模型评估与选择

经验误差与过拟合

- 过拟合无法避免

评估方法

测试集上的测试误差作为泛化误差的近似

留出法hold-out

- 两个互斥集合
- 两个集合数据分布尽可能一致【分层抽样】

  避免数据划分引入额外的偏差
- 一般采用若干次留出法，去平均作为评估结果
- 2/3~4/5样本用于训练，剩余样本用于测试

  训练集占比高，测试集偏差大，不准确；
  训练集占比低，测试集方差大，保真性低

交叉验证法k-fold cross validation

- k个互斥子集
- 随机使用p次划分，常见10次10折交叉验证
- 留一法

  每个样本为一个子集，不受随机样本划分影响，训练集与整体数据接近，结果比较准确，但计算成本高
  根据"没有免费的午餐"定理，估计结果不一定永远比其他评估方式好

自助法bootstrapping

- 有抽样放回
- 外包估计：36.8%样本不会出现在采样数据集中
- 使用场景
  - 👍适用于数据集较小、难以有效划分训练/测试集时
  - 👍集成学习
- 😐改变了数据集分布，引入估计偏差

调参与最终模型

- 对每个参数选定一个范围和变化步长
- 模型选择完成后，应使用整个数据集重新训练模型

- **性能度量**

  模型的好坏是相对的，什么样的模型是好的，不仅取决于算法和数据，还决定于任务需求

  - 错误率与精度
    - 错误率：分类错误样本数/样本总数
    - 精度：分类正确样本数/样本总数
  - 查准率、查全率与F1
    - 查准率precision=TP/(TP+FP)
      - 预测正例的样本中，真正为正例的比例
    - 查全率recall=TP/(TP+FN)
      - 真正为正例中，正确预测为正例的比例
    - 查准率高，查全率往往偏低；查全率高，查准率往往偏低

      若希望好瓜尽可能多选出来（查全率），可以增加选瓜数量；但是坏瓜也会被选多了，导致查准率低
    - P-R曲线
      -

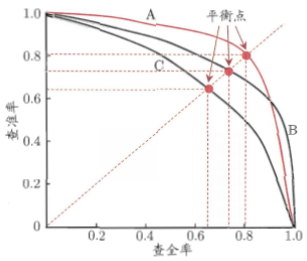
      

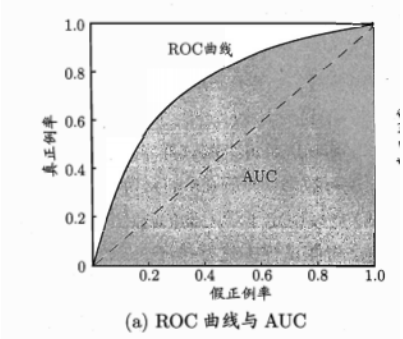
      图 2.3 P-R曲线与平衡点示意图
      - 比较方法
        - P-R曲线下面积大小
        - 平衡点Break-Even Point：查准率=查全率
        - F1度量
          -

          $$F1 = \frac{2 \times P \times R}{P + R}$$
          -

          $$\frac{1}{F1} = \frac{1}{2}(\frac{1}{P} + \frac{1}{R})$$

- 为什么用调和平均：分子相同，分母不同，把分母调成平均数再当分母。
- 表达对查准率/查全率不同的偏好
  - $$F_\beta = \frac{(1+\beta^2) \times P \times R}{(\beta^2 \times P) + R}$$
  - $$\frac{1}{F1} = \frac{1}{1+\beta^2}(\frac{1}{P} + \frac{\beta^2}{R})$$
  - β>1时查全率更重要

- ROC与AUC
  - 排序本身的质量好坏，体现了综合考虑学习器在不同任务下的“期望泛化性能”的好坏
  - ROC图像
    - 横坐标：假正例率FPR：FP/（TN+FP），反例当中有多少被预测为正例
    - 纵坐标：真正例率TPR：TP/（TP+FN），正例当中有多少被预测为正例
    - 左上角为理想模型



(a) ROC 曲线与 AUC

  - AUC：ROC曲线下的面积
- 代价敏感错误率与代价曲线
  - 为权衡不同类型错误所造成的不同损失，可为错误赋予“非均等代价”
  - 代价敏感错误率
    - $$E(f;D;cost) = \frac{1}{m}\left(\sum_{\boldsymbol{x}_i \in D^+} \mathbb{I}(f(\boldsymbol{x}_i) \neq y_i) \times cost_{01} + \sum_{\boldsymbol{x}_i \in D^-} \mathbb{I}(f(\boldsymbol{x}_i) \neq y_i) \times cost_{10}\right)$$

- 代价曲线
  - 横轴：正例概率代价

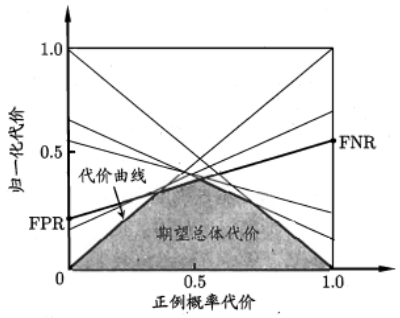
    p为样例为正例的概率
    - $$P(+)cost = \frac{p \times cost_{01}}{p \times cost_{01} + (1-p) \times cost_{10}}$$
  - 纵轴：归一化代价
    - $$cost_{norm} = \frac{\mathrm{FNR} \times p \times cost_{01} + \mathrm{FPR} \times (1-p) \times cost_{10}}{p \times cost_{01} + (1-p) \times cost_{10}}$$
  - ROC曲线上每一点对应了代价平面的一条线段，取所有线段下界，围成的面积即为期望总体代价
    - 

      图 2.5 代价曲线与期望总体代价

比较检验

- 假设检验
  - 在测试集观察到学习器A比B好，则A的泛化性能是否在统计上优于B
  - 假设学习器泛化错误率为某个值，用测试错误率进行假设检验
  - t检验
- 交叉验证t检验
  - 成对t检验，假设A错误率=B错误率
  - 交叉检验不满足错误率独立采样的前提，使用5×2交叉验证
- McNemar检验
  - 二分类问题，两个算法进行比较
  - 卡方分布

-

表 2.4 两学习器分类差别列联表

算法 B	算法 A	
	正确	错误
正确	e_{00}	e_{01}
错误	e_{10}	e_{11}

若我们做的假设是两学习器性能相同, 则应有 $e_{01} = e_{10}$, 那么变量 $|e_{01} - e_{10}|$ 应当服从正态分布, 且均值为 1, 方差为 $e_{01} + e_{10}$. 因此变量

$$\tau_{\chi^2} = \frac{(|e_{01} - e_{10}| - 1)^2}{e_{01} + e_{10}} \tag{2.33}$$

- Friedman检验与Nemenyi后续检验
  - 多个算法进行比较

偏差与方差

- 偏差：拟合能力
- 方差：数据扰乱
- 噪声：问题难度