



机器学习

Machine Learning



主讲人：张敏 清华大学长聘副教授



机器学习

MACHINE LEARNING-MIN ZHANG

Unit.02

机器学习实验方法与原则 (II)

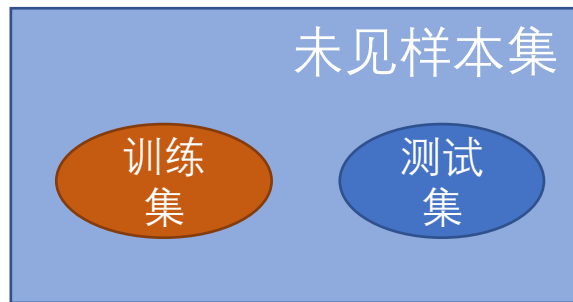
*图片均来自网络或已发表刊物

机器学习实验方法与原则 2

- 评价指标
- 训练集、验证集与测试集
- 随机重复实验
- K折交叉验证
- 统计有效性检验

训练集与测试集

- 训练集 ([作业](#))：模型可见样本标签，用于训练模型，样本数有限
 - 在训练集上表现好的模型，在其他未见样本上一定表现好吗？小心**过拟合**！
- 未见样本 ([所有没做过的题](#)) 往往有指数级别或无穷多个
- 测试集 ([考试](#))：用于评估模型在可能出现的未见样本上的表现
 - 尽可能与训练集**互斥**，即测试样本尽量**不在训练集中出现**，为什么？
 - **估计**模型在整个未见样本上的表现



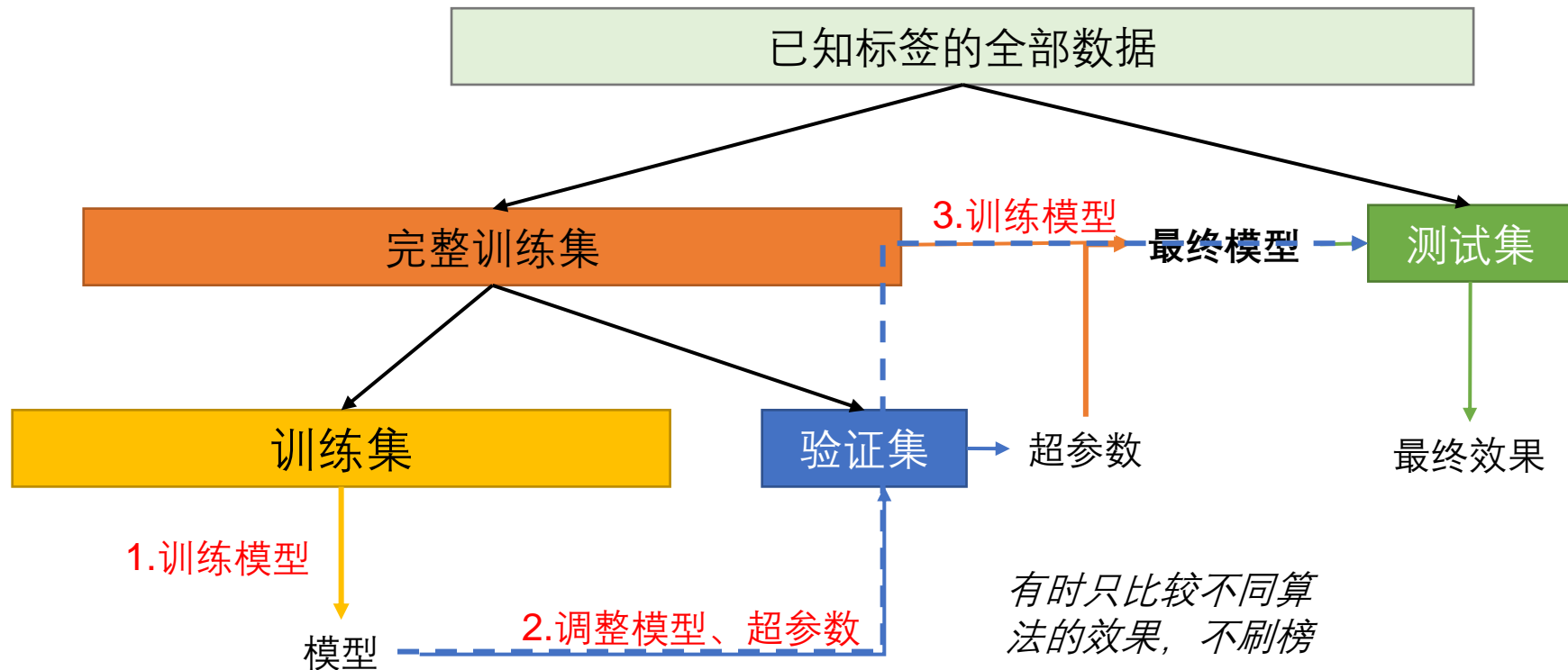
训练集与测试集的划分方式

- 随机划分
 - 按比例，例如9:1、8:2
 - 固定数目，例如测试集从全部样本中采样1w个，其余为训练集
- 留一化分 (leave-one-out)
 - 一个样本作测试，其余样本训练：常用于K近邻等算法的性能评估
- 特殊划分
 - 按时间划分，例如1-5月气象数据作训练，6月气象数据作测试
 - 推荐系统中，常把每个用户交互序列的最后一个样本作测试，其余作训练
 -

验证集

- 从训练集中额外分出的集合，一般用于超参数的调整
 - 训练轮次、正则化权重、学习率等等
- 为什么不在训练集上调整超参数？过拟合训练集
- 为什么不在测试集上调整超参数？过拟合测试集
 - 针对当前测试集调出的参数可能只在当前测试集上较好
 - 使得测试集结果偏高，不能反映实际在所有未见样本上的效果
 - 类比：针对某场考试的知识点分布作重点复习，不能准确反映学生对所有知识的掌握程度。
 - 举例：机器学习竞赛中，针对公开部分的测试数据过度调参，不一定在隐藏的全部测试数据上表现好。

