# 人工智能与机器学习

## 第十三讲 算法的评估与比较

倪东 叶琦
工业控制研究所

杭州 • 浙江大学 • 2022

# 概述

- 对学习算法的精度进行评估是机器学习中的基本问题
- 本章用统计方法估计算法精度，主要解决以下三个问题：
  - 已知一个假设在有限数据样本上观察到的精度，怎样估计它在其它实例上的精度？

    即如何评估一个学习算法在给定问题上的期望误差率？
  - 如果一个算法在某些数据样本上好于另一个，那么一般情况下该算法是否更准确？

    即给定两个学习算法，如何就给定应用来判断一个算法的误差率比另一个低。
  - 当数据有限时，怎样高效地利用这些数据，通过它们既能学习到假设，还能估计其精度？
- 统计的方法，结合有关数据基准分布的假定，可以用有限数据样本上的观察精度来逼近整个数据分布上的真实精度。

# 动机

- 对学习到的假设进行尽可能准确地性能评估十分重要
  - 为了知道是否可以使用该假设
  - 是许多学习方法的重要组成部分
- 当给定的数据集有限时，要学习一个概念并估计其将来的精度，存在两个很关键的困难：
  - 估计的偏差
    - 使用与训练样例和假设无关的测试样例
  - 估计的方差
    - 即使假设精度在独立的无偏测试样例上测量，得到的精度仍可能与真实精度不同。
    - 测试样例越少，产生的方差越大
- 重点讨论对学到的假设的评估、对两个假设精度的比较、两个学习算法精度的比较

# 学习问题的框架

- 有一所有可能实例的空间X，其中定义了多个目标函数，假定X中不同实例具有不同的出现频率。一种合适的建模方式是，假定存在一未知的概率分布D，它定义了X中每一实例出现的概率。
- 学习任务是在假设空间上学习一个目标概念（目标函数）
- 目标函数的训练样例中的每一个实例按照分布D独立地抽取，然后连同正确的目标值提供给学习器。

# 评估假设的问题

- 给定假设h和包含若干按D分布抽取的样例的数据集，如何针对将来按同样分布抽取的实例，得到对h的精度最好估计
- 这一精度估计的可能的误差是多少

# 样本错误率和真实错误率

- 定义：假设h关于目标函数f和数据样本S的样本错误率（标记为$error_S(h)$）

$$error_s(h) = \frac{1}{n}\sum_{x \in S} \delta(f(x), h(x)) \qquad \delta(f(x), h(x)) = \begin{cases} 1 & f(x) \neq h(x) \\ 0 & otherwise \end{cases}$$

$$n = |S|$$

- 定义：假设h关于目标函数f和分布D的真实错误率（标记为$error_D(h)$）

$$error_D(h) = \Pr_{x \in D}[f(x) \neq h(x)]$$

# 样本错误率和真实错误率（2）

- 想知道的是假设的真实误差，因为这是在分类未来样例时可以预料到的误差。
- 能测量的只是样本错误率，因为样本数据是我们知道的。
- 要考虑的问题是：样本错误率在何种程度上提供了对真实错误率的估计？

# 离散值假设的置信区间

- 先考虑离散值假设的情况，比如：
  - 样本S包含n个样例，它们的抽取按照概率分布D，抽取过程是相互独立的，并且不依赖于假设h
  - n>=30
  - 假设h在这n个样例上犯了r个错误
- 根据上面的条件，统计理论可以给出以下断言：
  - 没有其它信息的话，真实错误率$error_D(h)$最可能的值是样本错误率$error_S(h)=r/n$
  - 有大约95%的可能性，真实错误率处于下面的区间内：

$$error_S(h) \pm 1.96\sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$$

# 举例说明

- 数据样本S包含n=40个样例，并且假设h在这些数据上产生了r=12个错误，这样样本错误率为$error_S(h)=12/40=0.3$
- 如果没有更多的信息，对真实错误率$error_D(h)$的最好的估计即为0.3
- 如果另外收集40个随机抽取的样例S’，样本错误率$error_{S'}(h)$将与原来的$error_S(h)$存在一些差别
- 如果不断重复这一实验，每次抽取一个包含40样例的样本，将会发现约95%的实验中计算所得的区间包含真实错误率
- 将上面的区间称为$error_D(h)$的95%置信区间估计

# 置信区间表达式的推广

- 常数1.96是由95%这一置信度确定的
- 定义$z_N$为计算N%置信区间的常数（取值见下），计算$error_D(h)$的N%置信区间的一般表达式（公式5.1）为：

$$error_S(h) \pm z_N \sqrt{\frac{error_S(h)(1-error_S(h))}{n}} \qquad (5.1)$$

| confidence level | 50% | 68% | 80% | 90% | 95% | 98% | 99% |
|---|---|---|---|---|---|---|---|
| z-score | 0.67 | 1.00 | 1.28 | 1.64 | 1.96 | 2.33 | 2.58 |

