习器的显著变化，若训练数据自身非全局特性被学到则会发生过拟合。



泛化误差与偏差、方差的关系示意图

小结

- 经验误差与过拟合
  - 过拟合，欠拟合
- 评估方法
  - 留出法，交叉验证，自助法，验证集调参
- 性能度量
  - 均方误差，精度，查准率/查全率/F1，AUC，Cost敏感
- 比较检验
  - T-检验
- 偏差与方差
  - 了解基本原理