《机器学习》读书笔记 第2章 模型评估与选择

```
《机器学习》读书笔记 第2章 模型评估与选择
     错误率 (error rate)
     精度 (accuracy)
     误差 (error)
     泛化误差(generalization error)
     过拟合(overfitting)
     欠拟合 (underfitting)
     测试误差(testing error)
  数据集的划分
     划分类型
        训练集(training set)
        测试集 (testing set)
            测试误差(testing error)
            训练误差(training error)
        验证集 (verify set)
     划分方法
         留出法(hold-out)
        交叉验证法(cross validation)
         自助法 (bootstrapping)
```

性能度量

```
均方误差(mean square error)错误率(error rate)精度(accuracy)查准率、查全率和 F<sub>1</sub>
查准率(precision)查全率(recall)
平衡点(Balance-Event Point)
F1
Fβ
真正率(True Positive Rate)
假正例率(False Positive Rate)
ROC 和 AUC
代价敏感错误率与代价曲线
代价敏感错误率
```

错误率(error rate)

$$E = \frac{a}{m}$$

意为: 在m个样本中,有a个样本分类错误

精度 (accuracy)

$$1 - E = 1 - \frac{a}{m}$$

误差(error)

学习器的实际预测输出 与 样本的真实输出之间的差异

泛化误差 (generalization error)

学习器在新样本上产生的误差

过拟合 (overfitting)

无法彻底避免,只能缓解或减小其风险

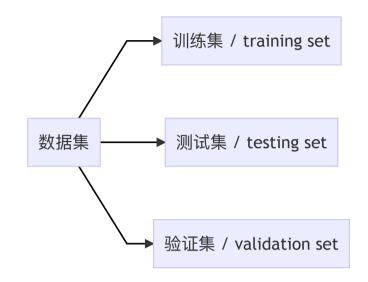
欠拟合 (underfitting)

与过拟合相对

测试误差(testing error)

数据集的划分

划分类型



训练集(training set)

测试集 (testing set)

用于测试学习器对新样本的判别能力

测试误差 (testing error)

在测试集上得出的误差,可作为泛化误差的近似

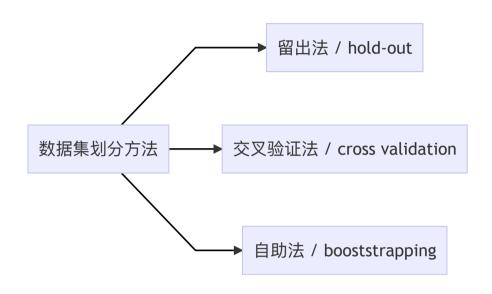
训练误差 (training error)

亦称: 经验误差 (empirical error)

学习器在训练集上产生的误差

验证集 (verify set)

划分方法



留出法(hold-out)

该方法直接将数据集 D 划分为两个互斥的集合,其中一个 集合作为训练集 S,另一个作为测试集 T, 即 $D=B\cup T,S\cap T=\varnothing$.在 S 上训 练出模型后,用 T来评估其测试误差。

交叉验证法 (cross validation)

又称 k 折交叉验证(k-fold cross validation)

先将数据集 D 划分为 k 个大小相似的互斥子集, 即

$$D = D_1 \cup D_2 \cup \ldots \cup D_k$$

其中

$$D_i \cap D_i = \varnothing, \quad (i \neq j)$$

每个子集 D_i 都 尽可 能保持数据分布的一致性,即从 D中通过分层采样得到。然后,每次用 k-1 个子集的并集作为训练集,余下的那个子集作为测试集。

自助法 (bootstrapping)

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' 。 D' 用过训练集,D/D' 用作测试集

性能度量

均方误差 (mean square error)

对于样例集D,有

$$E(f;D) = rac{1}{m} \sum_{i=1}^m \left(f\left(oldsymbol{x}_i
ight) - y_i
ight)^2$$

更一般的可描述为

$$E(f;\mathcal{D}) = \int_{oldsymbol{x}\sim\mathcal{D}} (f(oldsymbol{x}) - y)^2 p(oldsymbol{x}) \mathrm{d}oldsymbol{x}$$

错误率 (error rate)

$$E(f;D) = rac{1}{m} \sum_{i=1}^{m} \mathbb{I}\left(f\left(oldsymbol{x}_i
ight)
eq y_i
ight)$$

更一般的有

$$E(f;\mathcal{D}) = \int_{oldsymbol{x}\sim\mathcal{D}} \mathbb{I}(f(oldsymbol{x})
eq y) p(oldsymbol{x}) \mathrm{d}oldsymbol{x}$$