- 可以求得同样情况下的68%置信区间，从直觉上可以看出68%置信区间要小于95%置信区间，因为减小了要求$error_D(h)$落入的概率

# 置信区间表达式的推广（2）

- 公式5.1只能应用于离散值假设，它假定样本S抽取的分布与将来的数据抽取的分布相同，并且假定数据不依赖于所测试的假设；
- 公式5.1只提供了近似的置信区间，这一近似在至少包含30个样例，并且$error_S(h)$不太靠近0或1时很接近真实情况
- 判断这种近似是否接近真实的更精确规则是：

$$n \cdot error_S(h)(1-error_S(h)) \geq 5$$

# 采样统计中的基本定义和概念

- 随机变量
- 某随机变量Y的概率分布
- 随机变量Y的期望值或均值
- 随机变量的方差
- Y的标准差
- 二项分布
- 正态分布
- 中心极限定理
- 估计量
- Y的估计偏差
- N%置信区间

# 错误率估计和二项比例估计

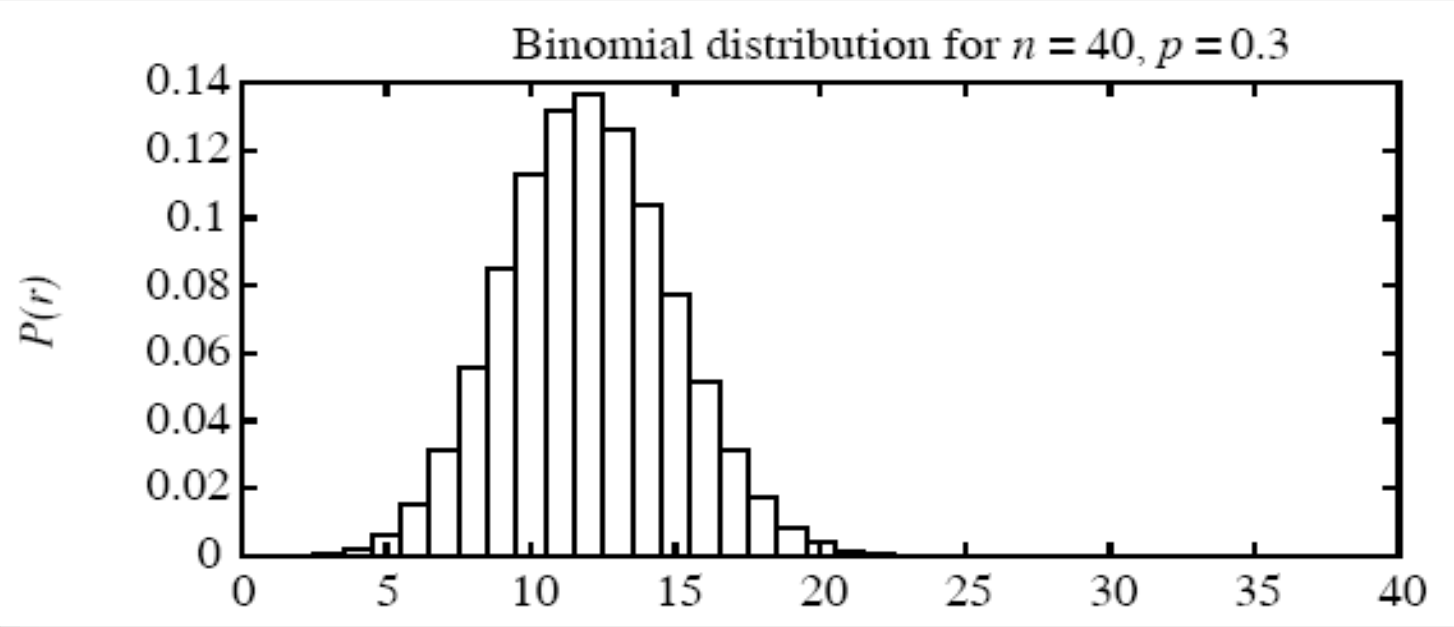
- 样本错误率和真实错误率之间的差异与数据样本大小的依赖关系如何？
- 给定从总体中随机抽取的某些样本的某个属性的观察比例，估计该属性在总体的比例
- 此处，感兴趣的属性是：假设h对实例错误分类

# 错误率估计和二项比例估计（2）

- 测量样本错误率相当于在作一个有随机输出的实验
- 从分布D中随机抽取n个独立的实例，形成样本S，然后测量样本错误率 $error_S(h)$
- 将实验重复多次，每次抽取大小为n的不同的样本$S_i$，得到不同的$error_{S_i}(h)$，取决于$S_i$的组成中的随机差异
- $error_{S_i}(h)$被称为一随机变量，一般情况下，可以将随机变量看成一个有随机输出的实验。随机变量值即为随机实验的观察输出