训练集、验证集与测试集

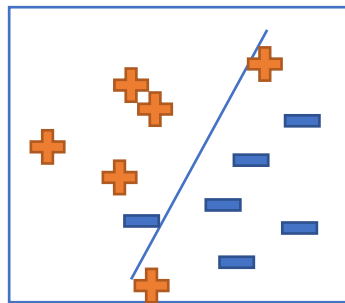
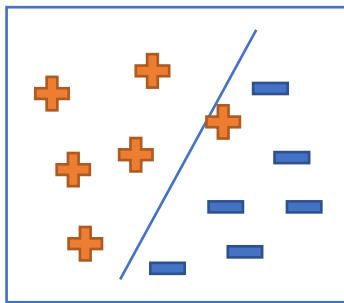


机器学习实验方法与原则 3

- 评价指标
- 训练集、验证集与测试集
- 随机重复实验
- K折交叉验证
- 统计有效性检验

随机重复实验

- 测一次就足够了吗？
 - 极端情况：二分类中分类器**随机**输出，**恰好**测试集都对了（效果最好？）
- **数据**随机性
 - 由数据集划分带来的评价指标波动
- **模型**随机性
 - 由模型或学习算法本身带来的评价指标波动
 - 例如：神经网络初始化、训练批次生成



随机重复实验

- 数据随机性

- (数据足够多时) 增多测试样本
- (数据量有限时) 重复多次划分数据集

- 模型随机性

- 更改随机种子重复训练、测试

- **注意**: 保持每次得到的评价指标**独立同分布(iid)**

- 报告结果: 评价指标的**均值** $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

- **样本标准差**(个体离散程度, 反映了个体对样本均值的代表性) $S = \sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)}$
- **标准误差**(样本均值的离散程度, 反映了样本均值对总体均值的代表性)

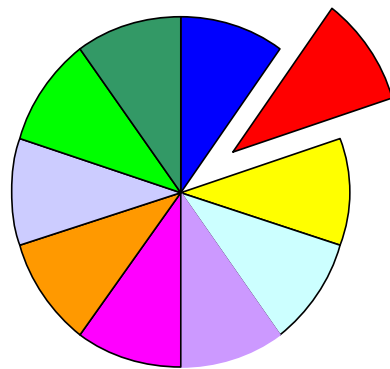
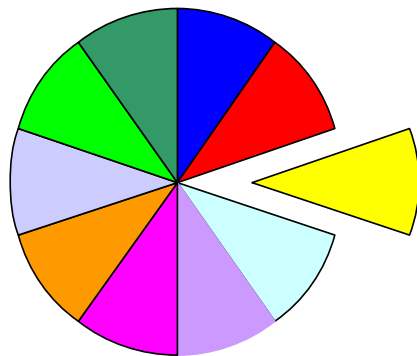
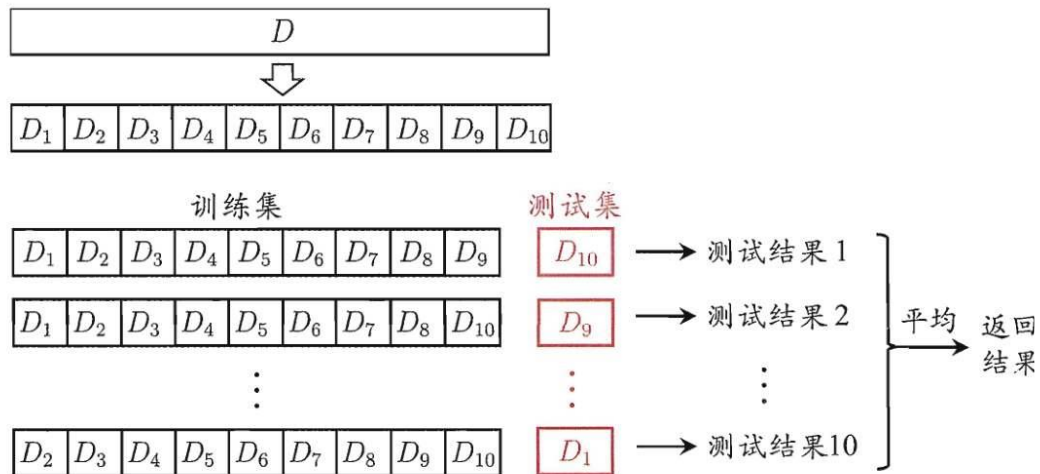
$$\text{Standard Error of the Mean, SEM} = \frac{S}{\sqrt{n}} = \sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 / (n(n - 1))}$$

机器学习实验方法与原则 4

- 评价指标
- 训练集、验证集与测试集
- 随机重复实验
- **K折交叉验证**
- 统计有效性检验

K折交叉验证

- 随机把数据集分成K个相等大小的不相交子集



K折交叉验证

- 优点：数据利用率高，适用于数据较少时
- 缺点：训练集互相有交集，每一轮之间并不满足独立同分布
- 增大K，一般情况下：
 - 所估计的模型效果偏差 (bias) 下降
 - 所估计的模型效果方差 (variance) 上升
 - 计算代价上升，更多轮次、训练集更大
- K一般取5、10

