第五章 分类

第五章 分类

5.1 模型评估和性能度量
5.2 决策树
5.3 贝叶斯分类
5.4 k最近邻分类
5.5 组合分类

分类问题

分类：把未标签样本分配到预先定义好的类别； 未标签样本又叫新样本或测试样本。

模型训练：基于标签样本（训练样本）建立用于预测新样本类别的模型的过程。

训练集 ： 带标签的数据集

id	属性1	属性2	标签
1	2.3	4.5	1
2	1.2	5.6	1
3	2.3	4.5	1
4	-1.2	-3.4	2
5	-2.3	-5.6	2
6	-4.5	-2.3	3
7	-3.4	-4.2	3

测试集： 需要预测标签的数据集

id	属性1	属性2	标签
1	3.4	2.5	?
2	1.2	4.5	?
3	-2.3	-3.5	?

模型是什么？它可以是一系列规则、一个决策树、一个带参数的线性/非线性函数。

主要分类算法：决策树、贝叶斯分类、 k最近邻、 SVM, 逻辑回归、神经网络,…

分类的基本概念和术语

- 数学描述

给定训练集$\chi_{train} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$，其中$(\boldsymbol{x}_i, y_i)$叫做元组(tuple)，代表一个样本向量$\boldsymbol{x}_i$及对应类标签$y_i$。模型训练或模型学习的过程是建立$\mathcal{X} \rightarrow \mathcal{Y}$的映射$f$。一旦建立，可以用$y = f(\boldsymbol{x})$来预测新样本的标签。这个映射叫做分类的预测模型或分类器。

- 二分类 vs 多分类

若$|\mathcal{Y}| = 2$，如$\mathcal{Y}$={−1, 1} 或者 {1, 0}，则为二分类问题。若$|\mathcal{Y}| > 2$，则为多分类。

- 监督学习

训练集中样本的类标签是一种最常用的监督信息。监督信息的数量和质量对模型学习起到关键作用。

什么样的模型是好的？

用于模型选择、参数设置。

分两个方面：

怎么评估一个模型的好坏？--模型评估方法

怎么衡量有多好？--性能度量

第五章 分类

问题一：模型评估

分类任务的目标：训练好的模型对未知样本的分类尽可能准确 –泛化能力

泛化误差：在“未来”样本上的误差 ➡ 理想的评估方法

然而，在训练模型的时候并不知道未知样本，不能直接评估泛化误差

要找一个近似的方法。。。

训练误差：在训练集上的误差，亦称“经验误差”合适吗？

- 泛化误差越小越好
- 经验误差是否越小越好？

NO! 因为会出现“**过拟合**”(overfitting)

过拟合 (overfitting) vs. 欠拟合 (underfitting)

过拟合：关注太多细节，要求太苛刻—太敏感。

欠拟合：对重要信息描述不到位—太迟钝。

泛化误差的近似：测试误差

老师给小明做了10道线性代数的题目，他想知道小明是否掌握了相关知识点？
做法1：老师遮住拿刚才的10道题目的答案，让小明再回答一遍，他都答对了！训练误差为0。

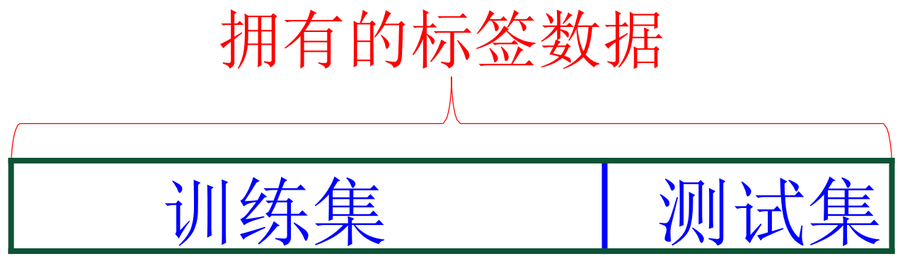
小明真的掌握这些知识点了吗？
其实，小明的记忆力超强，把这10题的答案都记住了。