# 错误率估计和二项比例估计（3）

- 设想要运行k个这样的随机实验，得到k个随机变量值，以图表的形式显示观察到的每个错误率值的频率；
- 当k不断增长，该图表将呈现二项分布。



# 二项分布

- 有一非均质硬币，要估计在抛硬币时出现正面的概率p；
- 投掷硬币n次并计算出现正面的次数r，那么p的一个合理估计是r/n；
- 如果重新进行一次实验，生成一个新的n次抛硬币的集合，出现正面的次数r可能与前不同，得到对p的另一个估计；
- 二项分布描述的是对任一可能的r值，这个正面概率为p的硬币抛掷n次恰好出现r次正面的概率。

# 二项分布（2）

- 从抛掷硬币的随机样本中估计p与在实例的随机样本上测试h以估计$error_D(h)$是相同的问题
- 一次硬币抛掷对应于从D中抽取一个实例并测试它是否被h误分类
- 一次随机抛掷出现正面的概率p对应于随机抽取的实例被误分类的概率$error_D(h)$
- 二项分布给出了一个一般形式的概率分布，无论用于表示n次硬币出现正面的次数还是在n个样例中假设出错的次数
- 二项分布的具体形式依赖于样本大小n以及概率p或$error_D(h)$

# 应用二项分布的条件

- 有一基本实验，其输出可被描述为一随机变量Y，随机变量Y有两种取值
- 在实验的任一次尝试中Y=1的概率为常数p，它与其它实验尝试无关，因此Y=0的概率为1-p
- p为预先未知，面临的问题是如何估计
- 基本实验的n次独立尝试按序列执行，生成一个独立同分布的随机变量序列
- 随机变量R表示n次实验中出现$Y_i$=1的次数，它取特定值r的概率由二项分布给出

$$\Pr(R=r)=\frac{n!}{r!(n-r)!}p^r(1-p)^{n-r}$$

$$P(r)=\frac{n!}{r!(n-r)!}error_{\mathcal{D}}(h)^r(1-error_{\mathcal{D}}(h))^{n-r}$$

# 均值

- 期望值是重复采样随机变量得到的值的平均
- 定义：考虑随机变量Y可能的取值为$y_1 \ldots y_n$，Y的期望值E[Y]定义如下：

$$E[Y] = \sum_{i=1}^{n} y_i \Pr(Y = y_i)$$

- 如果随机变量Y服从二项分布，那么可得E[Y]=np

# 方差

- 方差描述的是概率分布的宽度或散度，描述了随机变量与其均值之间的差有多大
- 定义：随机变量Y的方差Var[Y]定义如下：

$$Var[Y] = E[(Y - E(Y))^2]$$

描述了从Y的一个观察值估计其均值E[Y]的误差平方的期望

- 随机变量Y的标准差σY

$$\sigma_Y = \sqrt{E[(Y - E[Y])^2]}$$

- 若随机变量Y服从二项分布，则方差和标准差分别为：

Var[Y]=np(1-p)

$$\sigma_Y = \sqrt{np(1-p)}$$

# 估计量、偏差和方差

- 回到问题：我们得出了随机变量$error_S(h)$服从二项分布，那么$error_S(h)$和$error_D(h)$之间可能的差异是多少？
- 用5.2式定义的二项分布，可得

$$error_S(h)=r/n$$

$$error_D(h)=p$$

- 统计学中将$error_S(h)$称为$error_D(h)$的一个估计量
- 估计量是用来估计总体的某一参数的随机变量，最关心的是它平均来说是否能产生正确估计

# 估计量、偏差和方差（2）

- 估计偏差衡量估计量的期望值同真实参数值之间的差异
- 定义：针对任意参数p的估计量Y的估计偏差是：E[Y]-p
- 如果估计偏差为0，称Y为p的无偏估计量，在此情况下，由多次重复实验生成的Y的多个随机值的平均将收敛于p
- 由于$error_S(h)$服从二项分布，因此$error_S(h)$是$error_D(h)$的一个无偏估计量

# 估计量、偏差和方差（3）

- 对估计偏差的补充说明：
  - 要使$error_S(h)$是$error_D(h)$的无偏估计，假设h和样本S必须独立选取
  - 估计偏差不能与第2章介绍的学习器的归纳偏置相混淆
- 估计量的另一重要属性是它的方差，给定多个无偏估计量，选取其中方差最小的
- 由方差的定义，所选择的应为参数值和估计值之间期望平方误差最小的

