# 西瓜书概念整理（chapter 1-2）熟悉机器学习术语 - 梦里风林

括号表示概念出现的其他页码，如有兴趣协同整理，请到issue中认领章节

完整版见我的github：ahangchen

觉得还不错的话可以点个star ^_^

- Page2：假设(269)(hypothesis)

  学得模型对应了数据的某种潜在的规律，因此亦称假设
- Page2：示例(instance)

  数据集中的每条记录是关于某个事件或对象的描述，称为一个“示例”或“样本”
- Page2：属性(attribute)

  反映事务或对象在某方面的表现或性质的事项，如“色泽”，称为属性或特征
- Page2：属性空间(attribute space)

  属性长成的空间称为属性空间，样本空间，或输入空间
- Page2：数据集(data set)

  数据记录的集合称为一个数据集
- Page2：特征(247)(feature)

  同属性
- Page2：学习器(learner)

  学习过程就是为了找出或逼近真相，有时将模型称作学习器
- Page2：训练(training)

  同学习
- Page2：训练集(training data)

  训练过程中使用的数据称为“训练集”，其中每个样本称为一个“训练样本”，训练样本组成的集合称为训练集
- Page2：训练样本(training sample)

  见训练集
- Page2：样本空间(sample space)

  同属性空间
- Page3：标记空间(label space)

所有标记的集合称为标记空间或输出空间

- Page3: 测试(testing)

学得模型后，使用其进行预测的过程称为测试，被预测的样本称为测试样本

- Page3: 测试样本(testing sample)

见测试

- Page3: 独立同分布(267)（independent and identically distributed）

我们获得的每个样本都是独立的从一个分布上采样获得的，即“独立同分布”

- Page3: 多分类（multi-class classification）

预测值涉及多个类别时，称为“多分类”

- Page3: 二分类（binary classification）

预测值设计两个分类的任务

- Page3: 泛化（121，350）（generalization）

学得模型适用于新样本的能力，称为“泛化”能力

- Page3: 分类（classification）

如果预测的是离散值，此类学习任务称为分类

- Page3: 回归（regression）

如果预测的值是连续值，此类学习任务称为回归

- Page3: 监督学习（supervised learning）

根据训练数据是否拥有标记信息，学习任务可以大致分为两大类：监督学习和无监督学习，分类和回归是前者的代表，聚类是后者的代表

- Page3: 聚类(197)（clustering）

见簇

- Page3: 无监督学习(197)（unsupervised learning）

见有监督学习

- Page4: 概念学习(17)（concept learning）

广义的归纳学习大体相当于从样例中学习，而狭义的归纳学习则要求从训练数据中学得概念，因此亦称为概念学习或概念形成

- Page4: 归纳学习(11)（inductive learning）

从样例中学习

- Page5: 版本空间（version space）

存在着一个与训练集一致的假设集合，称之为“版本空间”

- Page6: 归纳偏好（inductive bias）

机器学习算法在学习过程中对某种类型假设的偏好，称为归纳偏好

- Page6: 偏好

同归纳偏好

- Page7: 奥卡姆剃刀(17)（Occam's razor）

若有多个假设与观察一致，则选最简单的那个

- Page10: 符号主义(363)（symbolism）

基于逻辑表示

- Page10: 连接主义（connectionism）

基于神经网络

- Page11: 机械学习

信息存储与检索

- Page11: 示教学习

从指令中学习

- Page12: 统计学习(139)

如SVM，核方法

- Page16: WEKA

# 第二章 模型评估与选择

- Page23: 错误率(error rate)

分类错误的样本数占样本总数的比例称为错误率，即如果在m个样本中有a个样本分类错误，则错误率E = a/m；相应的，1-a/m称为精度。

- Page23: 泛化误差（generalization error）

在新样本上的误差称为泛化误差

- Page23: 过拟合(104,191,352)（overfitting）

当学习器把训练样本学得太好了的时候，很可能已经把训练样本自身的一些特点当作了所有潜在样本都会具有的一般性质，这样就会导致泛化性能下降，这种现象称为过拟合

- Page23: 过配

同过拟合

- Page23: 经验误差(267)(empirical error)