做法2：拿小明没有做过的题目考他。

如何近似地评估模型的泛化能力？

把带标签数据分成两部分，一部分用于训练，一部分用于测试。由于测试数据在训练过程中不出现，因此测试样本类似“未知样本”，测试误差近似泛化误差。

训练集(training set)和测试集(test set)的划分



我们需要考虑的问题：

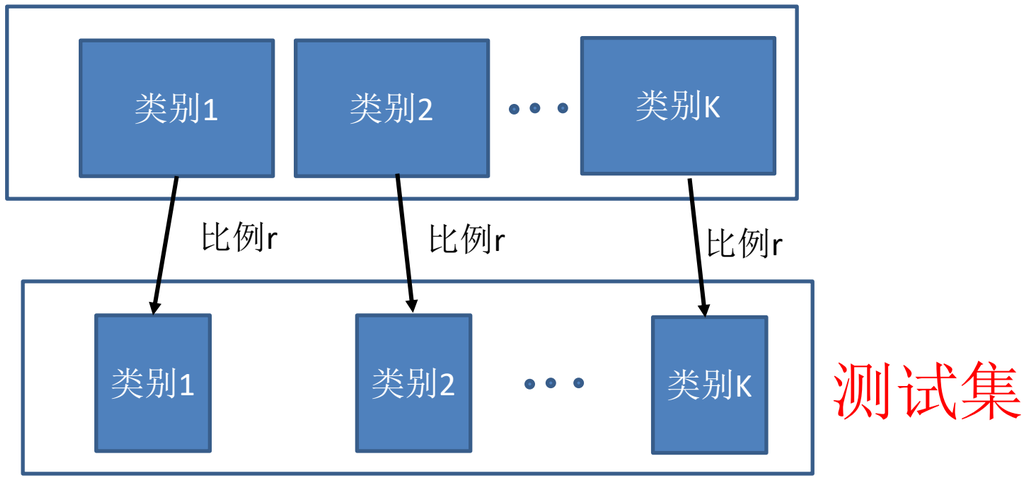
- 为了减少过拟合风险，测试集应该与训练集“互斥”
- 训练集和测试集都尽可能大—矛盾

常见方法：

- 留出法 (hold-out)
- 交叉验证法 (cross validation)
- 自助法 (bootstrap)

留出法（holdout）

拥有的数据集



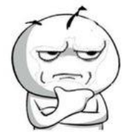
测试集

注意：

- 保持数据分布一致性 （例如：分层采样，即从每个类里面随机抽样同比例的样本)
- 多次重复划分 (例如：100次随机划分)
- 测试集不能太大、不能太小，为什么？一般在 1/5~1/3

k-折交叉验证法(k-fold cross validation)

若 k = n，则得到“留一法”
(leave-one-out, LOO)

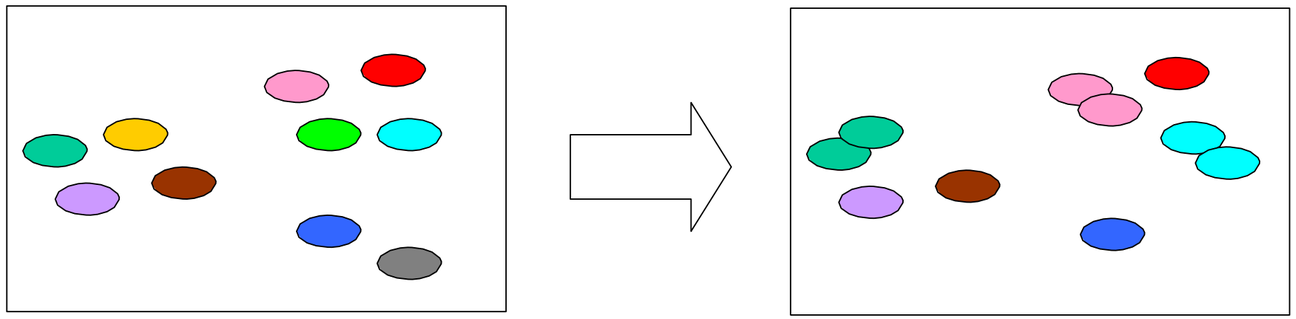


思考：留一法的优缺点是什么？

自助法（Bootstrapping）

基于“自助采样”：从n个样本中每次随机抽一个，然后放回，重复n次。

亦称“有放回采样”、“可重复采样”



