机器学习

绪论

- 什么是机器学习
  - 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。
- 监督学习
  - 有无标记信息
  - 分类
    - 离散值，二分类：好瓜坏瓜
    - 区别：1 输出不同：（1）分类输出离散值，回归输出连续值 （2）分类输出的是类别，回归输出具体的值 （3）分类定性，回归定量 2 目的不同：分类寻找决策边界，得到一个决策面对样本分类；回归用于寻找最优拟合曲线
  - 回归
    - 连续值，成熟度：0.5 0.9
  - 聚类
- 无监督学习
- 泛化能力
  - 该模型适用于新样本的能力
- 假设空间
  - 我们可以把学习过程看作一个在所有假设组成的空间中进行搜索的过程，搜索目标是找到与训练集匹配的假设，即能够将训练集中的元判断正确的假设。
- 版本空间
  - 学习过程是基于有限样本训练集进行的，存在着一个与训练集一致的假设集合
- 奥卡姆剃刀
  - 常用的，自然科学研究中最基本的原则，即若有多个假设与观察一致，则选最简单的那个。当存在多条曲线与有限样本训练集一致时，选择更简单平滑的曲线
- 没有免费的午餐NFL
  - 脱离具体问题，空泛的谈论什么学习算法更好毫无意义，要谈论算法的相对优劣，必须要针对具体的学习问题，在某些问题上表现好的学习算法，在另一些问题上却可能不尽如人意

模型评估与选择

- 过拟合
  - 学习能力过强，避免困难：训练误差大，测试误差大
  - 神经网络：优化目标增加正则项，early stop 决策树：预剪枝，后剪枝
- 欠拟合
  - 学习能力过弱，容易克服：训练误差小，测试误差大
  - 神经网络：增加训练次数 决策树：拓展分支
- 评估方法
  - 留出法：留out的数据作为测试集，剩下of作为训练集。
    - 优点：深度学习最常用的方法，简单易懂好操作，计算成本还小
    - 缺点：但是结果可能可信度不算很高
  - k折交叉验证法：将原数据切成大小相同的K份，然后开始进行K重循环，第次循环使用第i份的数据作为测试集，剩余K-1份数据作为训练集
    - 优点：可信度高 缺点：计算量比较大，适合传统机器学习模型，不适合大型神经网络
  - 自助法：假设原数据集为D，大小为m，我们有放回地从D中依次采样m次，得到D'，将D'作为训练集，剩下的样本作为测试集
    - 优点:数据集较小、难以有效划分训练/测试集时比较有效
    - 缺点：改变了原数据的分布，引入了估计偏差，在数据集较大时会非常明显
- 性能度量
  - 均方误差：回归任务最常用的性能度量
  - 错误率与精度：分类任务最常用的性能度量
  - 查准率、查全率与F1
    - 查准率：在预测结果中，预测出的真实正例所占所有预测正例中的比例
    - 查全率：在真实情况中，预测出的真实正例所占所有真实情况中的比例
    - F1：基于查准率与查全率的调和平均
- 偏差与方差
  - 泛化误差可分解为偏差、方差与噪声之和
    - 偏差度量了学习算法的期望预测与真实结果的偏离程度，即刻画了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变化，即刻画了数据扰动所造成的影响
    - 噪声则表达了在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，即刻画了学习问题本身的难度

线性模型

- 基本形式
  - 一般形式：f（x）=w1x1+w2x2+...wdxd+b
  - 向量形式：f（x）=wTx+b
  - 优点：（1）形式简单，易于建模（2）线性模型引入层级结构或高维映射可得到非线性模型（3）可解释性强
- 线性回归
  - 基于均方误差最小化来进行模型求解的方法称为最小二乘法。在线性回归中，最小二乘法就是试图找到一条直线，使所有样本到直线上的欧式距离之和最小。
- 对数几率回归
  - 最理想的函数：单位阶跃函数 预测值大于零就判为正例，小于零就判为反例，预测值为临界值零则可任意判别
  - 由于单位阶跃函数不连续，因此用对数几率函数替代
- 线性判别分析LDA
  - 极大似然方法可以来估计w和b，求解最优解可以使用梯度下降法和牛顿法
  - 二分类问题：主要用于降维 有标签用LDA 无标签用PCA
- 多分类学习
  - 一对一OVO
  - 一对其余OVR
  - 多对多MVM
    - 纠错输出码：对N个类别做M次划分，每次划分将一部分类别划为正类，一部分划为反类，从而形成一个二分类训练集
    - 海明距离：与预测示例不同的次数
    - 欧氏距离：差值的平方和
    - 对于同一个学习任务，ECOC编码越长，纠错能力越强。对同等长度的编码，任意两个类别之间的编码距离越远，则纠错能力越强
- 类别不平衡问题
  - 类别不平衡是指分类任务中不同类别的训练样例数目差别很大的情况
  - 欠采样：去除一些反例使得正、反例数目接近
  - 过采样：增加一些正例使得正、反例数目接近
  - 阈值移动：直接基于原始训练集进行学习，但在用训练好的分类器进行预测时，将公式嵌入到决策过程中

