NLP方向总结-机器学习

论文

XGBoost： A Scalable Tree Boosting System

- XGBoost
  - 定义
    - 是一个Tree Boosting模型，scalable end-to-end tree boosting system
    - 使用示例：商店销售预测；高能物理事件分类；网络文本分类；客户行为预测；运动检测；广告点击率预测；恶意软件分类；产品分类；危险风险预测；大规模在线课程退学率预测
    - 优点
      - 可扩展性：运行速度快，支持分布式和并行计算，核外计算
      - 可处理处理稀疏数据
  - 模型
    - 使用K个加法函数预测输出
      - 例如图1：一条数据的预测结果y^hat是所有树的预测结果之和，一条数据经过一个树后得到一个分数score
      - 树的权重，这里就是分数，wi
    - 训练目标加了正则化项（公式2）
      - 预测值和真实值的差最小+正则化项
      - 防止过拟合
      - 梯度提升的解释，这次的预测结果是由上一次的预测结果相加得到的（公式2.1）
        - 最终简化目标函数为公式3
          - t代表第t次迭代
          - gi表示第i个样本的一阶梯度
          - ft(Xi)是第i个样本的树函数
          - hi是第i个样本的二阶梯度
          - Ωf(t)是正则化项
            - 展开正则化项
              - 得到一个树的叶子节点的权重（分数）的计算方式，公式5
                - 简化后得到公式6，每颗树的损失函数
                - T表示树中叶子的数量
                - gama表示树的叶子节点上做进一步分区所需的最小损失减少
                - 程序中默认设置为0
                - lamda是L2正则化项的超参数
                - 图2的Obj就是这颗数的损失函数，T已经换成了3，因为有3个叶子节点
    - 权重缩小
      - 每一步 tree boosting之后将新增加权重按系数η进行缩减
      - 防止过拟合
    - 列下采样
      - 不仅可以降低过拟合，还可以减少计算量
      - 借鉴RandomForest的思想，采取列的下采样
    - 加权量化简图
      - 寻找树的分割点的近似算法，解决数据量过大时，超过内存的问题
        - 变种1：全局变体在构建树的初始阶段提出所有的候选分裂
        - 变种2：局部变体在每次分裂后重新提出建议
        - 变种1和变种2的比较（图3）
        - 具体操作：附录中的合并和修剪运算部分
      - 对比精确贪婪算法，列举了所有可能的分割点
    - 平行树的稀疏感知算法
      - 加速运算
        - 使计算复杂度与输入的非缺失项数量成线性关系
          - 加快50倍
      - 处理稀疏数据，1）数据中存在缺失值；2）统计中经常出现零条目；3）特征工程的人工处理，如one-hot编码
      - 如图4，设定一个默认的方向，指向一个叶子节点，图中的红色线就是如果数据确实，那么数据的默认走向的叶子节点
        - 如果用XGB模型绘图plot_tree时，路径上显示的missing就是表示如果数据这个特征缺失，那么数据的被分的方向
      - 的并行是特征粒度上的，在计算特征的Gain的时并行
    - 缓存感知的块结构（图6）
      - 高速缓存和数据压缩
        - 加速运算
      - 将数据存储在内存单元称为块
        - 每个块中的数据都以压缩列（CSC）格式存储
        - 每一列都按相应的特征值排序
      - 减少数据排序的成本

Concept Paper Kano Model Integration with Data Mining to Predict Customer Satisfaction

- 加诺Kano模型与机器学习
  - 对影响客户满意度的因素进行分析
    - 基本（必备）型质量——Must-be Quality/ Basic Quality
    - 期望（意愿）型质量——One-dimensional Quality/
    - Performance Quality 兴奋（魅力）型质量—Attractive Quality/ Excitement Quality
    - 无差异型质量——Indifferent Quality/Neutral Quality
    - 反向（逆向）型质量——Reverse Quality，亦可以将'Quality' 翻译成"质量"或"品质"。
  - 难点：将Kano模型与特征选择技术相结合，选出与客户满意度有关的最突出的属性
  - 使用的模型
    - 逻辑回归
    - 决策树回归
    - 随机森林回归
    - Adaboost回归
    - XGBoost回归
  - 模型性能评估
    - MAE平均绝对误差
    - RMSE均方根误差
    - R-Square即R2
  - 特征选择的方法
    - PCA主成分分析
    - 特征子集的相关性： 皮尔森系数等，斯皮尔曼系数等
    - Relief方法
    - EFA：探索性因数分析
    - SFFS： TFS监督下的前向特征选择
    - Filter-Wrapper
    - 信息增益（IG）
    - 卡方检验Chi-square（CHI）
    - Odds Ratio（OR）
    - Anova: 方差分析（ANOVA）又称"变异数分析"或"F检验"
    - 基于模型的预测的重要性
  - 其它类似加诺模型方法
    - PRCA
      - 惩罚-回报对比分析
    - IGA
      - 重要性网格分析
        - 对顾客明确陈述的重要性和研究人员隐含得出的重要性进行比较
    - 关键事件技术
      - 通过比较顾客在积极背景或消极背景下提及某一属性的频率来对属性进行分类
    - 直接分类法和定性数据方法
  - 数据集
    - 学生的满意度模型
    - 646份样本
    - 共36个特征
    - 回归问题: 满意度问卷
    - 分类问题：加诺问题问卷（表3）