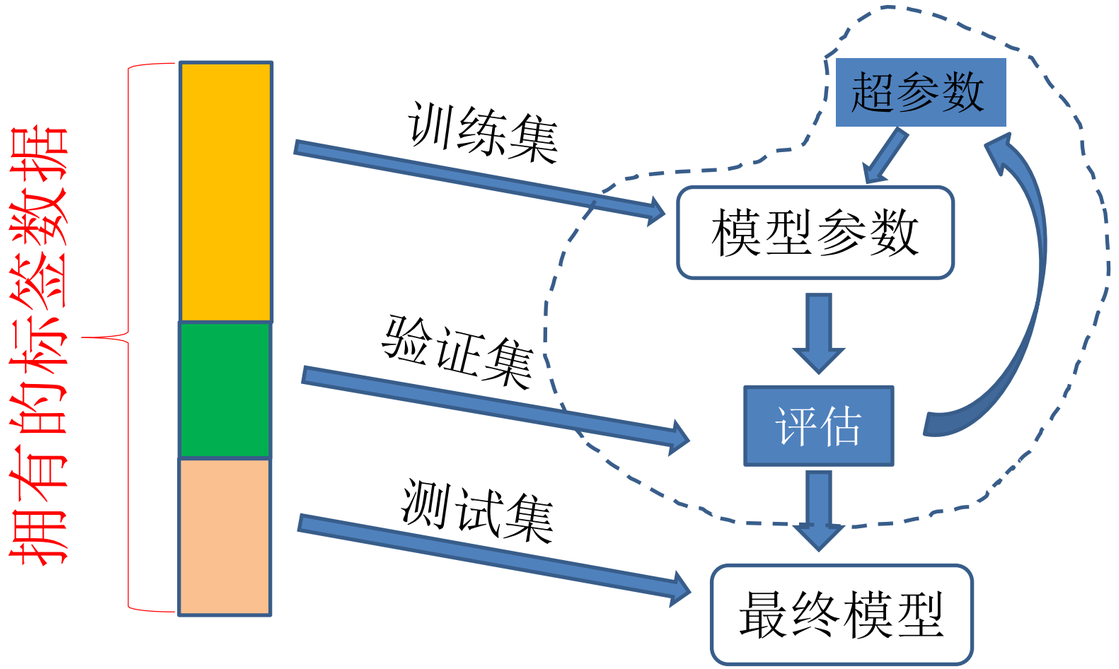
$$\lim_{n\to\infty}\left(1-\frac{1}{n}\right)^{n}\to\frac{1}{e}$$

$$\approx 0.368$$

“包外估计”(out-of-bag estimation)

优点：训练集与原样本集同规模
缺点：数据分布有改变

“调参”与最终模型



模型训练，即学习模型参数之前，除了给定训练集，一般还需要人工设定算法特有的参数，称“超参数”。

超参数怎么选择呢？

区别：训练集 vs. 测试集 vs. 验证集 (validation set)

算法参数选定后，要用“训练集+验证集”重新训练最终模型

问题二：性能度量

性能度量是衡量模型泛化能力的评价标准

什么样的模型是“好”的，不仅取决于算法和数据， 还取决于任务需求

- 分类任务用错误率或准确率

$$\mathrm{E}(f,D)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(f(\boldsymbol{x}_i)\neq y_i)$$

$$\begin{aligned}\mathrm{acc}(f,D)&=\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(f(\boldsymbol{x}_i)=y_i)\\&=1-\mathrm{E}(f,D)\end{aligned}$$

- 回归任务常用均方误差：

$$\mathrm{E}(f,D)=\frac{1}{n}\sum_{i=1}^{n}(f(\boldsymbol{x}_i)-y_i)^2$$

查准率 vs. 查全率

混淆矩阵（二分类）

真实情况	预测结果	
	正例	反例
正例	TP (真正例)	FN (假反例)
反例	FP (假正例)	TN (真反例)

例：已知一个包含10个样本的测试集，前4个为正例，后6个为反例。现有一个算法预测出来的结果是：1 0 1 0 1 0 0 0 1 0，其中1代表正例，0代表反例。 计算查准率和查全率。

- 查准率(precision)： $P = \frac{TP}{TP + FP}$
- 查全率/ 召回率 （recall）： $R = \frac{TP}{TP + FN}$

F1度量：查准率和查全率的调和平均

$$F1 = \frac{2 \times P \times R}{P + R} = \frac{2 \times TP}{\text{样例总数} + TP - TN}$$

