研究生算法课课堂笔记

上课日期：11.7 第(2)节课

组长学号及姓名：李冠成 1601214512

组员学号及姓名：王东 1601214560 王靖博 1601214562

注意：请提交 Word 格式文档

XGBoost 介绍（续）

01 分类问题（续）

Q：为什么叶子的值不是概率，而是一个具体的值？
A:如果叶子上存储的是概率值,那么各棵树的预测值加起来很有可能超出[0,1)的范围。所以需要存储范围为$(-\infty,+\infty)$的 z 值。叶子上的值 z 会被带入到 sigmoid 函数，起到了 z->∞，正例的可能性会趋于 1，z->-∞，反例的可能性趋于 1

多分类问题

情景：已知病人的外观特征与病例，预测病人是什么病

数据格式：特征大多数为外观特征，如外观病变得严重程度等。除此之外还有家族史、年龄等。

数据预处理：加载数据文件、分测试集

不一样的参数：

- ‘objective’ = ‘multi：softmax’意思是使用多分类任务当中常用的 softmax Loss(本质是 logistic Loss 的多分类版本)。
- ‘nthread’ = 4 意思是使用四线程加速训练

Demo

1. 数据预处理，按照 7：3 的比例分 train 和 test
2. 数据训练

回归问题

情景： CPU 性能预测问题

相关特征： 厂商(离散型特征)、主频、最大最小内存等特征(连续型特征)

Demo

1. 数据预处理，离散特征 one hot encoding，连续性原封不动
2. 数据训练 objective 为 reg:linear（线性回归）

参考资料在 ppt 有

大作业

汽车投保风险指数预测

场景:

车辆保险公司具有评价投保汽车风险的需求，即根据汽车的各项指标对汽车的风险进行打分。因此我们需要训练一个模型拟合 y=f(x)，x 为描述汽车的 32 个指标，y 为汽车的风险指数，汽车风险指数是一个 0～70 之间的正整数，数值越大汽车的风险越高。

训练特征:

汽车的 32 个特征

训练目标：

投保风险指数

性能指标：

RMSE=sqrt((y-yˆ)ˆ2/N)，即越接近真实越好

数据介绍

1. train.csv 包含了车辆特征以及投保风险指数
2. test.csv 仅包含车辆的特征, 需要提交车辆投保风险指数的预测值
3. test_sample.csv 提交格式样例

数据格式

训练数据

- 4W 组数据
- 每行代表一个汽车，34 列，第一列为车辆 id，第二列为风险值，3～34 为 32 个特征（x 值）
- 特征中有数值型、有类别型

测试数据

- 每行代表一个汽车，33 列，第一列是车辆 id，第 2 到 33 列是汽车特征（x 值）
- 特征中有数值型、有类别型

任务流程

1. 特征提取，将 csv 转换成 xgboost 能处理的数据格式，主要是类别型转数值型，例如 one-hot encoding。
2. 模型训练，要把数据分为两部分，一部分为 training set，另一部分为 validation set。目标为学到一个模型，输入为训练特征 x，输出为训练目标 y。可以选用不同的参数训练多个模型，找一个最好的。
3. 使用训练得到的模型对 test.csv 的数据进行预测
4. 鼓励大家多试验不同的方法（或工具），如对 y 值取 log，sqrt，pow 等、possion 回归、logistics 回归、对多个模型进行组合等。

Hint

- Xgboost 优化目标 objective：线性回归 reg:liner,泊松回归 count:poisson（优化目标为最大似然估计）
- maxdepth 不要太大，以防止 overfitting
- learning rate 要较小 如 0.1
- 合理设置 nthread

大作业提交内容

- 对 test.csv 进行预测，按照 test_sample.csv 的格式提交
- 文件名为 组长学号_predict.csv
- 提交一个报告，叙述关键点
- 代码（包括如何执行）

QA

- 如果调用了比较牛逼的第三方库，要搞明白其原理
- 不允许到网上查解题报告
- 缺省状态下 3 个人分数一样，如有特殊情况写到报告里
- 不同的组间不允许抄袭
- 什么库都可以用