学习器在训练集上的误差称为“训练误差”

- Page23: 欠配（underfitting）

欠拟合，对训练样本的一般性质尚未学好

- Page23: 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称为误差
- Page23: 训练误差（trainning error）

  同经验误差
- Page24: 模型选择(model selection)

  选择学习算法与参数配置
- Page25: 分层采样(stratified sampling)

  如果从采样的角度看待数据集的划分过程，则保留类别比例的采样方式通常称为“分层采样”
- Page25: 留出法（hold-out）

  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，在S上训练出模型后，用T来评估其测试误差，作为对泛化误差的估计。
- Page26: k折交叉验证（k-fold cross validation）

  交叉验证先将数据集D划分为k个大小相似的互斥子集，每个自己都尽可能保持数据分布的一致性，即从数据集中分层采样得到，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，这样就可以获得k组训练/测试集，最终返回k个测试结果的均值，交叉验证评估结果的稳定性和保真性很大程度上取决于k的取值，通常称之为k折交叉验证，最常用的k是10
- Page26: 交叉验证法（cross validation）

  同k折交叉验证
- Page27: 包外估计(179)（out of bag estimate）

  用于测试的样本没在训练集中出现，这样的测试结果称为包外估计
- Page27: 自助法(bootstrapping)

  以自主采样法为基础，给定包含m个样本的数据集D，对它采样产生数据集D’：每次随机从D中挑选一个样本，将其考本放入D’， 然后再将该样本放回D中，下次可能再被采到，这个过程执行m次后，得到包含m个样本的数据集D’,m足够大时，有36.8%的样本不会被采到，于是可以用没采到的部分做测试集。
- Page28: 参数调节（parameter tuning）

  大多数学习算法有些参数需要设定，参数配置不同，学得模型的性能往往有显著差别，因此，在进行模型评估与选择时，除了要对适用学习算法进行选择，还需要对算法参数进行设定，这就是参数调节或者调参。
- Page28: 验证集(105)（validation set）

  通常把学得模型在实际使用中遇到的数据称为测试数据，为了加以区分，为了加以区分，模型评估与选择中用于评估测试的数据集常称为“验证集”。
- Page29: 均方误差(54)（mean squared error）

  回归任务最常用的性能度量是均方误差（几何距离）

- Page30：查全率（recall）

  预测为真且正确的结果占所有预测正确的结果的比例。
- Page30：查准率（precision）

  预测为真且正确的结果占所有预测结果的比例。
- Page30：混淆矩阵（confusion matrix）

| 真实情况 | 预测为正例 | 预测为反例 |
| --- | --- | --- |
| 正例 | TP（真正例） | FN（假反例） |
| 反例 | FP（假正例） | TN（真反例） |

- Page30：召回率

  同查全率
- Page31：平衡点（break-even point，bep）

  查准率=查全率时的取值。平衡点大的学习模型可以认为综合性能更好
- Page32：F1

  查准率和查全率的调和平均，比算术平均（求和除以2）和几何平均（平方相乘开方）更重视较小值。

  $1/F1 = 1/2 \ (1/P + 1/R)$

  $1/F\beta = 1/(1+\beta) \ (1/P + \beta^2/R)$
- Page32：宏F1(macro-F1)

  如果进行多次训练/测试，每次得到一个混淆矩阵，或是在多个数据集上进行训练/测试，可以在n个混淆矩阵上综合考察查准率和查全率

  $macro\text{-}P = 1/n(\sum Pi)$

  $macro\text{-}R = 1/n(\sum Ri)$

  $1/macro\text{-}F1 = 1/2*(1/macro\text{-}P + 1/macro\text{-}R)$
- Page32：微查准率

  将各混淆矩阵的对应元素进行平均，再去计算
- Page32：微查全率
- Page33：ROC曲线(46)

  真正例率（True Positive Rate，TPR）和假正例率（FPR）的关系曲线

TPR = TP/(TP+FN)