# 估计量、偏差和方差（4）

- 一个例子
  - n=40个随机样例
  - r=12个错误
  - $error_S(h)$的标准差
- 一般地，若在n个随机选取的样本中有r个错误，$error_S(h)$的标准差是：

$$\sigma_{error_S(h)} = \frac{\sigma_r}{n} = \sqrt{\frac{p(1-p)}{n}}$$

近似地

$$\sigma_{error_S(h)} = \sqrt{\frac{error_S(h)(1-error_S(h))}{n}} \qquad (5.9)$$

# 置信区间

- 通常描述某估计的不确定性的方法是使用置信区间，真实的值以一定的概率落入该区间中，这样的估计称为置信区间估计
- 定义：某个参数p的N%置信区间是一个以N%的概率包含p的区间
- 由于估计量$error_S(h)$服从二项分布，这一分布的均值为$error_D(h)$，标准差可由式5.9计算，因此，为计算95%置信区间，只需要找到一个以$error_D(h)$为中心的区间，它的宽度足以包含该分布全部概率的95%
- 这提供了一个包围$error_D(h)$的区间，使$error_S(h)$有95%机会落入其中，同样它也指定了$error_D(h)$有95%的机会落入包围$error_S(h)$的区间的大小

# 置信区间（2）

- 对于二项分布，计算置信区间很烦琐，多数情况下，计算它的近似值
- 对于足够大的样本，二项分布可以由正态分布来近似，而正态分布的置信区间容易得到
- 如果随机变量Y服从均值为μ，标准差为σ的一个正态分布，那么Y的任一观察值y有N%的机会落入下面的区间

$$\mu \pm z_N \sigma$$

- 相似地，均值μ有N%的机会落入下面的区间

$$y \pm z_N \sigma$$

# 置信区间（3）

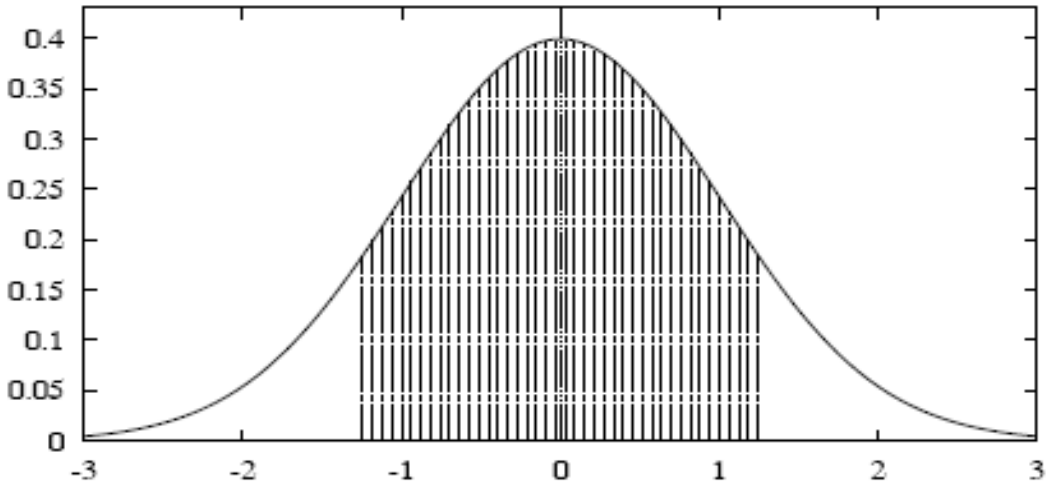
$$error_S(h) \pm z_N \sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$$

- 式子5.1的三步推导过程
  - $error_S(h)$遵从二项分布，其均值为$error_D(h)$，标准差如式5.9所示
  - 对于足够大的样本n，二项分布非常近似于正态分布
  - 式5.1告诉我们如何根据正态分布的均值求出N%置信区间
- 式子5.1的推导中有两个近似
  - 估计$error_S(h)$的标准差，我们将$error_D(h)$近似为$error_S(h)$
  - 用正态分布近似二项分布
- 统计学的一般规则表明，这两个近似在n>=30或np(1-p)>=5时工作得很好，对于较小的n值，最好使用列表的形式给出二项分布的具体值