决策树

- 划分选择
  - 信息增益
    - 信息熵度量样本集合纯度最常用的一种指标。Ent(D)值越小，D纯度越高
  - 增益率
    - 信息增益 Gain越大，意味着使用属性a来进行划分所获得的纯度提升越大
    - 信息增益准则对可取值数目较多的属性有所偏好，为减少这种偏好可能带来的不利影响，使用增益率来选择最优划分属性。增益率准则对可取值数目较少的属性有所偏好，a取值数目越多，IV（a）值越大
  - 基尼指数
    - 数据集D的纯度可用基尼值来度量，Gini(D)越小，数据集D的纯度越高
- 剪枝处理
  - 预剪枝
    - 优点：（1）降低过拟合风险（2）显著减少决策树的训练时间开销和测试时间开销
    - 缺点：欠拟合风险：有些分支的当前划分虽不能提升泛化性能，甚至可能导致泛化性能暂时下降，但在其基础上进行的后续划分却有可能导致性能显著提高
  - 后剪枝
    - 优点：后剪枝决策树比预剪枝保留了更多的分支。欠拟合风险小，泛化性能往往优于预剪枝决策树
    - 缺点：训练时间开销大，后剪枝过程是在生成完全决策树之后进行的，而且要自底向上地对树中的所有非叶结点进行逐一考察

神经网络

- 感知机与多层网络
  - 感知机
    - 感知机由两层神经元组成，输入层接受外界输入信号后传递给输出层，输出层是M-P神经元，亦称阈值逻辑单元 容易实现与或非运算
  - 多层前馈神经网络
    - 每层神经元与下一层神经元全互联，神经元之间不存在同层链接，也不存在跨层链接
- 误差逆传播算法BP
  - 误差逆传播算法是最成功的训练多层前馈神经网络的学习算法
  - 标准BP算法：每次更新只针对单个样例，参数更新非常频繁，而且对不同样例进行更新的效果可能出现抵消现象。需要多次迭代
  - 累积BP算法：其优化的目标是最小化整个训练集上的累计误差
- 全局最小与局部最小

支持向量机

- 间隔与支持向量
  - 间隔：两个异类支持向量到超平面的距离之和
  - 支持向量：距离超平面最近的训练样本点使等式成立的向量
  - 支持向量机SVM就是利用间隔最大化求最优分离超平面
- 核函数
  - 从原始空间向高维空间映射解决线性不可分问题，使得样本在这个特征空间内线性可分
- 软间隔与正则化
  - 软间隔不用所有样本必须满足约束，不用都划分正确
  - 正则化：结构风险描述模型某些性质，经验风险描述模型与训练数据的契合程度
- 核方法

