Task01

- 1&2Chapter

  基本概念和术语

- 绪论
  - 概述
    - 机器学习致力于研究如何通过计算的手段，利用经验来改善系统自身的性能。“经验”通常以“数据”形式存在。
    - 机器学习所研究是关于在计算机上从数据中产生“模型”的算法，即“学习算法”。
    - Mitchell，1997 给出了一个更形式化的定义:假设用P来评估计算机程序在某任务类T上的性能，若一个程序通过利用经验E在T中任务上获得了性能改善，则我们就说关于T和P，该程序对E进行了学习。
  - 术语:
    - 一组记录的集合叫数据集dataset,其中每条记录都是一个事件或对象，称为instance or sample,反映这个事件、对象某方向的表现或性质的事项叫属性(attributes)或特征（feature),属性上的取值称为属性值（attributes value）。属性张成的空间称为属性空间（attributes space） 或样本空间(sample space).每个示例都能在空间中对应一个坐标向量，这个示例的向量称为“特征向量”（feature vector）
    - 从数据中学习模型的过程称为learning or training,训练过程的数据称为training data,其中每个样本称为training sample,样本组成的集合称为training set。
    - 预测是离散值称为分类classification，预测的是连续值，任务称为“回归”regression.聚类“clastering”
    - 根据训练数据是否拥有标记信息，学习任务分为“监督学习”和“无监督学习”（supervised learning and unsupervised learning.）
  - 假设空间
    - 归纳induction和演绎deduction是科学推理的两大基本手段。
  - 归纳偏好
    - inductive bias 任何一个有效的机器学习算法必有其归纳偏好。
    - 奥卡姆剃刀：若有多个假设与观察一致，则选择最简单的那个。
    - “没有免费的午餐No free lanch theorem”定理:无论学习算法a多聪明、算法b多笨拙，它们的期望性能相等，脱离具体的问题，谈什么学习算法更好毫无意义。
- 模型评估与选择
  - 经验误差与过拟合
    - 分类错误的样本占样本总数的比例称为“错误率”error rate
    - 1 - error rate = accuracy精度，即精度= 1 - 错误率。

    - 误差：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”，在training set上的误差叫training error 或者empirical error. 在新样本上的误差叫generalization error(泛化误差)
    - 过拟合overfitting:学习能力太好，欠拟合underfitting：学习能力太低下。欠拟合比较好解决，比如在决策树学习中拓展分支、在神经网络学习中增加训练轮数等。
- 评估方法
    - 使用测试集testing set 来测试学习器对新样本的判别能力，以测试误差作为泛化误差的近似。
    - 留出法（hold-out）：把数据集D划分为两个互斥的集合，一个训练集S，一个测试集T. 训练/测试的划分尽可能保持数据分布的一致性。
    - 交叉验证法：将D划分为k个大小相似的互斥自己。每个子集都尽可能保持数据分布的一致性（从D中通过分层采样得到）。然后每次用k-1个子集的并作为训练集，余下的子集作为测试机。这样就可以获得k组训练集和测试集，从而进行k次训练和测试，最终返回的是这个k个测试结果的均值。
        - 优点：减小因样本划分不同而引入的差别，k折交叉验证通常要随机使用不同的划分重复p次，最终个评估结果是p次k折交交叉验证结果的均值。
    - 留一法（Leave-One-Out）
    - 自助法（bootstrapping）：以自助采样法boostrap sampling。
        - 包含m个样本的数据集D中，采样产生数据集D'：每次随机从D中挑选样本放入D'，然后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踩到。这个过程重复执行m次后，就得到了包含m个样本的数据集D'。通过自助采样，初始数据集。中约有36.8%的样本未出现在采样数据集D'. 实际评估的模型与期望评估的模型都使用馆个训练样本,而我们仍有数据总量约1 /3的、没在训练集中出现的样本用于测试.这样的测试结果，亦称“包外估计”(out-of-bag estimate).
        - 适用于数据集较小、难以有效划分数据集时很有用。
        - 但会改变初始数据集的分布，会引入估计偏差。
    - 调参与最终模型
        - 除了对适用学习算法的选择，还需要对算法参数进行设定。这就是参数调parameter tuning.
        - 我们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，模型评估与选择中用于评估测试的数据集常称为“验证集" (validation set).例如，在研究对比不同算法的泛化性能时，我们用测试集上的判别效果来估计模型在实际使用时的泛化能力，而把训练数据另外划分为训练集和验证集,基于验证集上的性能来进行模型选择和调参.
- 性能度量Performance measure
    - 均方误差mean sqaure error
    - 分类任务中常用的性能度量
        - 错误率与精度

- 查准率precision、查全率recall与F1
  - TP、FP、TN、FN
  - confusion matrix.
  - Precision = TP/(TP+FP)
  - recall = TP/(TP+FN)
  - P-R曲线 查准率-查全率曲线，去线下的面积的大小一定程度上表征了学习器在查准率和查全率上取得相对“双高”的比例。
  - BEP(break-event point) 平衡点，查全率=查准率时的取值。
  - F1 = 2 * TP/(Samples+TP-TN)
- ROC与AUC
  - 学习器产生的实值或概率预测，与一个分类阈值threshold进行比较，若大于阈值则分为正类，否则为反类。
  - ROC曲线Receiver Operating Characteristic。受试者工作特征。
  - ROC纵轴是“True Positive Rate”,横轴是“False Postive Rate”
  - AUC(Area Under ROC curve)
- 代价敏感错误率与代价曲线
  - 设定一个“代价矩阵”cost matrix

表 2.2　二分类代价矩阵

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

- 比较检验
  - 假设检验hypothesis test
  - 交叉验证t检验
  - McNemar检验
  - Friedman检验与Nemenyi后续检验
- 偏差与方差bias-variance decomposition
  - 期望输出与真实标记的差别称为偏差(bias),刻画学习算法本身的拟合能力
  - 方差度量了同样大小的训练集的变动所导致的学习性能的辩护啊，刻画的是数据扰动所造成的的影响

> - 噪声是当前任务上任何学习算法所能达到的期望泛化误差的下界，刻画的是学习问题本身的难度。
> - 泛化误差可分解为偏差、方差与噪声之和.
> - 好的泛化性能，则需使偏差较小，即能够充分拟合数据，并且使方差较小，即使得数据扰动产生的影响小.