

# 学习笔记

乔林波

徐晓明

刘栋

团队：木头

LINBOQIAO@GMAIL.COM

HUMIXU0208@MOBIKE.COM

XXX

## 1. 赛题描述

### 1.1. 简介

对股票价格趋势的预测是金融领域极为复杂和极为关键的问题，有效市场假说认为股票价格趋势不可能被预测，然而真实市场由于各种因素的存在并不完全有效，这对于股票市场而言相当于一种“错误”。这里我们为参赛者提供了大规模的股票历史数据，从而可以通过集合大家的智慧来纠正股票市场的这些“错误”。

### 1.2. 数据

数据集包括训练数据集和测试数据集两部分。训练数据集是一个以逗号分隔的文本文件(csv)，其中：**id**列为数据唯一标识编码，**feature**列为原始数据经过变换之后得到的特征，**weight**列为样本重要性，**label**列为待预测二分类标签，**group**列为样本所属分组编号，**era**列为时间区间编号(取值1-20为时间顺序)。测试数据集是一个以逗号分隔的文本文件(csv)。测试数据集不包括**weight**列、**label**列和**era**列。

- **weight**取值最小为0，最大取到63。
- **group**取值共28组。
- **era**取值20组。

### 1.3. 评价标准

虚拟股票趋势预测比赛的评价指标类比一般二分类问题的评价方式，将最终的**logloss**值作为最终选手排名的依据，**logloss**的计算方法如下：

$$l(y_p; y_t) = - \sum_{i=1}^N (w_i * (y_t^i * \ln(y_p^i) + (1 - y_t^i) * \ln(1 - y_p^i))) \quad (1)$$

其中， $y_t = \{y_t^0, \dots, y_t^i, \dots, y_t^N\}$  是样本的真实标签， $y_p = \{y_p^0, \dots, y_p^i, \dots, y_p^N\}$  是样本预测为正类的概率， $w_i$  是第*i*个样本的样本权重， $N$ 是测试集样本数量。

## 2. 求解

### 2.1. 使用简单二分类求解

原问题视为一个二分类问题,

### 2.2. note

对于交叉验证, 建议按照训练数据`era`列随机抽取一个或若干个`era`进行交叉验证, 而不是在全部训练样本上进行随机采样进行交叉验证, 因为后者会导致严重的过拟合问题, 这也是我们加入了`era`列的主要目的。