贝叶斯分类器

- 贝叶斯决策论
  - 判别式模型：决策树、BP神经网络、支持向量机
    - 给定x通过直接建模p（c|x)来预测c
  - 生成式模型：贝叶斯分类器
    - 先对联合概率分布p（x，c）建模，然后由此获得p（c|x）
  - 贝叶斯定理：p（c|x)=p(c)p(x|c)/p(x)，p（c）先验概率，p（x）证据，p（x|c）似然，p（c|x）后验概率
- 极大似然估计
- 朴素贝叶斯分类器
  - 采用属性条件独立性假设，对已知类别，假设所有属性相互独立，假设每个属性独立的对分类结果产生影响
  - 拉普拉斯修正：为了避免其他属性携带的信息被训练集中未出现的属性值"抹去"，在估计概率值时要进行平滑，避免了因训练集样本不充分而导致概率值为零的问题。
- 半朴素贝叶斯分类器
  - 为了降低贝叶斯公式中估计后验概率的困难，朴素贝叶斯分类器采用了属性条件独立性假设，对属性条件独立性假设进行一定程度的放松，由此产生了一类成为半朴素贝叶斯分类器的算法
  - 最常用的一种策略：独依赖估计，假设每个属性在类别之外最多仅依赖于一个其他属性
- 贝叶斯网
  - 贝叶斯网，亦称信念网，借助有向无环图刻画属性之间的依赖关系，并使用条件概率表来描述属性的联合概率分布
  - 贝叶斯网结构有效地表达了属性间的条件独立性，给定父结点，贝叶斯网假设每个属性与它的非后裔属性独立
  - 贝叶斯网学习的首要任务就是根据训练数据集来找出结构最恰当的贝叶斯网，用评分函数来评估贝叶斯网与训练数据的契合程度
- EM算法
  - 常用的估计参数隐变量的利器，是一种迭代式方法，基本想法是若参数一致则可根据训练数据推断出最优隐变量z的值，反之若z的值已知，则可方便的对参数做极大似然估计

集成学习

- 个体与集成
  - 集成学习通过构建并结合多个学习器来完成学习任务，有时也被称为多分类器系统
  - 同质：同种类型的个体学习器，也称基学习器，对应的算法成为基学习算法
  - 异质：不同类型的个体学习器，成为组件学习器
  - 集成个体应好而不同
- boosting 强依赖，串行生成
  - （1）先从初始训练集训练出一个基学习器；
  - （2）根据基学习器的表现对训练样本分布进行调整，使得先前基学习器分错的训练样本在后续得到更多关注。然后再基于调整后的样本分布来训练下一个基学习器；
  - （3）重复（2），直到基学习器数目达到指定值T，最终将这T个基学习器进行加权结合。

  从偏差-方差分解角度看，Boosting 主要关注降低 偏差
- bagging 与随机森林 不存在强依赖，并行化生成
  - （1）Bagging的基本流程：（放回？ 放空？）
    通过自助采样法采样出T个含m个训练样本的采样集。然后基于每个采样集训练出一个基学习器。再将这些基学习器进行结合。
    （2）Bagging采用 自助采样法 包外估计
    （3）从偏差-方差分解的角度看，Bagging 主要关注降低 方差
    （4）Bagging的包外估计应用：回归型算法
    Bagging回归的包外采用：简单平均法

    随机森林（RF）是Bagging的一个扩展变体

    RF在 以决策树为基学习器 构建Bagging集成的基础上，进一步在决策树的训练过程中引入了 随机属性选择 。

    随机森林的多样性体现在：采样随机性；属性选择随机性。
  - 随机森林最突出的 优点：（1）训练高度并行化，提高速度（2）随机选择决策树节点划分特征，得到高效模型（3）随机采样，训练模型方差小，泛化能力强 缺点：在某些噪音比较大的样本集上，RF模型容易陷入过拟合

聚类

- 聚类任务
  - 在无监督学习中，训练样本的标记信息是未知的，目标是通过对无标记训练样本的学习来揭示数据的内在性质及规律，为进一步的数据分析提供基础
  - 聚类试图将数据集中的样本划分为若干个通常不相交的子集
- 性能度量
  - 外部指标：将聚类结果与某个参考模型进行比较
    - Jaccard系数：JC=a/a+b+c
    - FM指数
    - Rand指数
  - 内部指标：直接考察聚类结果而不利用任何参考模型
    - DB指数，DBI越小越好
    - Dunn指数，DI越大越好
- 距离计算
  - 函数dist基本性质
    - 非负性
    - 同一性
    - 对称性
    - 直递性
  - 闵可夫斯基距离公式
    - p=2时，欧氏距离
    - p=1时，曼哈顿距离
- k均值算法

  K均值聚类算法

  算法流程：

  1. 指定需要划分的簇的个数K。
  2. 随机选择K个数据对象作为初始聚类中心，这些数据对象不一定是已有的样本点。
  3. 逐一计算其他各个数据点到这K个初始聚类中心的距离，把数据对象划分到距离它最近的那个中心所在的簇中。
  4. 调整新的簇的聚类中心。
  5. 循环执行步骤2和步骤3，观察聚类中心是否收敛（也就是样本不发生变化），如果收敛或达到最大迭代次数则终止循环过程。

  备注：距离的计算也可以使用曼哈顿距离，闵可夫斯基距离等（泛化的欧氏距离），