机器学习实验方法与原则 5

- 评价指标
- 训练集、验证集与测试集
- 随机重复实验
- K折交叉验证
- 统计有效性检验

假设的评估检验：问题1

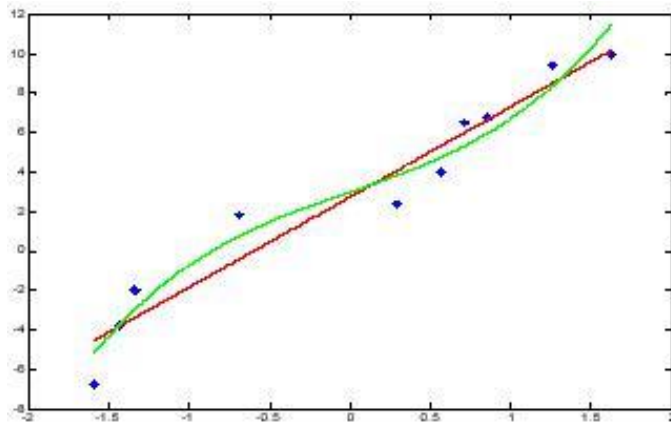
- 效果估计

- 给定一个假设在**有限量数据**上的准确率
- 该准确率是否能准确估计**在其它未见数据上**的效果？



假设的评估检验：问题2

- h_1 在数据的一个样本集上表现优于 h_2
- h_1 总体上更好的概率有多大？



抽样理论基础



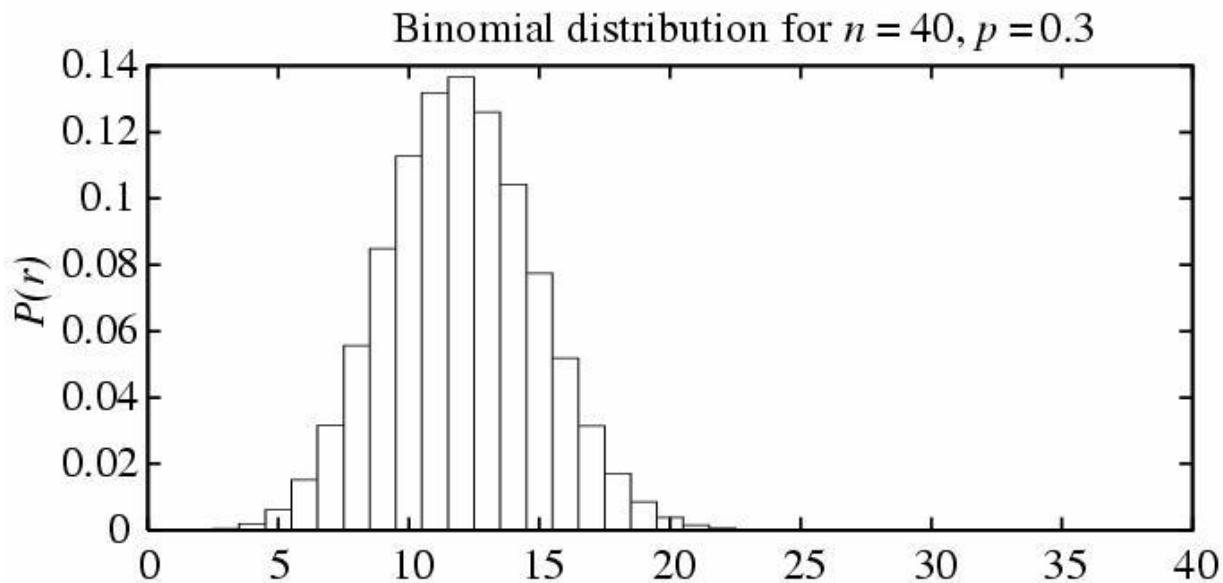
- 伯努利实验
 - 只有 2 种输出:
成功概率: p , 失败概率: $q=1-p$
 - 用随机变量 X 记录成功的次数

- 伯努利分布:

- 抛硬币: 正面朝上的概率为 p , 抛 n 次, 观察到 r 次正面朝上
- 若计 $X \sim B(n, p)$, $Pr(X=r) = P(r)$, 则

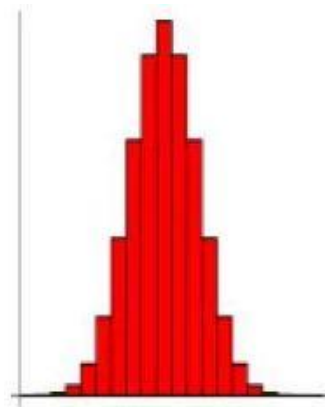
$$P(r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$

二项分布 (Binomial Distribution)



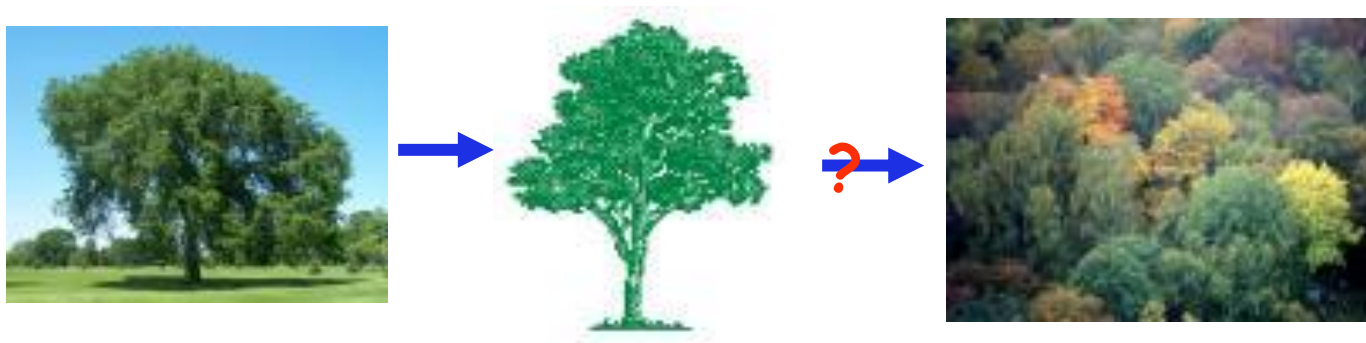
二项分布的应用场景

- 两个可能的输出 (成功/失败) ($Y=0$ 或 $Y=1$)
- 每次尝试成功的概率相等 $Pr(Y=1) = p$, 其中 p 是一个常数
- n 次独立尝试
 - 随机变量 Y_1, \dots, Y_n ,
 - iid (independent identically distribution, 独立同分布)
 - R : 随机变量, n 次尝试中 $Y_i = 1$ 的次数,
- $Pr(R=r) \sim$ 二项分布
- 平均 (期望值): $E[R], \mu$
 - 二项分布: $\mu = np$
- 方差: $Var[R]=E[(R-E[R])^2], \sigma^2$ (标准差 σ)
 - 二项分布: $\sigma^2 = np(1-p)$