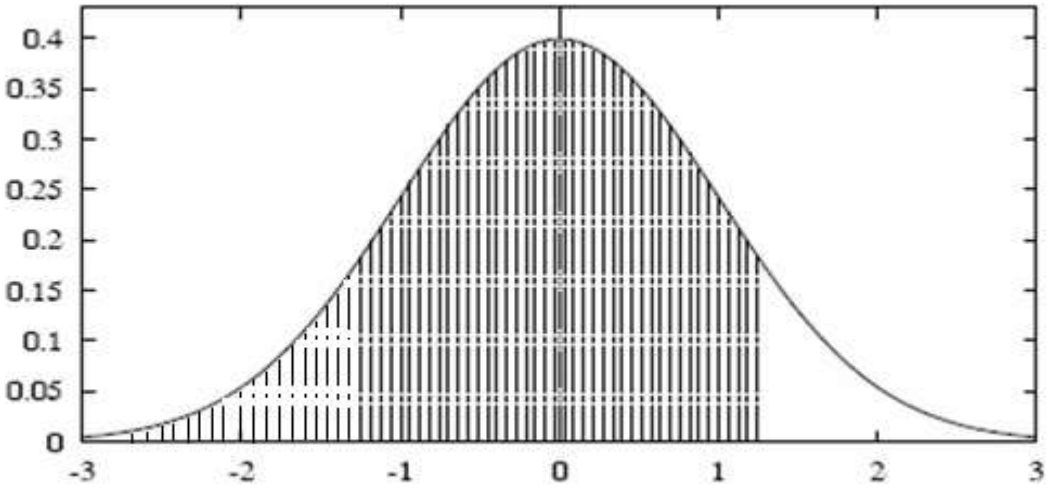
# 双侧和单侧边界

- 上述的置信区间是双侧的，有时用到单侧边界
- 例如问题“$error_D(h)$至多为U的概率”，在只要限定h的最大错误率，而不在乎真实错误率是否小于估计错误率时，很自然提出这种问题
- 由于正态分布关于其均值对称，因此，任意正态分布上的双侧置信区间能够转换为相应的单侧区间，置信度为原来的两倍
- 由一个有下界L和上界U的100(1-α)%置信区间，可得到一个下界为L且无上界的100(1- α/2)%置信区间，也得到一个有上界U且无下界的100(1-α/2)%置信区间

# 均值为0，标准差为1的正态分布




**80%双侧置信区间**




**90%单侧置信区间**

# 推导置信区间的一般方法

- 前面介绍的是针对一特定情况推导置信区间估计：基于独立抽取的n个样本，估计离散值假设的$error_D(h)$
- 下面介绍的方法是在许多估计问题中用到的通用的方法
- 基于大小为n的随机抽取样本的均值，来估计总体均值的问题

# 通用的过程的步骤

- 确定基准总体中要估计的参数p，例如$error_D(h)$
- 定义一个估计量Y（如$error_S(h)$），它的选择应为最小方差的无偏估计量
- 确定控制估计量Y的概率分布DY，包括其均值和方差
- 通过寻找阈值L和U确定N%置信区间，以使这个按DY分布的随机变量有N%机会落入L和U之间

# 思考题

- 如果假设h在n=65的独立抽取样本上出现r=10个错误，真实错误率的90%置信区间是多少？95%的单侧置信区间（上界）是多少？90%的单侧区间是多少？

# 中心极限定理

- 考虑如下的一般框架
  - 在n个独立抽取的且服从同样概率分布的随机变量$Y_1 \ldots Y_n$中观察试验值
  - 令μ代表每一变量$Y_i$服从的未知分布的均值，并令σ代表标准差，称这些变量$Y_i$为独立同分布随机变量
  - 为了估计$Y_i$服从的分布的均值μ，我们计算样本的均值

$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

  - 中心极限定理说明在$n \to \infty$时，$\bar{Y}_n$所服从的概率分布为一正态分布，而不论$Y_i$本身服从什么样的分布
  - $\bar{Y}_n$服从的分布均值为μ，而标准差为 $\frac{\sigma}{\sqrt{n}}$

# 中心极限定理（2）

- 定理5.1（中心极限定理）考虑独立同分布的随机变量$Y_1 \ldots Y_n$的集合，它们服从一任意的概率分布，均值为μ，有限方差为$\sigma^2$，定义样本均值为$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$，当n→∞时，式子 $\frac{\bar{Y}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$ 服从正态分布，均值为0且标准差为1
- 中心极限定理说明在不知道独立的$Y_i$所服从的基准分布的情况下，我们可以得知样本均值$\bar{Y}$的分布形式，说明了怎样使用$\bar{Y}$的均值和方差来确定独立的$Y_i$的均值和方差
- 中心极限定理说明了任意样本均值的估计量服从的分布在n足够大时可以近似为正态分布