- 哪些应用更看中查准率？ 哪些应用更看中查全率？
- 当分类器返回的不是直接对应类别的离散值，而是分布概率时，怎么办？

若对查准率/查全率有不同偏好：

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

$\beta > 1$ 时查全率有更大影响; $\beta < 1$ 时查准率有更大影响

宏xx vs. 微xx

若能得到多个混淆矩阵：例如多分类的两两混淆矩阵

宏(macro-)查准率、查全率、F1

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

$$\text{macro-}R = \frac{1}{n}\sum_{i=1}^{n} R_i \ ,$$

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

微(micro-)查准率、查全率、F1

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

$$\text{micro-}R = \frac{\overline{TP}}{\overline{TP} + \overline{FN}} \ ,$$

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

非均等代价

犯不同的错误往往会造成不同的损失

此时需考虑“非均等代价”(unequal cost)

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

- 代价敏感(cost-sensitive)错误率：

$$\mathrm{E}(f, D, cost) = \frac{1}{n}\left(\sum_{\boldsymbol{x}_i \in D^+} \mathbb{I}(f(x_i) \neq y_i) \times cost_{01} + \sum_{\boldsymbol{x}_i \in D^-} \mathbb{I}(f(x_i) \neq y_i) \times cost_{10}\right)$$

显著性检验

有了“模型评估”和“性能度量”就能知道哪个模型好了吗？

若以下为两个模型5折交叉验证的错误率：
模型A：0.20， 0.33， 0.31， 0.18，0.15
模型B：0.19， 0.22， 0.28，0.20，0.21

➡ 哪个模型表现好？还是差不多一样好？

成对t检验(paired t-test)来比较检验两个学习器的性能

假设H（hypothesis）： 学习器A和 B性能相同

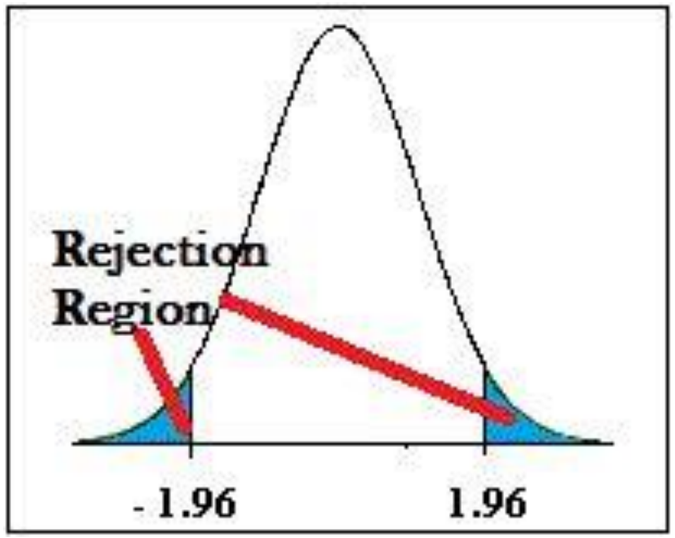
1. 计算每次错误率差值： Δ_1，Δ_2，$\cdots\Delta_k$,其中$\Delta_i=\epsilon_i^A-\epsilon_i^B$ 。

2. 计算k个差值的均值$\mu=\frac{1}{k}\sum_{i=1}^{k}\Delta_i$和方差$\sigma^2=\frac{1}{k-1}\sum_{i=1}^{k}(\Delta_i-\mu)^2$, 以及 $\tau_t=\left|\frac{\sqrt{k}\mu}{\sigma}\right|$ 。

3. 若 $\tau_t<t_{\frac{\alpha}{2},k-1}$ (自由度为k-1 的t分布上尾部累积分布为的$\alpha/2$ 的临界值)

则假设H不能被拒绝，即两个学习器性能相同，否则平均错误率较小的那个分类器性能较好（一般取α=0.05，对应95%置信度）。

t分布与假设检验



- t分布类似正态分布，不过用在样本量小的情况下，如小于30。
- 样本很大时，趋向于正态分布。

单侧	75%	80%	85%	90%	95%	97.5%	99%
双侧	50%	60%	70%	80%	90%	95%	98%
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764

假设有10组评估结果，则自由度为10-1=9， 确定置信度95%（双侧）或97.5%（单侧），查表得到对应t临界值为2.262。

该图上的中心区域是接受区域，尾部（两边蓝色部分）是拒绝区域或区域