回顾 – 问题1

- 效果估计
 - 给定一个假设在有限量数据上的准确率
 - 该准确率是否能准确估计在其它未见数据上的效果？



估计假设准确率 – Q1.1解答

如何对一个假设 h 在来自同一分布的未见样本上的准确率作出最好的估计？

n 个随机样本中有 r 个被误分类的概率 – 二项分布 $P(r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$

$$error_D(h) = p \quad error_S(h) = r/n \quad \mathbf{n=100, r=12 \quad 12\%} \quad \mathbf{n=25, r=3 \quad 12\%}$$

$$E[r] = np, \quad E[error_S(h)] = E\left[\frac{r}{n}\right] = \frac{np}{n} = p = error_D(h)$$

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

$$\sigma_{error_S(h)} = \frac{\sigma_r}{n} \approx \sqrt{\frac{error_S(h)(1-error_S(h))}{n}} \quad \mathbf{3.2\%} \quad \mathbf{6.5\%}$$

估计的两个重要性质

• 估计偏差 (Bias)

- 如果 S 是训练集, $error_S(h)$ 是有偏差的 (偏乐观),

$$bias \equiv E[error_S(h)] - error_D(h)$$

- 对于无偏估计($bias = 0$), h 和 S 必须独立不相关地产生

→ 不要在训练集上测试!

• 估计方差 (Varias)

- 即使是 S 的无偏估计, $error_S(h)$ 可能仍然和 $error_D(h)$ 不同

- E.g. 之前的例子 (3.2% vs. 6.5%)

- 需要选择无偏的且有最小方差的估计

估计假设准确率 – Q1.2解答

准确率的估计可能包含多少错误？

($error_S(h)$ 对 $error_D(h)$ 的估计有多好?)

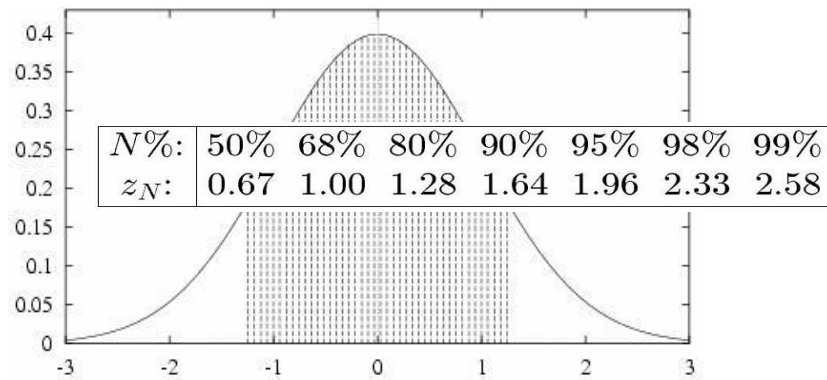
- 抽样理论: **confidence interval** (置信区间)
- 定义:
 - 参数 p 的 $N\%$ 置信区间是一个以 $N\%$ 的概率包含 p 的区间, $N\%$: 置信度
- ✓ 90.0%的置信度, 年龄: [12, 24]
- ✓ 99.9%的置信度, 年龄: [3, 60]



