# 信息内容安全课程报告

——股市分析

作者：周汈

学号：18307110272

**一、选题背景**

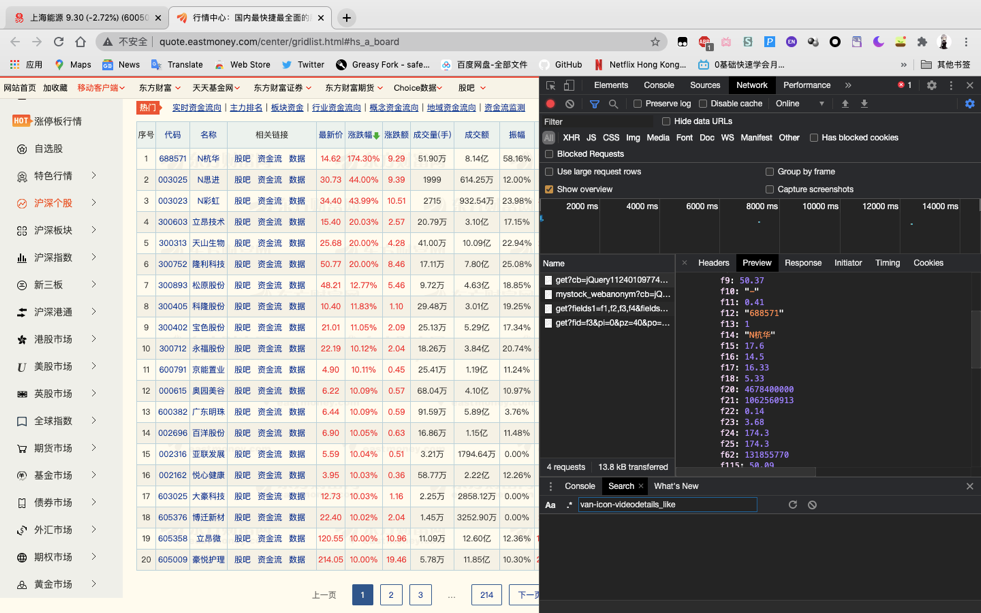
改革开放以来，随着市场经济的发展，虚拟经济迅速崛起。作为经济的重要组成部分，虚拟经济是市场信息经济发展的必然产物，对实体经济的发展和社会的繁荣有着巨大的促进作用。股票、基金、外汇等越来越多的金融产品也逐渐进入了人们的日常生活。对于普通人来说，学会利用金融产品理财毫无疑问是非常重要的。学会理财意味着能够将资本用于产生资本，使得收益率能够抗衡甚至超过通货膨胀的比率，避免由于静置资本导致的资本隐形缩水。

