第二章　模型评估与选择

0.1　评估方法

0.1.1　留出法 (hold-out)

将数据集 D 划分为两个互斥集合：训练集 S 和测试集 T，即：$D = S \cup T, S \cap T = \phi$。数据集划分时，采用分层采样 (stratified sampling)，单次留出法得到的结果往往不稳定可靠，所以一般若干次随机划分，重复进行实验评估后取平均值作为最终评估结果。

常用的划分比例是：将约 $\frac{2}{3} \sim \frac{4}{5}$ 的样本用于训练，剩余样本用于测试。

0.1.2　交叉验证法 (cross validation)

将数据集 D 划分为 k 个大小相似的互斥子集，即 $D = D_1 \cup D_2 \cup \cdots \cup D_k, D_i \cap D_j = \phi(i \neq j)$，每个子集通过分层采样得到尽可能数据分布一致的样本，每次使用 k-1 个子集的并集作为训练集，剩下的一个子集作为测试集，共得到 k 个测试结果，取平均值。通常把交叉验证法称为 k 折交叉验证 (k-fold cross validation)。k 最常用取值为 10。

特例：**留一法** (Leave-One-Out，简称 LOO)。数据集 D 有 m 个样本，且 k=m。留一法的评估结果通常较为准确，但未必永远更优且计算开销很大。

0.1.3　自助法 (bootstrapping)

对于含有 m 个样本的数据集 D，采样产生含有 m 个样本的数据集 D'：每次随机从 D 中挑选一个样本放入 D' 中，并将此样本放回 D 中，重复 m 次。某个样本不出现在 D' 中的概率是：$(1 - \frac{1}{m})^m$，取极限得到：

$$\lim_{m \to \infty} (1 - \frac{1}{m})^m \mapsto \frac{1}{e} \approx 0.368$$

即初始数据集 D 中约有 36.8% 的样本未出现在采样数据集 D' 中，于是用 D' 作训练集，$D \setminus D'$ 坐测试集。

自助法在数据集小，难以有效划分训练集和测试集时很有用；且自助法能从初始数据集中产生多个不同的训练集，对**集成学习**等有很大优势。但自助法改变了初始数据集的样本分布，会引入估计误差。

0.2　机器学习的参数

机器学习常涉及两类参数：算法参数和模型参数。前者也称为“超参数”，数目常在 10 以内，由人工设定；后者数目可能很多，通过学习产生。

0.3　性能度量

0.3.1　均方误差 (mean squared error)

给定样例集 $D = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$，其中 y_i 是示例 x_i 的真实标记，令 $f(x_i)$ 是 x_i 在学习器 f 上的预测结果。

回归任务：

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

更一般的，对于数据分布 D 和概率密度函数 $p(\cdot)$，均方误差可表示为：

$$E(f;D)=\int_{x\sim D}(f(x_i)-y_i)^2p(x)dx \tag{2}$$

0.3.2 查准率 (precision) 与查全率 (recall)

查准率亦称"准确率"，查全率亦称"召回率"。

TP, FP, TN, FN 分别表示真正例 (true positive)，假正例 (false positive)，真反例 (true negative)，假反例 (false negative)，有 $TP+FP+TN+FN=$ 样例总数。

查准率：$P=\frac{TP}{TP+FP}$，查全率：$R=\frac{TP}{TP+FN}$。查准率和查全率是一对矛盾的度量。一般说，当查准率高时，查全率往往偏低；当查全率高时，查准率往往偏低。只有在一些简单的任务中，才可能使查准率和查全率都很高。

"平衡点"(Break-Even Point，简称 BEP) 是当"查准率 = 查全率"时的取值，即 $BEP=P=R$。

0.3.3 F_1 度量

F_1 是基于查准率和查全率的调和平均 (harmonic mean) 定义的：

$$\frac{1}{F_1}=\frac{1}{P}+\frac{1}{R}\Longrightarrow F_1=\frac{2\times P\times R}{P+R}=\frac{2\times TP}{\text{样例总数}-TP-TN} \tag{3}$$

F_β 则是加权调和平均：

$$\frac{1}{F_\beta}=\frac{1}{1+\beta^2}(\frac{1}{P}+\frac{\beta^2}{R})\Longrightarrow F_\beta=\frac{(1+\beta^2)\times P\times R}{(\beta^2\times P)+R} \tag{4}$$