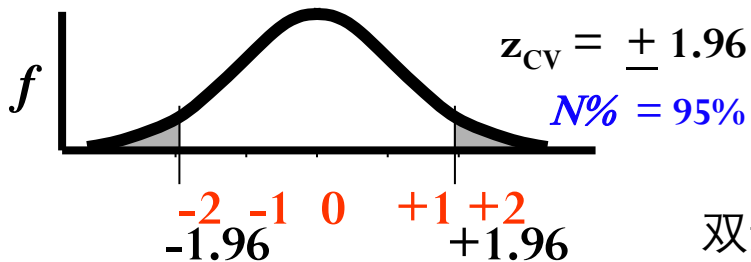
置信度与置信区间

- 如何得到置信区间？
 - 坏消息: 对二项分布来说很难
 - 好消息: 对正态分布来说很简单
 - 通过正态分布的某个区间(面积) 来获得

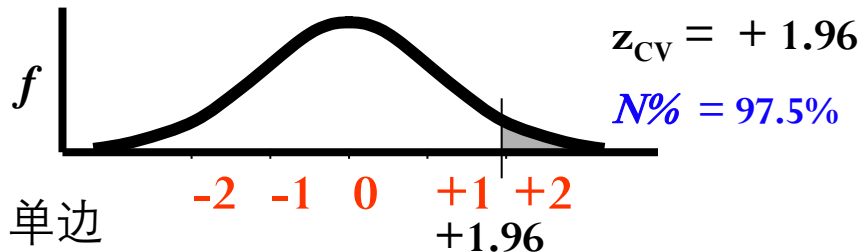
• 正态分布的置信度和置信区间



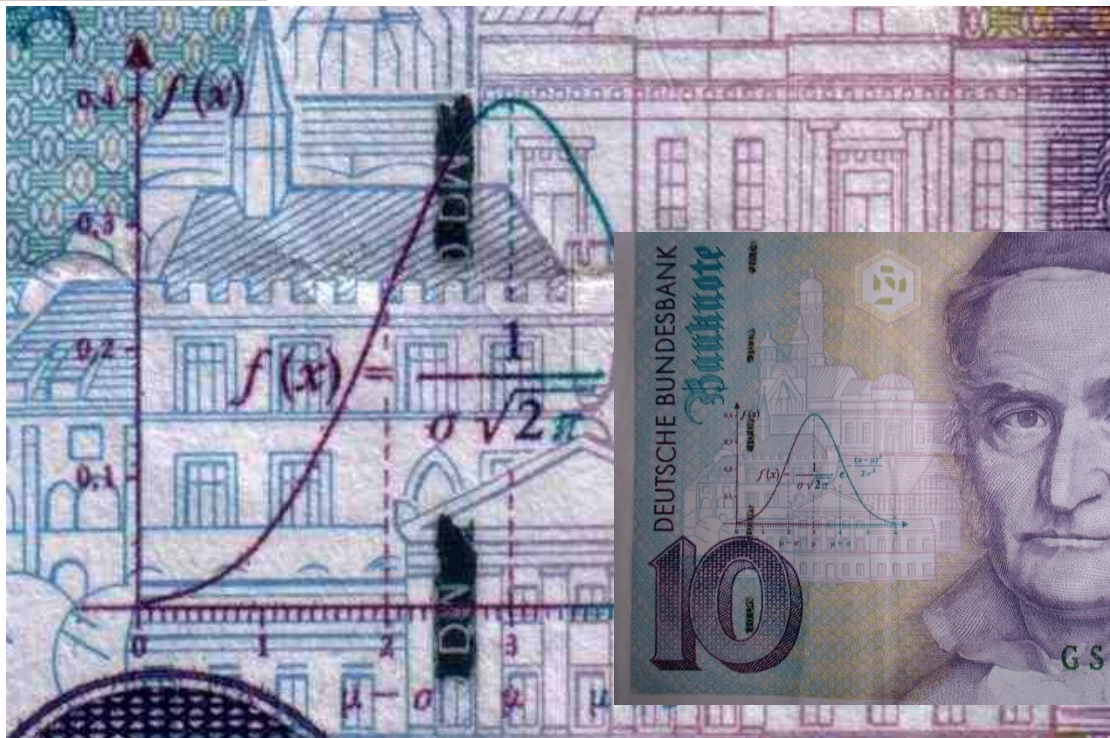
- 指标 y 有 $N\%$ 的可能性落在区间 $\mu \pm z_N \sigma$
- 等价于, 均值 μ 有 $N\%$ 的可能性落在区间 $y \pm z_N \sigma$



双边 vs 单边



正态分布



正态分布 & 二项分布

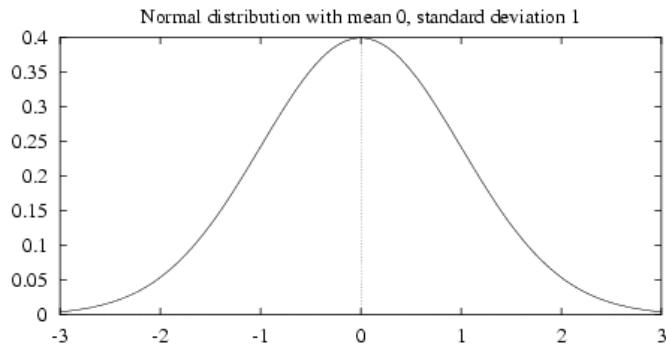
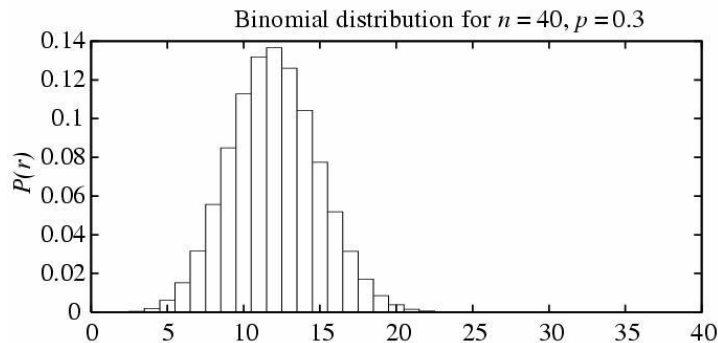
- 对于足够大的采样大小

二项分布



可以通过正态分布近似表示

- 经验法则: $n > 30$, $np(1-p) > 5$

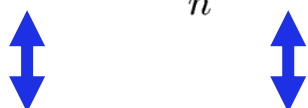


估计假设准确率 – Q1.2解答

- 如果满足以下条件，估计更准确：
 - S 包含 $n \geq 30$ 个样本，与 h 独立产生，且每个样本独立采样
- 那么有大约95%的概率 $error_S(h)$ 落在区间

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

- 等价于， $error_D(h)$ 落在区间

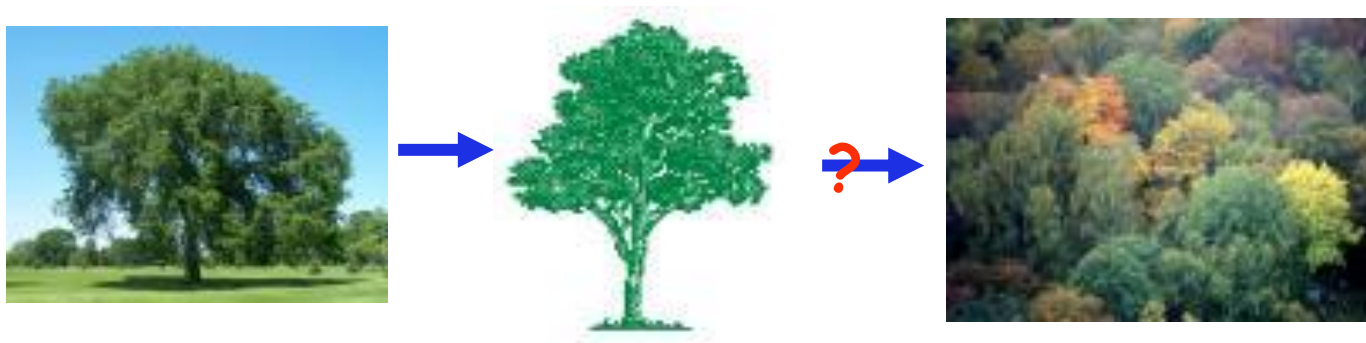
$$error_S(h) \pm 1.96 \sqrt{\frac{error_D(h)(1 - error_D(h))}{n}}$$


- 近似等于，

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

回顾 – 问题1

- 效果估计
 - 给定一个假设在有限量数据上的准确率
 - 该准确率是否能准确估计在其它未见数据上的效果?