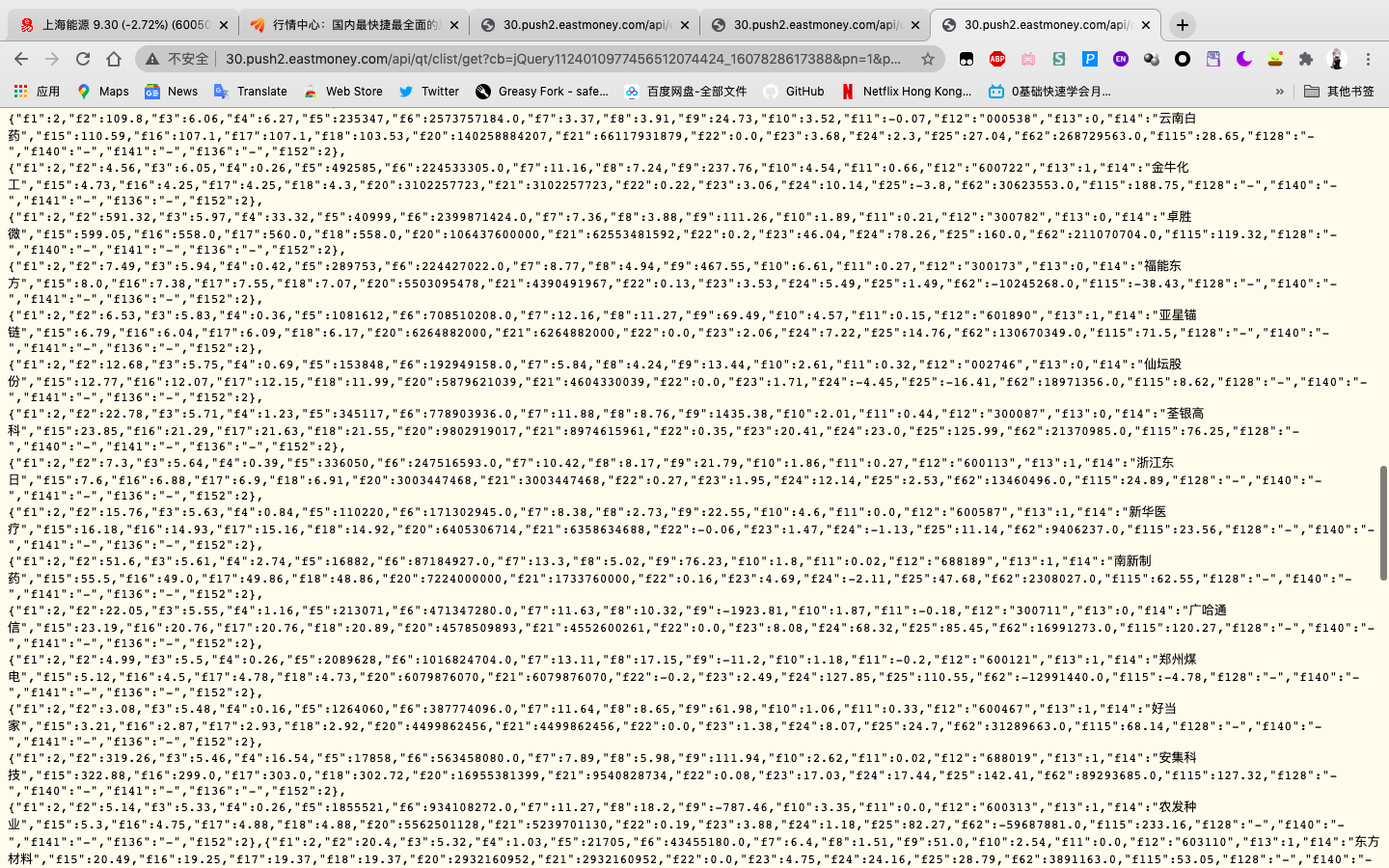
但是，理财并不是简单的拿钱取钱，正确的理财实践起来实际上是比较困难的。正所谓股市有风险，投资需谨慎，参与证券交易的盈亏状况虽不像彩票那般具有完全的随机性，但股市、基金、外汇市场的价格波动通常十分复杂，人们也仍难以通过简单的直觉对证券市场作出相对准确的长期预测。然而理论上，我们仍然是可以通过基于大数据分析的机器学习模型来，结合合理、科学的函数模型，利用计算机，对未来一段时间内股价的涨跌概率做出相对准确的预测。