其中 $\beta>0$ 度量了查全率对查准率的相对重要性。当 $\beta=1$ 时，退化为标准的 F_1；当 $\beta>1$ 时，查全率有更大影响；当 $\beta<1$ 时，查准率有更大影响。

0.3.4 ROC 与 AUC

ROC 全称为：受试者工作特征 (Receiver Operating Characteristic) 曲线。根据学习器的预测结果对样例进行排序，按此顺序**逐个把样本作为正例**进行预测，每次计算 FPR 和 TPR，分别以它们作为横/纵坐标作图。横轴 $FPR=\frac{FP}{TN+FP}$ 是"假正例率"(False Positive Rate)，纵轴 $TPR=\frac{TP}{TP+FN}$ 是"真正例率"(True Positive Rate).

绘制 ROC 图的过程：给定 m^+ 个正例和 m^- 个反例，根据学习器预测结果对样例进行排序，然后把分类阈值设为最大，即把所有样例均预测为反例，此时 TPR 和 FPR 均为 0，在坐标 (0,0) 处标记一个点。然后**将分类阈值依次设为每个样例的预测值，即依次将每个样例划分为正例**。设前一个标记点坐标为 (x,y)，当前若为真正例，则对应标记点的坐标为 $(x,y+\frac{1}{m^+})$；若当前为假正例，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即可。

通过比较 ROC 曲线下的面积，即 AUC(Area Under ROC Curve)，来比较学习器的优劣。AUC 可估算为；

$$AUC=\frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1}-x_i)\cdot(y_i+y_{i+1}) \tag{5}$$

AUC 考虑的是样本预测的排序质量。给定 m^+ 个正例和 m^- 个反例，令 D^+ 和 D^- 分别表示正/反例集合，则排序损失 (rank loss) 定义为；

$$\ell_{rank}=\frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(\mathbb{I}(f(x^+)<f(x^-))+\frac{1}{2}\mathbb{I}(f(x^+)=f(x^-))\right) \tag{6}$$

即考虑每一对正反例。若正例的预测值小于反例，则代价加一，若相等则代价加 0.5。ℓ_{rank} 表示的是 ROC 曲线上的面积：若一个正例在 ROC 曲线上对应标记点的坐标为 (x, y)，则 x 恰是排序在其之前的反例所占的比例，即假正例率。因此有：

$$AUC=1-\ell_{rank} \tag{7}$$

0.4 比较检验

0.4.1 假设检验

假设检验的逻辑是：**全称命题只能被否证而不能被证明**。因为个案不足以证明全称命题，但是可以否定全称命题。因此通过否证假设的反面（虚无假设），或者"拒绝"假设的反面，来证明假设。由于抽样的原因，样本并不可能绝对地否证虚无假设。在个案中，小概率事件可以等同于不可能发生的事件。我们在这个意义上去在一定的事先约定的概率水平上去拒绝虚无假设。

在包含 m 个样本的测试集上，泛化错误率为 ϵ 的学习器被测得测试错误率为 $\hat{\epsilon}$ 的概率：

$$P(\hat{\epsilon};\epsilon) = \binom{m}{\hat{\epsilon}\times m}\epsilon^{\hat{\epsilon}\times m}(1-\epsilon)^{m-\hat{\epsilon}\times m} \tag{8}$$

求导可得：

$$\frac{\partial P(\hat{\epsilon};\epsilon)}{\partial \epsilon} = \binom{m}{\hat{\epsilon}\times m}\epsilon^{\hat{\epsilon}\times m-1}(1-\epsilon)^{m-\hat{\epsilon}\times m-1}(\hat{\epsilon}-\epsilon)\times m \tag{9}$$

所以当 $\hat{\epsilon}=\epsilon$ 时，$P(\hat{\epsilon};\epsilon)$ 最大，当 $|\epsilon-\hat{\epsilon}|$ 增大时，$P(\hat{\epsilon};\epsilon)$ 减小。