问题1解答总结

- 问题设定:
 - S : n 随机独立样本, 且独立于假设 h
 - $n \geq 30$ & h 有 r 个错误
- 真实错误率 $error_D$ 落在以下区间有 $N\%$ 置信度:

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

$N\%$:	50%	68%	80%	90%	95%	98%	99%
z_N :	0.67	1.00	1.28	1.64	1.96	2.33	2.58

推导置信区间的一般方法

- 一般地,
 - 确定需要估计的变量 p , e.g. $error_D(h)$
 - 确定估计量 Y (偏差, 方差), e.g. $error_S(h)$
 - 希望: 小方差, 无偏估计
 - 确定 Y 的分布 D (包括均值 & 方差)
 - 确定 $N\%$ 置信区间 ($L..U$)
 - 可能有 $L=-\infty$ or $U=\infty$
 - E.g. (对于正态分布) 利用 z_n 表查询相关值
- 也可应用在其他问题上

推导置信区间的一般方法

- 一般地,
 - 确定需要估计的变量 p , e.g. $error_D(h)$
 - 确定估计量 Y (偏差, 方差), e.g. $error_S(h)$
 - 希望: 小方差, 无偏估计
 - 确定 Y 的分布 D (包括均值 & 方差)
 - 确定 $N\%$ 置信区间 ($L..U$)
 - 可能有 $L=-\infty$ or $U=\infty$
 - E.g. (对于正态分布) 利用 z_n 表查询相关值
- 也可应用在其他问题上



中心极限定理

- 简化了求解置信区间的过程
- 问题设定
 - 独立同分布Independent, identically distributed (iid) 的随机变量 Y_1, \dots, Y_n ,
 - 未知分布, 有均值 μ 和有限方差 σ^2
 - 估计均值: $\bar{Y} \equiv \frac{1}{n} \sum_{i=1}^n Y_i$
- 中心极限定理
 - \bar{Y} 服从正态分布 ($n \rightarrow \infty$)
 - 均值 μ , 方差 σ^2/n
 - 可以被归一化到标准正态分布, 即 $\mu = 0, \sigma = 1$

中心极限定理...

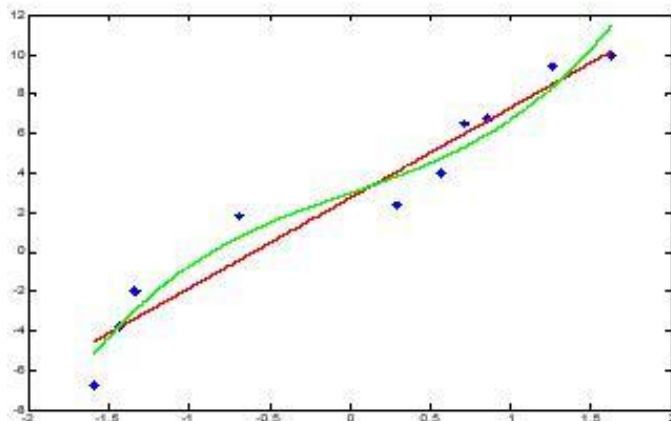
- 样本均值 \bar{Y} 的分布
 - 是已知的
 - 即使 Y_i 的分布是未知的
 - 可以用来确定的 Y_i 均值方差
- 提供了估计的基础
 - 估计量的分布
 - 一些样本的均值

问题2

- h_1 在数据的一个样本集上表现优于 h_2
 - h_1 总体上更好的概率有多大？



假设之间的差异



假设间的差异

- 在样本集合 S_1 (n_1 个随机样本) 上测试 h_1 , 在 S_2 (n_2) 上测试 h_2
- 选择要估计的参数 $d \equiv error_{\mathcal{D}}(h_1) - error_{\mathcal{D}}(h_2)$
- 选择估计量
 - 无偏的 $\hat{d} \equiv error_{S_1}(h_1) - error_{S_2}(h_2)$
- 确定估计量 \hat{d} 所服从的概率分布
 - $error_{S_1}(h_1)$, $error_{S_2}(h_2)$ 近似服从正态分布
 - \hat{d} 也近似正态分布
 - 均值 = d
 - 方差: 加和

* 证明: http://en.wikipedia.org/wiki/Sum_of_normally_distributed_random_variables

假设间的差异

- 在样本集合 S_1 (n_1 个随机样本) 上测试 h_1 , 在 S_2 (n_2) 上测试 h_2
- 选择要估计的参数 $d \equiv error_D(h_1) - error_D(h_2)$
- 选择估计量
 - 无偏的 $\hat{d} \equiv error_{S_1}(h_1) - error_{S_2}(h_2)$

- 确定估计量所服从的正态分布

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

- 确定区间 (L, U) 满足 $N\%$ 的概率落在区间

$$\hat{d} \pm z_N \sqrt{\frac{error_{S_1}(h_1)(1 - error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)(1 - error_{S_2}(h_2))}{n_2}}$$