基于以上理由，我们打算通过构建机器学习模型，结合互联网上记录的历史股价数据，建立股价涨跌预测程序，来实现对股市未来走向的初步预测，为理财的人提供一个参考。

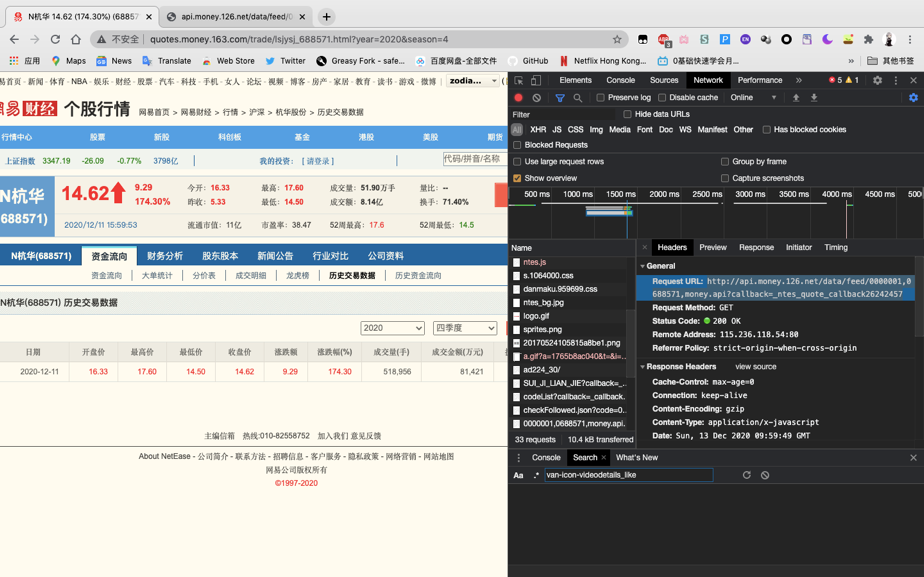
**二、数据抓取**

进行机器学习训练之前，我们将需要获取大量的数据来对模型进行训练，我所采取的方法是对一系列股票的10年以来每天的市值的变化来进行爬取，从而得到相当数量的数据。

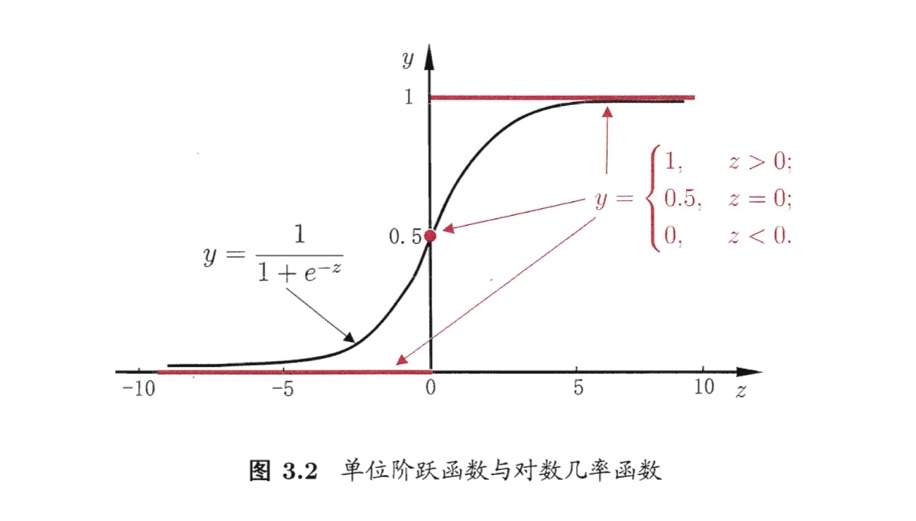
首先，我们来获取股票代码。我们从东方财富网沪深个股行情中心(<http://quote.eastmoney.com/center/gridlist.html#hs_a_board>)爬取股票代码，该页面内容由代码动态生成得到，因此对其页面进行解析得到数据的接口为[http://7.push2.eastmoney.com/api/qt/clist/get?cb=jQuery112407838651393251979\_1607852052096&pn=1&pz=20&po=1&np=1&ut=bd1d9ddb04089700cf9c27f6f7426281&fltt=2&invt=2&fid=f3&fs=m:0+t:6,m:0+t:13,m:0+t:80,m:1+t:2,m:1+t:23&fields=f1,f2,f3,f4,f5,f6,f7,f8,f9,f10,f12,f13,f14,f15,f16,f17,f18,f20,f21,f23,f24,f25,f22,f11,f62,f128,f136,f115,f152&\_=1607852052131](http://7.push2.eastmoney.com/api/qt/clist/get?cb=jQuery112407838651393251979_1607852052096&pn=1&pz=20&po=1&np=1&ut=bd1d9ddb04089700cf9c27f6f7426281&fltt=2&invt=2&fid=f3&fs=m:0+t:6,m:0+t:13,m:0+t:80,m:1+t:2,m:1+t:23&fields=f1,f2,f3,f4,f5,f6,f7,f8,f9,f10,f12,f13,f14,f15,f16,f17,f18,f20,f21,f23,f24,f25,f22,f11,f62,f128,f136,f115,f152&_=1607852052131)，经过分析其接口后我们又会发现pn所对应的是页数，而pz所对应的是一次爬取的条数，其他参数没有任何影响，因此我们将pn保持不变，而pz修改为4000，可以一次将所有的数据全部爬取，将整个json文件保存后我们将对其尝试解析。



