XGBoost在weka的实现与分析

**引言与背景**

XGBoost的特点

Xgboost是什么？

Xgboosttree的特点是什么？与其他树区别是什么

Weka中引入算法

Weka开发机器学习算法的简介

开发的产出是要得到什么

如何实验

算法实现

Xgboost中，已经实现依据nominal/numeric class 类型来判断损失函数是Logloss还是SquaredError，和对损失函数的一阶、二阶求导。针对logic类型，还进行了Sigmoid函数转换。

代码流程图和模块图如下：

在XGBoostTree中，接收到一阶和二阶导数，可以参与到树的决策运算中。

数学原理简要介绍如下：

对于Boost集成算法，我们的目标是把一个目标函数尽量减小。

Obj(Θ)=L(Θ)+Ω(Θ)

Gradient Boosting在每次迭代中将新的树加到过往的树来进行纠错，也就使最小化损失函数L(Θ)，

Xgboost使用泰勒展开公式来操作损失函数，将损失函数近似取到二阶展开。

损失函数可以表示为：∑i[L(yi,y^K−1i)+gifK(xi)+12hif2K(xi)]

在树中有这样的映射关系：

fK(x)=wq(x)

目标函数（移除常数项后）就可以用叶子节点表达为：

∑i[giwq(xi)+12hiw2q(xi)]+12λ∑Tj||wj||2+γT

最终得到

∑Tj=1[Gjwj+12(Hj+λ)w2j]+γT

对wj求导，然后带入极值点，可以得到一个极值：

w∗=−GjHj+λ

对正则项则取Ω(fK)=12λ∑Tj||wj||2+γT，t是叶子节点数，w是叶子值。

随机数引入

工作原理和意义

实现细节

实验

Introduction

In this study, we implemented the XGBoostTree class in Java for the Weka framework. This class was then utilized within the provided XGBoost class, aiming to facilitate to construct a comprehensive XGBoost learner in Weka.

XGBoost, standing for eXtreme Gradient Boosting, is a widely used gradient boosting framework. Its core principle is using the Taylor expansion of the loss function, specifically first and second derivative, to refine model predictions in each iteration, seeking to minimize the overall objective function. XGBoost trees incorporate regularization to reduce the risk of overfitting. The tree structures are optimized by gradient boosting framework and evaluate splits using the ‘Gain’.

Weka offers a platform for developing new algorithms and models. Integrating algorithms into Weka can facilitate research, provide toolkit for graphical interface. In this study the primary output is the algorithm module designed for integration with Weka and Java.

To evaluate the implemented XGBoost tree learning algorithm, we will compare it with the original XGBoost by performing experiments using Weka experimenter and then analyse the capabilities and potential refinements.