FPR = FP/(TN+FP)

- Page35: 代价(47)(cost)

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”

- Page35: 代价矩阵

| 真实情况 | 预测为0类 | 预测为1类 |
| --- | --- | --- |
| 0类 | 0 | cost01 |
| 1类 | cost10 | 0 |

- Page36: 代价敏感(67)(cost-sensitive)

在损失函数中考虑了非均等代价

- Page36: 代价曲线

正例概率代价（横轴）和归一化代价（纵轴）的曲线

正例概率代价： P(+)cost = p * cost01 /( p * cost01 + (1 - p) * cost10)，p是样例为正例的概率

归一化代价： cost_norm = (FNR * p * cost01 + FPR * (1-p) * cost10)/(p * cost01+ (1-p) * cost10)

- Page36: 规范化(183)(normalization)

将不同变化范围的值映射到相同的固定范围中，常见的是[0,1]，此时亦称归一化

- Page36: 归一化（regular）

同规范化

- Page36: 总体代价

错误率是直接计算错误次数，并没有考虑不同错误会造成不同的后果，在非均等代价下，我们所希望的不再是简单的最小化错误次数，而是希望最小化总体代价

- Page37: 假设检验(hypothesis test)

假设是对学习器泛化错误率分布的某种判断或猜想，用测试错误率估计泛化错误率，以检查学习器性能。

- Page38: 二项检验（binomial test）

二项分布检验，根据收集到的样本数据，推断总体分布是否服从某个指定的二项分布。泛化错误率为e的学习器被测得测试错误率为e’的概率是服从二项分布的。

- Page38: 置信度(confidence)

估计总体参数落在某一区间时，可能不犯错误的概率，一般用符号1-α表示。

- Page40: 交叉验证成对t校验（paired t-tests）

  对两个学习器A和B，使用k折交叉验证法分别得到k个测试错误率，如果两个学习器性能相同，则使用相同训练/测试集时测试错误率应该相同，求两个学习器的k个测试错误率的差，若abs(sqrt(k)*μ/σ)＜临界值则认为两个学习器性能相同。
- Page41: 5x2交叉验证

  由于交叉验证中，不同轮次的训练集之间有一定程度的重复，会过高估计假设成立的概率，因此做5次2折交叉验证，每次验证前将数据打乱，对5次2对2个学习器的测试错误率求差值，对所有差值求方差，对前两次差值求均值，再进行临界值判断。
- Page41: McNemar检验

  两个学习器分类差别列联表

| 算法B\A | 正确 | 错误 |
| --- | --- | --- |
| 正确 | e00 | e01 |
| 错误 | e10 | e11 |

检验变量|e01-e10|是否服从正态分布，服从则认为两学习器性能相同等同于检查 $\tau x^2 = (|e01-e10|-1)^2/(e01+e10)$ 是否服从自由度为1的卡方分布（标准正态分布变量的平方）

- Page41: 列联表(187)

  见McNemar检验
- Page42: Friedman检验

  有多个数据集多个学习器进行比较时使用，对各个算法在各个数据集上对测试性能排序，对平均序值计算 $\tau x^2$ 和 $\tau F$,并进行临界值检验。
- Page43: Nemenyi后续检验(Nemenyi post-hoc test)

  学习器性能性能显著不同时，进行后续检验来进一步区分各算法，临界值域：CD=qα* sqrt(k*(k+1)/6N)
- Page44: 偏差-方差分解(177)

  对学习算法的期望泛化错误率进行拆解，学习算法在不同训练集上学得的结果很可能不同，真实输出与期望输出的差别称为偏差(bias)，使用样本数相同的不同训练集产生的输出的方差为var(x)，有：$E(f;D) = bias^2(x) + var(x) + \varepsilon^2$

- 根目录是因为手机／app有root权限吧？是不是因为SQLite没有考虑在root的手机上运行所有才会这样写？
- --lxc902
- 图挺好看的，请问是用什么软件绘制的?
- --oylz