假设检验

- 某些陈述可能是真的概率

- E.g. 例如 $e_D(h_1) > e_D(h_2)$ 的概率

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

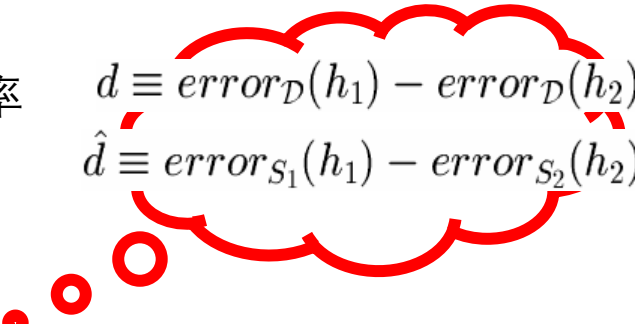
- 例子 ($n_1=n_2=100$)

- $e_{S_1}(h_1) = 0.3$, $e_{S_2}(h_2) = 0.2$, 求 $e_D(h_1) > e_D(h_2)$ 的概率

- 给定 $\hat{d} = 0.1$, 求 $e_D(h_1) > e_D(h_2)$ 的概率

假设检验

- 某些陈述可能是真的概率
 - E.g. 例如 $e_D(h_1) > e_D(h_2)$ 的概率
- 例子 ($n_1=n_2=100$)
 - $e_{S1}(h_1) = 0.3, e_{S2}(h_2) = 0.2$
 - 给定 $\hat{d} = 0.1$, 求 $e_D(h_1) > e_D(h_2)$ 的概率
 - 给定 $\hat{d} = 0.1$, 求 $d > 0$ 的概率



$$d \equiv error_D(h_1) - error_D(h_2)$$
$$\hat{d} \equiv error_{S1}(h_1) - error_{S2}(h_2)$$

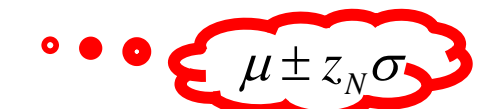
假设检验

- 某些陈述可能是真的概率
 - E.g. 例如 $e_D(h_1) > e_D(h_2)$ 的概率

- 例子 ($n_1=n_2=100$)

- $e_{S1}(h_1) = 0.3, e_{S2}(h_2) = 0.2$
- 给定 $\hat{d} = 0.1$, 求 $e_D(h_1) > e_D(h_2)$ 的概率
- 给定 $\hat{d} = 0.1$, 求 $d > 0$ 的概率
- \hat{d} 在区间 $d + 0.1 > \hat{d}$ 的概率
 - 注意: d 是 \hat{d} 概率分布的均值
- \hat{d} 在区间 $\hat{d} < \mu_{\hat{d}} + 0.1$ 的概率


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


$$\mu \pm z_N \sigma$$

统计有效性检验: (z检验) 举例 (1)

- $n_1 = n_2 = 100$, $acc_{S_1}(h_1) = 0.3$, $acc_{S_2}(h_2) = 0.2$
- 则 $\hat{d} \equiv acc_{S_1}(h_1) - acc_{S_2}(h_2) = 0.1$
- 求 $d \equiv acc_D(h_1) - acc_D(h_2) > 0$ 的置信度
- 即求 $\hat{d} < d + 0.1$ 的概率
- 又有 $\sigma_{\hat{d}} \approx \sqrt{\frac{1}{n_1} acc_{S_1}(h_1) (1 - acc_{S_1}(h_1)) + \frac{1}{n_2} acc_{S_2}(h_2) (1 - acc_{S_2}(h_2))} = 0.061$
- 则 $\hat{d} < d + 0.1 \rightarrow \hat{d} < d + 1.64 \sigma_{\hat{d}}$
- 即 $z_N = 1.64$, 查正态分布表可知, 双边置信度为90%
- 则单边置信度为95%
- 即 $acc_D(h_1) > acc_D(h_2)$ 的置信度为95%

$N\%$:	50%	68%	80%	90%	95%	98%	99%
z_N :	0.67	1.00	1.28	1.64	1.96	2.33	2.58

统计有效性检验: (z检验) 举例 (2)

- $n_1 = n_2 = 30$, $acc_{S_1}(h_1) = 0.3$, $acc_{S_2}(h_2) = 0.2$
- 则 $\hat{d} \equiv acc_{S_1}(h_1) - acc_{S_2}(h_2) = 0.1$
- 求 $d \equiv acc_D(h_1) - acc_D(h_2) > 0$ 的置信度
- 即求 $\hat{d} < d + 0.1$ 的概率
- 又有 $\sigma_{\hat{d}} \approx \sqrt{\frac{1}{n_1} acc_{S_1}(h_1) (1 - acc_{S_1}(h_1)) + \frac{1}{n_2} acc_{S_2}(h_2) (1 - acc_{S_2}(h_2))} = 0.111$
- 则 $\hat{d} < d + 0.1 \rightarrow \hat{d} < d + 0.90 \sigma_{\hat{d}}$
- 即 $z_N = 0.90$, 查正态分布表可知, 双边置信度为 68%
- 则单边置信度为 84%
- 即 $acc_D(h_1) > acc_D(h_2)$ 的置信度为 84%