# 两个假设错误率间的差异

- 问题：
  - 考虑某离散目标函数的两个假设h1和h2，h1在一拥有n1个随机抽取的样例的样本S1上测试，h2在一拥有n2个从相同分布中抽取的样例的样本S2上测试，要估计这两个假设的真实错误率间的差异

$$d=error_D(h1)-error_D(h2)$$

# 两个假设错误率间的差异（2）

- 使用5.4节中描述的四个步骤来推导d的置信区间估计
  - 确定待估计的参数，如上所述的d
  - 定义一估计量，
  - $\hat{d}$是d的无偏估计量，即E[$\hat{d}$]=d。由于对于较大的n1和n2，$error_{S1}(h1)$和$error_{S2}(h2)$都近似遵从正态分布，两个正态分布的差仍为正态分布，方差为两个正态分布的方差的和

$$\hat{d} = error_{S_1}(h_1) - error_{S_2}(h_2)$$

  - 现在知道了$\hat{d}$服从均值为d、方差为σ2的正态分布，因此d的N%置信区间是

$$\sigma_{\hat{d}}^2 \approx \frac{error_{S_1}(h_1)(1-error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)(1-error_{S_2}(h_2))}{n_2} \quad (5.13)$$

$$\hat{d} \pm z_N \sigma$$

# 两个假设错误率间的差异（3）

- 上面分析的是h1和h2在相互独立的数据样本上测试的情况，如果在同一个样本上测试h1和h2，那么也可以使用公式5.13计算置信区间
- 这种情况下的方差通常小于式子5.12给出的方差，这是因为单个样本消除了两个样本组合带来的随机差异，这样，由式子5.13给出的置信区间一般来说偏于保守，但结果是正确的

# 假设检验

- 有时感兴趣的是某个特定猜想正确的概率，而不是对某参数的置信区间估计。比如：$error_D(h_1)>error_D(h_2)$的可能性有多大？
- 例子，假定分别用大小为100的独立样本S1和S2测量h1和h2的样本错误率为0.30和0.20，给定$\hat{d}=0.10$，问$error_D(h_1)>error_D(h_2)$的概率是多少？d>0的概率是多少？
- 概率Pr(d>0)等于$\hat{d}$对d的过高估计不大于0.1的概率，也就是这个概率为 $\hat{d}$ 落入单侧区间

  $\hat{d}$<d+0.10=$\mu_{\hat{d}}$+0.10的概率

# 假设检验（2）

- 对于$\hat{d}$落入单侧区间$\hat{d}$ <$\mu_{\hat{d}}$+0.10的概率，可以通过计算$\hat{d}$分布在该区间的概率质量来确定$\hat{d}$落入这个单侧区间的概率
- 将区间$\hat{d}$<$\mu_{\hat{d}}$+0.10用允许偏离均值的标准差的数目来重新表示，根据式5.12可得$\sigma_{\hat{d}}$≈0.061，所以这一区间可近似表示为

$$d < \mu_{\hat{d}} + \frac{0.10}{\sigma_{\hat{d}}} \cdot \sigma_{\hat{d}} = \mu_{\hat{d}} + 1.64\sigma_{\hat{d}}$$

- 查表5-1知，$\hat{d}$关于均值的1.64标准差对应置信度$\hat{d}$90%的双侧区间，因此这个单侧区间具有95%的置信度
- 因此给定观察$\hat{d}$=0.1，$error_D(h1) > error_D(h2)$的概率约为95%。使用统计学术语表述为：接受$error_D(h1) > error_D(h2)$假设的置信度是95%

# 学习算法比较

- 有时感兴趣的是比较两个学习算法的性能，而不是两个具体的假设本身
  - 如何近似地检验多个学习算法？
  - 如何确定两个算法之间的差异在统计上是有意义的？
- 假定有$L_A$和$L_B$两个算法，要确定为了学习一特定目标函数f，平均来说那个算法更好
- 定义“平均”的一种合理方法是，从一基准实例分布中抽取包含n个样例的训练集合，在所有这样的集合中测量两个算法的平均性能，即

$$\underset{S \subset D}{E}[error_D(L_A(S)) - error_D(L_B(S))] \quad (5.14)$$

# 学习算法比较（2）

- 在实**际的学习算法比较中，我们只有一个有限的样本$D_0$，把它分割成训练集合$S_0$和测试集合$T_0$**，使用下式比较两个学习到的假设的准确度

$$error_{T_0}(L_A(S_0)) - error_{T_0}(L_B(S_0)) \quad （5.15）$$

- 上式与5.14有两个重要的不同
  - 使用$error_{T0}(h)$来近似$error_D(h)$
  - 错误率的差异测量是在一个训练集合$S_0$上，而不是在从分布D中抽取的所有样本S上计算的期望值
- 改进5.15式的一种方法是，将数据$D_0$多次分割为不相交的训练和测试集合，然后在其中计算这些不同的实验的错误率的平均值$\bar{\delta}$

# 学习算法比较（3）

## K-Fold 交叉验证

Randomly partition data $D$ into $k$ disjoint equal-sized

subsets $P_1 \ldots P_k$

For $i$ from 1 to $k$ do:

Use $P_i$ for the test set and remaining data for training

$S_i = (D - P_i)$

$h_A = L_A(S_i)$

$h_B = L_B(S_i)$

$\delta_i = \text{error}_{Pi}(h_A) - \text{error}_{Pi}(h_B)$

Return the average difference in error: $\bar{\delta} = \frac{1}{k}\sum_{i=1}^{k}\delta_i$

# 学习算法比较（4）

- 算法返回的$\bar{\delta}$可看作下式的估计

$$\underset{S\subset D_0}{E}[error_D(L_A(S))-error_D(L_B(S))] \quad （5.17）$$

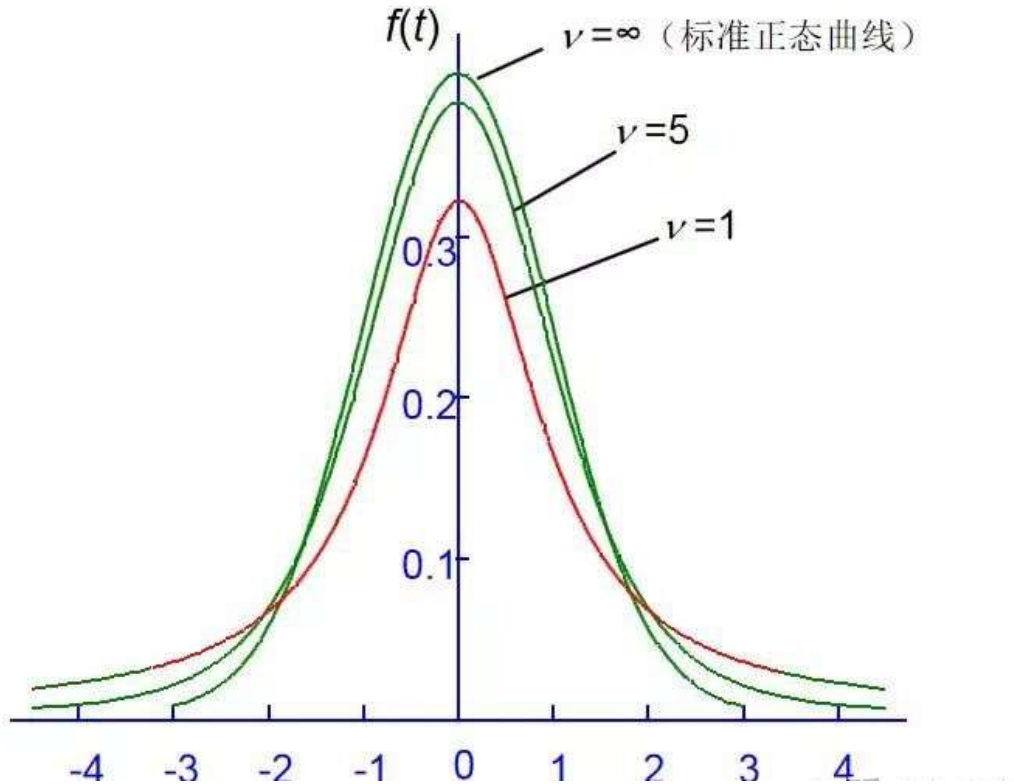
- 估计式5.17的近似的N%置信区间可表示成 $\bar{\delta}\pm t_{N,k-1}s_{\bar{\delta}}$（5.18），其中$t_{N,k-1}$是一常数，其意义类似于前面的$z_N$，第一个下标表示所需的置信度，第二个下标表示自由度，常记作v，它与生成随机变量$\bar{\delta}$的值时独立的随机事件数目相关。而$s_{\bar{\delta}}$代表$\bar{\delta}$所服从的概率分布的标准差的估计，定义如下:

$$s_{\bar{\delta}}=\sqrt{\frac{1}{k(k-1)}\sum_{i=1}^{k}(\delta_i-\bar{\delta})^2} \quad (5.19)$$

- 注意当自由度v→∞时，$t_{N,v}$的值趋向常数$z_N$。

# t分布

- 类似于正态分布的钟形分布，但更宽且矮，以反映使用$\sigma_{\bar{Y}}$近似真实的标准差$s_{\bar{Y}}$时带来的更大方差。



# 学习算法比较（5）

- 这里描述的比较学习算法的过程要在同样的测试集合上测试两个假设，这与前面描述的比较两个用独立测试集合评估过的假设不同。
- 使用相同样本来测试假设被称为配对测试，配对测试通常会产生更紧密的置信区间，因为在配对测试中任意的差异都来源于假设之间的差异。
- 若假设在分开的数据样本上的测试，两个样本错误率之间的差异也可能部分来源于两个样本组成的不同。