精度(accuracy)

$$egin{align} \operatorname{acc}(f;D) &= rac{1}{m} \sum_{i=1}^m \mathbb{I}\left(f\left(oldsymbol{x}_i
ight) = y_i
ight) \ &= 1 - E(f;D) \end{split}$$

更一般的有

$$egin{align} &\operatorname{acc}(f;\mathcal{D}) = \int_{m{x}\sim\mathcal{D}} \mathbb{I}(f(m{x}) = y) p(m{x}) \mathrm{d}m{x} \ &= 1 - E(f;\mathcal{D}) \end{split}$$

查准率、查全率和 F_1

	预测结果	预测结果	
真实情况	正例	反例	
正例	TP(真正例 $)$	FN(假反例 $)$	
反例	<i>FP</i> (假正例)	TN(真反例 $)$	

TP、FN、FP 和 TN 分别代表对应样例数,由此可知:TP + FN + FP + TN 等于全部样例数

查准率和查全率是一对矛盾的度量。一般来说,查准率 高时,查全率往往偏低;而查全率高时,查准率往往偏 低。

查准率 (precision)

$$P = \frac{TP}{TP + FP}$$

$$macro -P = \frac{1}{n} \sum_{i=1}^{n} P_i$$

$$\operatorname{micro} -P = \frac{\overline{TP}}{\overline{TP} + \overline{FP}}$$

查全率(recall)

$$R=rac{TP}{TP+FN}$$
 $ext{macro-}\,R=rac{1}{n}\sum_{i=1}^{n}R_{i}$
 $ext{micro-}\,R=rac{\overline{TP}}{\overline{TP}+\overline{FN}}$

平衡点 (Balance-Event Point)

其值为当 查全率 = 查准率 时 的取值,其值越大,说明学习器的性能越好。

F1

$$F1 = rac{2 imes P imes R}{P+R} = rac{2 imes TP}{$$
 样例总数 $+TP-TN$ macro- $F1 = rac{2 imes ext{macro} - P imes ext{macro} - R}{ ext{macro} - P + ext{macro} - R}$ micro- $F1 = rac{2 imes ext{micro} - P imes ext{micro} - R}{ ext{micro} - P + ext{micro} - R}$

 $F\beta$

$$F_{eta} = rac{\left(1 + eta^2
ight) imes P imes R}{\left(eta^2 imes P
ight) + R}$$

其中 $\beta > 0$ 度量了查全率对查准率的相对重要性。

eta=1 时退化为标准的 F_1 ; eta>1 时查全率有更大影响; eta<1 时查准率有更大影响。

真正率 (True Positive Rate)

$$ext{TPR} = rac{TP}{TP + FN}$$

假正例率(False Positive Rate)

$$ext{FPR} = rac{FP}{TN + FP}$$

ROC 和 AUC

ROC 全称是"受试者工作特征" (Receiver Operating Characteristic) 曲线

AUC 的估计如下

$$ext{AUC} = rac{1}{2} \sum_{i=1}^{m-1} \left(x_{i+1} - x_i
ight) \cdot \left(y_i + y_{i+1}
ight)$$

$$\ell_{ ext{rank}} \ = rac{1}{m^+m^-} \sum_{oldsymbol{x}^+ \in D^+} \sum_{oldsymbol{x}^- \in D^-} \left(\mathbb{I}\left(f\left(oldsymbol{x}^+
ight) < f\left(oldsymbol{x}^-
ight)
ight) + rac{1}{2} \mathbb{I}\left(f\left(oldsymbol{x}^+
ight) = f\left(oldsymbol{x}^-
ight)
ight)
ight)$$

$$AUC + \ell_{rank} = 1$$

代价敏感错误率与代价曲线

	预测类别	预测类别	
真实类别	第0类	第1类	
第0类	0	$cost_{01}$	
第1类	$cost_{10}$	0	

代价敏感错误率

$$E(f;D; ext{cost}) = rac{1}{m} \left(\sum_{oldsymbol{x}_i \in D^+} \mathbb{I}\left(f\left(oldsymbol{x}_i
ight)
eq v_i
ight) imes cost_{01} + \sum_{oldsymbol{x}_i \in D^-} \mathbb{I}\left(f\left(oldsymbol{x}_i
ight)
eq v_i
ight) imes cost_{10}
ight)$$

正例概率代价

$$P(+)\cos t = rac{p imes cost_{01}}{p imes cost_{01} + (1-p) imes cost_{10}}$$

其中 p 为样例为正的概率