$N\%$:	50%	68%	80%	90%	95%	98%	99%
z_N :	0.67	1.00	1.28	1.64	1.96	2.33	2.58

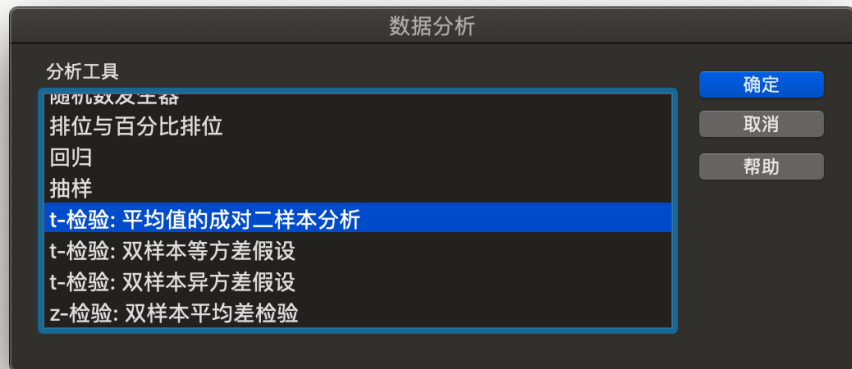
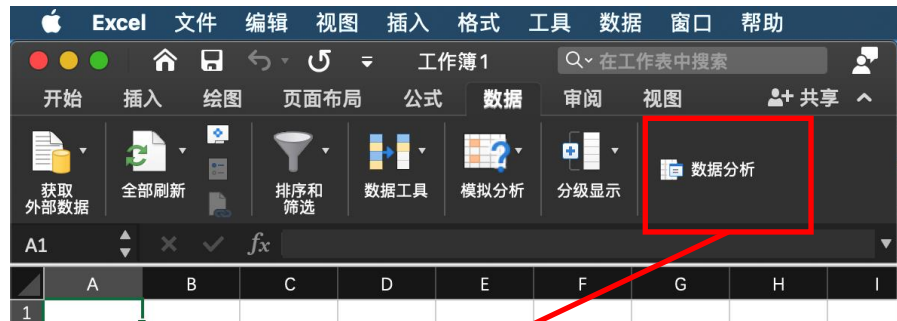
统计有效性检验：t检验

- 记模型 h_1 的 n_1 次重复实验结果为 $x_{11}, x_{12}, \dots, x_{1n_1}$ $\bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{1i}$, $s_1^2 = \frac{1}{n_1-1} \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2$
- 记模型 h_2 的 n_2 次重复实验结果为 $x_{21}, x_{22}, \dots, x_{2n_2}$ $\bar{x}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} x_{2i}$, $s_2^2 = \frac{1}{n_2-1} \sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2$
- 在样本量及方差均不相等的假设下有
 - 检验量 $t = (\bar{x}_1 - \bar{x}_2) / \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, 自由度 $df = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2 / \left(\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)} \right)$
- 若 $n_1 = n_2 = n$ 且 $d_i = x_{1i} - x_{2i}$ 独立且来自正态分布
 - 可采用配对t检验 (paired t-test), 例如两组结果测试集依次相同时
 - 即两个模型在同样划分的交叉验证或同样测试集的重复对比实验
 - 检验量 $t = (\bar{x}_1 - \bar{x}_2) / \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n(n-1)}}$, 自由度为 $n-1$
- 根据检验量和自由度查t分布表可得置信度 (类似根据 z_N 查正态分布表)

统计有效性检验

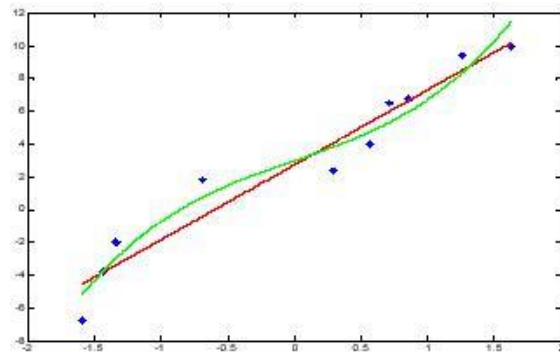
T 检验临界值表

自由度			显著性水平 (α)			自由度			显著性水平 (α)		
(df)	0.10	0.05	0.01	(df)	0.10	0.05	0.01	(df)	0.10	0.05	0.01
$n-m-1$				$n-m-1$				$n-m-1$			
1	6.314	12.706	63.657	301	1.650	1.968	2.592	301	1.650	1.968	2.592
2	2.920	4.303	9.925	302	1.650	1.968	2.592	302	1.650	1.968	2.592
3	2.353	3.182	5.841	303	1.650	1.968	2.592	303	1.650	1.968	2.592
4	2.132	2.776	4.604	304	1.650	1.968	2.592	304	1.650	1.968	2.592
5	2.015	2.571	4.032	305	1.650	1.968	2.592	305	1.650	1.968	2.592
6	1.943	2.447	3.707	306	1.650	1.968	2.592	306	1.650	1.968	2.592
7	1.895	2.365	3.499	307	1.650	1.968	2.592	307	1.650	1.968	2.592
8	1.860	2.306	3.355	308	1.650	1.968	2.592	308	1.650	1.968	2.592
9	1.833	2.262	3.250	309	1.650	1.968	2.592	309	1.650	1.968	2.592
10	1.812	2.228	3.169	310	1.650	1.968	2.592	310	1.650	1.968	2.592
11	1.796	2.201	3.106	311	1.650	1.968	2.592	311	1.650	1.968	2.592
12	1.782	2.179	3.055	312	1.650	1.968	2.592	312	1.650	1.968	2.592
13	1.771	2.160	3.012	313	1.650	1.968	2.592	313	1.650	1.968	2.592
14	1.761	2.145	2.977	314	1.650	1.968	2.592	314	1.650	1.968	2.592
15	1.753	2.131	2.947	315	1.650	1.968	2.592	315	1.650	1.968	2.592
16	1.746	2.120	2.921	316	1.650	1.967	2.591	316	1.650	1.967	2.591
17	1.740	2.110	2.898	317	1.650	1.967	2.591	317	1.650	1.967	2.591



统计有效性检验(总结)

- 比较算法A和B的优劣
 - 准确率均值高就一定好？有随机性
 - A比B高多少才能有把握说A算法更好？显著性检验
- 随机变量的样本个数较多时(一般 >30): **z检验**(利用中心极限定理)
 - 一般用于单次评测, 随机变量为**每个测试样本**的对错
- 随机变量的样本个数较少时(一般 ≤ 30): **t检验**
 - 一般用于多次评测如重复实验, 随机变量为**每次测试集**上的指标



小结

评价指标：回归任务， 分类任务， 特定任务

训练集、验证集与测试集：随机划分， 留一划分， 特殊划分

随机重复实验

K折交叉验证

统计有效性检验：z检验， t检验