模型

- CRF
  - 发展：Stochastic grammar->HMM->MEMM->CRF
  - 随机文法Stochastic grammar
  - 隐马尔可夫模型（HMM）
    - 生成模型: p(x,y)
      - 生成模型必须对观测值做出非常严格的独立性假设
      - 联合概率建模
    - HMM和随机文法是生成模型，为成对的观察序列和标注序列分配联合概率
    - HMM中是状态影响观测值
    - 1. HMM优化的是求解p(x,y)联合概率
      2. HMM中有向无环图
      3. HMM中是上一个时刻影响下一个时刻的状态
      4. HMM假设当前状态仅和前一个状态有关
  - 最大熵马尔可夫模型（MEMMs）
    - 条件概率序列模型 p(y|x)
      - 给定观测序列的可能标签序列的概率
      - 标签之间转移的概率可能不仅取决于当前观测，还取决于过去和未来的观测
      - MEMM中是观测值影响状态值
    - 每个源状态都有一个指数模型，该模型将观测特征作为输入，并输出下一个可能状态的分布
      - 转移分数是给定当前状态和观察序列的可能下一个状态的条件概率
      - 分数质量守恒，即到达一个状态的所有概率质量必须分布在可能的继承者之间，观察可以影响哪个目标状态获得质量，但不会影响传递的总质量
    - 缺点：标签偏差问题
      - 示例：给定观测序列为RIB，预测小写单词是rib还是rob
      - 如果这两个单词中有一个在训练集中稍微更常见，则从起始状态开始的转移将稍微倾向于其相应的转移，并且该单词的状态序列将始终获胜。
  - conditional random fields 条件随机场
    - 1. CRF优化的是求解p(y|x)条件概率
      2. CRF中是无向图
      3. CRF中是上下两个时刻的状态会影响当前时刻的状态值
      4. 条件随机场引入图模型(当前状态和相连状态都有关)+条件模型(当前状态和观测值有关)
    - 条件概率模型
      - 既具有MEMMs的所有优点，又有原则性地解决了标签偏差问题
      - CRF对给定观测序列的整个标签序列的联合概率使用单指数模型
      - 可以把CRF看作一个具有非标准化转移概率的有限状态模型
        - 根据实际观察结果进行对概率质量进行放大或衰减，并且对维特比路径的选择有更大的贡献。
      - CRF在可能的标签上分配了一个定义良好的概率分布，通过最大似然或映射估计进行训练。具有保证的全局最大似然收敛性。
      - 示例：假设词性标注任务，X是单词序列，即观察序列，Y是词性标注序列
        - G=（V，E）是一个Y=（Yv）的图，随机变量Yv服从关于图的马尔可夫性质，G是一条简单的链或线，如果Y的图G=（V，E）是一棵树（其中链是最简单的例子），那么它的团就是边和顶点

bayes

- 说说你对贝叶斯公式的理解？
  先验概率和后验概率的关系
  贝叶斯算法估计是一种从先验概率和样本分布情况来计算后验概率的一种方式。
  P(AB) AB 同时发生的概率，联合概率
- 先验概率P(A):在不考虑任何情况下，A事件发生的概率
  条件概率P(B|A):A事件发生的情况下，B事件发生的概率，这里把P(A)当作先验概率，如果把P(B)当作先验概率，那么P(B|A)就叫后验概率
  后验概率P(A|B):在B事件发生之后，对A事件发生的概率的重新评估
  P(A|B) = P(AB)/P(B)
- 从机器学习视角下，看待bayes
  X, Y 为随机变量
  P(Y|X) = P(X|Y)P(Y) / P(Y)
  我们可以把X理解为某种特征，把Y理解成类别标签
  P(所属类别|某种特征）= P(某种特征|所属类别) P(所属类别) / P(某种特征)
- 贝叶斯方法把计算"具有某特征的条件下属于某类"的概率转换成需要计算"属于某类的条件下具有某特征"的概率，属于有监督学习。
- 朴素贝叶斯理论源于随机变量的独立性，之所以称之为朴素是因为其思想基础的简单性：就文本分类而言，从朴素贝叶斯的角度来看，句子中的两两词之间的关系是相互独立的，即一个对象的特征向量中每个维度都是相互独立的。这是朴素贝叶斯理论的思想基础
- "朴素"原因
  - 简单性：就文本分类而言，从朴素贝叶斯的角度来看，句子中的两两词之间的关系是相互独立的，即一个对象的特征向量中每个维度都是相互独立的。
- 使用场景
  - 属于监督学习的生成模型，实现简单，没有迭代，学习效率高，在大样本量下会有较好的表现。但因为假设太强——假设特征条件独立，在输入向量的特征条件有关联的场景下并不适用。
- 贝叶斯误差
  - 分布是真实的，但是预测的输出只可能是一个值，不会列举出所有的情况，所以会存在误差，或者是预测准确结果的概率，也就是贝叶斯误差