# 配对t测试

- 前面主要讨论给定固定数据集时比较两个学习算法的过程
- 本节将论证公式5.18和5.19
- 为了理解5.18中的置信区间，考虑一下的估计问题
  - 给定一系列独立同分布的随机变量$Y_1 \ldots Y_k$的观察值
  - 要估计这些$Y_i$所服从的概率分布的均值$\mu$
  - 使用的估计量为样本均值

$$\bar{Y} = \frac{1}{k}\sum_{i=1}^{k} Yi$$

# 配对t测试（2）

- 这一基于样本均值估计分布均值的问题非常普遍（比如，早先的用$error_S(h)$估计$error_D(h)$）
- 由式5.18和5.19描述的t测试应用于该问题的一特殊情形，即每个单独的$Y_i$都遵循正态分布
- 考虑前面比较学习算法的过程的一个理想化形式，假定不是拥有固定样本数据$D_0$，而是从基准实例分布中抽取新的训练样例，使每一次循环需要的训练集$S_i$和测试集$T_i$是从基准实例分布中抽取
- 这一理想化方法能很好地匹配上面的估计问题，该过程所测量的$\delta_i$对应独立同分布的随机变量$Y_i$，其分布的均值$\mu$对应两学习算法错误率的期望差异。

# 配对t测试（3）

- 测试集$T_i$至少包含30个样例，因此，单独的$\delta_i$将近似遵循正态分布，因此，我们也要求$Y_i$服从近似的正态分布，样本均值 $\bar{Y}$ 也遵循正态分布
- 由此，可以考虑使用前面的计算置信区间的表达式。然而，该公式要求我们知道这个分布的标准差，但这个标准差未知
- t测试正好用于这样的情形，即估计一系列独立同正态分布的随机变量的样本均值
- 当k趋近于无穷时，t分布趋近于正态分布，即$t_{N,k-1}$趋近于$z_N$，因为样本规模k增加时，$S_{\bar{Y}}$收敛到真实的标准差，并且当标准差确切已知时可使用$z_N$。

# 实际考虑

- 上面的讨论说明了在使用样本均值来估计一个包含k个独立同正态分布的随机变量的样本均值时，可使用式5.18来估计置信区间；
- 这个结论假定对于目标函数的样例可进行无限存取，实际问题是随机变量之间并不独立，因为它们基于从有限子集中抽取的相互重叠的训练样例；
- 当只有一个有限的数据样本可用时，有几种重叠采用的方法。
  - 前面描述了k-fold方法(交叉检验)
  - 随机抽取至少有30个样例的测试集合，剩余样例组成训练集合，重复这一过程直到足够的次数

# 实际考虑（2）

- 随机方法的好处是能够重复无数次，以减少置信区间到需要的宽度，而k-fold方法受限于样例的总数
- 随机方法的缺点是，测试集合不再被看作是从基准实例分布中独立抽取，而k-fold交叉验证生成的测试集合是独立的，因为一个实例只在测试集合中出现一次
- 概括而言，统计学模型在数据有限时很少能完美地匹配学习算法验证中的所有约束。然而，它们确实提供了近似的置信区间。

# 小结

- 统计理论提供了一个基础，从而基于在数据样本S上的观察错误率，估计真实错误率。
- 估计置信区间的问题可通过一待估计的参数以及相对应的估计量来完成。由于估计量是一个随机变量，它可由其服从的概率分布来描述。置信区间的计算可通过确定该分布下包含所需概率质量的区间来描述。
- 估计假设精度中的一种可能误差为估计偏差。如果Y为对某参数p的估计量，Y的估计偏差为Y的期望值和p之间的差

# 小结（2）

- 估计产生误差的第二种原因是估计方差。即使对于无偏估计，估计量的观察值也可能在各实验中不同，估计量分布的方差描述了该估计与真实值的不同有多大。方差在数据样本增大时降低。
- 比较两个学习算法效果的问题在数据和时间无限时是一个相对容易的估计问题，但在资源有限时要困难得多。本章描述的一种途径是在可用数据的不同子集上运行学习算法，在剩余数据上测试学到的假设，然后取这些结果的平均值。
- 本章考虑的多数情况中，推导置信区间需要多个假定和近似。近似计算分布的方差，以及假定实例从一固定不变的概率分布中生成。

# 思考题

- 要测试一个假设h，其$error_D(h)$已知在0.2到0.6的范围内。要保证95%双侧置信区间的宽度小于0.1，最少应搜索的样例数是多少?

# 补充读物

- Billingsley et al.提供了对统计学的一个很简明的介绍，详尽讨论本章涉及的一些问题
- DeGroot、Casella & Berger、Duda & Hart在数值模式识别领域提出了对这些问题的解决方法
- Segre et al.、Etzioni & Etzioni、Gordon & Segre讨论了评估学习算法的统计意义测试
- Geman et al.讨论了在同时最小化偏差和最小化方差之间作出的折中。Dietterich讨论了在不同训练测试数据分割下使用配对差异t测试带来的风险