二项检验 (binomial test)：考虑假设"$\epsilon \leq \epsilon_0$"，则在 $1-\alpha$（此为置信度,confidence）的概率内所能观察到的最大错误率为：

$$\bar{\epsilon} = \max \epsilon \quad \text{s.t.} \quad \sum_{i=\epsilon_0\times m+1}^{m}\binom{m}{i}\epsilon^{i}(1-\epsilon)^{m-i} < \alpha \tag{10}$$

当接受假设"$\epsilon \leq \epsilon_0$"时，需满足测试错误率 $\hat{\epsilon}$ 小于临界值 $\bar{\epsilon}$，即能以 $1-\alpha$ 的置信度认为学习器的泛化错误率不大于 ϵ_0。

t 检验 (t-test)：假设得到了 k 个测试错误率 $\hat{\epsilon_1},\hat{\epsilon_2},\cdots,\hat{\epsilon_k}$，则平均测试误差 μ 和方差 σ^2 为：

$$\mu = \frac{1}{k}\sum_{i=1}^{k}\hat{\epsilon_i} \tag{11}$$

$$\sigma^2 = \frac{1}{k-1}\sum_{i=1}^{k}(\hat{\epsilon_i}-\mu)^2 \tag{12}$$

由于这 k 个测试错误率可看作泛化错误率 ϵ_0 的独立采样，则变量：

$$\tau_t = \frac{\sqrt{k}(\mu-\epsilon_0)}{\sigma} \tag{13}$$

服从自由度为 k-1 的 t 分布。**需要注意这里的 μ,σ^2,ϵ_0 分别为样本均值，样本方差和总体均值。**

0.4.2 交叉验证 t 检验

对于学习器 A 和学习器 B，使用 k 折交叉验证法得到的测试错误率分别为 $\epsilon_1^A,\epsilon_2^B,\cdots,\epsilon_k^A$ 和 $\epsilon_1^B,\epsilon_2^B,\cdots,\epsilon_k^B$，令 $\delta_i = \epsilon_i^A - \epsilon_i^B$，若两个学习器的性能相同，则差值均值应为 0。对假设"学习器 A 与学习器 B 性能相同"做 t 检验，计算差值的均值 μ 和方差 σ^2，在显著度 α 下，若：

$$\tau_t = |\frac{\sqrt{k}\mu}{\sigma}|$$

小于临界值 $t_{\alpha/2,k-1}$，则接受假设。其中 $t_{\alpha/2,k-1}$ 是自由度为 k-1 的 t 分布上尾部累积分布为 $\alpha/2$ 的临界值。

0.4.3 McdNemar 检验

对于学习器 A 和学习器 B 分别用 $e_{00},e_{01},e_{10},e_{11}$ 表示 A 和 B 的分类结果都正确，B 正确 A 错误，A 正确 B 错误，AB 均错误。如果两学习器性能相同，则 $e_{01}=e_{10}$，那么变量 $|e_{01}-e_{10}$ 应当服从正态分布. McNemar 检验考虑变量

$$\tau_{\chi^2} = \frac{(|e_{01}-e_{10}|-1)^2}{e_{01}+e_{10}}$$

服从自由度为 1 的 χ^2 分布，即标准正态分布变量的平方。分子中有-1，是因为 $e_{01}+e_{10}$ 通常很小，需要考虑连续性校正。给定显著度 α，当以上变量值小于临界值 χ^2_α 时，不能拒绝假设，即应认为两学习器的性能没有显著差别。

补充：**p 值 (p-value) 指的是：在原假设 H_0 成立时，出现现状或更差的情况的概率**。例如抛 20 次硬币，出现 18 次正面，则出现现状或更差的情况是：出现 18，19，20 次正面和出现 0，1，2 次正面，所以 p 值为：$p=(\frac{1}{2})^{20}\left(\binom{20}{18}+\binom{20}{19}+\binom{20}{20}+\binom{20}{0}+\binom{20}{1}+\binom{20}{2}\right)$。

0.4.4 Friedman 检验与 Nemenyi 后续检验

0.5 偏差与方差

泛化误差（在训练集上得到的模型在新样本上的输出误差）可分解为偏差（bias，期望输出与真实标记的差别），方差（variance，预测输出与期望输出的差别）与噪声（noise，数据集中标记和真实标记的差别）之和。