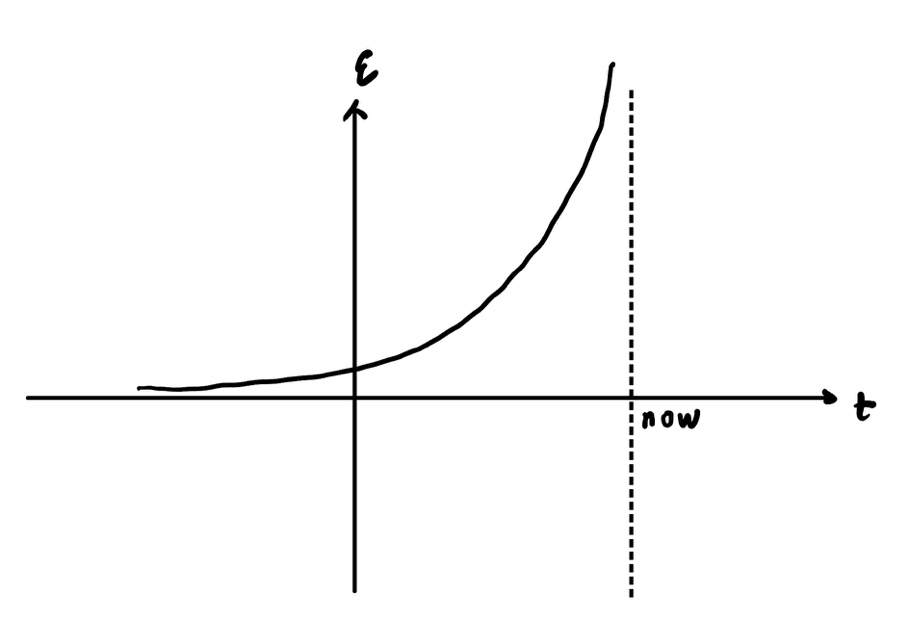
对json中的内容进行解析，由于所有数据保存方式完全相同，因此通过正则表达式可以方便提取其中的数据，我们所要的是股票的代码，因此通过python提取之后存放在mysql数据库中，这里我们一共得到了4000只股票的代码，这里考虑到在后续历史查询中可能有的股票无法查找到，又或者是有的股票诞生时间太短，因此我们选择相对较多的样本量。

在得到股票的代码之后，需要他们的历史记录，因此这次我们使用网易财经网站(http://quotes.money.163.com/trade/lsjysj\_600896.html?year=2018&season=4)的历史交易数据。观察之后可以得知，该网址中的600896为该股票的代码，year为年份，season为季度，而该页面属于静态页面（我后来发现貌似有数据传入的，其接口为<http://api.money.126.net/data/feed/0000001,0688571,money.api?callback=_ntes_quote_callback26242457>，但是直接通过抓取貌似逻辑更简单），那么抓取策略就成为了将三部分进行拼接得到url后爬取静态页面进行解析即可，这里我们选取2010年到2020年的数据进行爬取，得到十年内特定股票的每一天的收盘价。将收盘价，时间日期，股票的代码重新存入另一张表格中，作为之后机器学习的训练集。初步推测应有样本数4000 \* 10 \* 365 = 14600000。

**三、机器学习模型**

 这里我采用的是logistic回归模型进行训练，对其输出值的解释是输出值即为涨或者跌的概率，但是与传统的逻辑回归不同，我对其算法进行了自己的改造。传统的逻辑回归属于二分类算法，所采用的算法是对于一系列的变量x，通过赋予权重之后加和得到z，将z代入sigmoid函数中得到y的预测值，通过代价函数对权重进行更新从而得到代价函数的极值。

我们所要做的就是根据之前的数据来预测一只股票之后的涨跌趋势，针对传统逻辑回归的改造，首先我们来引入一个概念——影响力，该参数与时间相关，即一个因子距今时间越久，其对目前结果的预测是越小的，那么为了预测股票的涨落，我们选取每两天之间的差值作为我们的变量，当差值为负，代表股票跌了，当差值为正，代表股票涨了，而我们认为，距今时间越远，其结果对目前的结果的影响就越小，那么这个影响就是我们之前所说的影响力的概念，当把影响力施加在变量上后，影响力因子其实也就相当于传统逻辑回归中的权重，只不过影响力因子之间是有相互关联的，而权重彼此独立的，所以我把这种改造称之为非独立系数逻辑回归(dependent coefficient logistic regression)。

那么我们所要做的就是，首先给出影响力因子之间的相互关系，而且需满足随着时间接近现在，其值迅速增大，如下图所示，那么给出一个ε=f(θ, t)函数形式，作为权重计算，这里的∆pi是涨落差，再将z代入sigmoid函数中计算涨落的概率估计值，通过估计值与真实涨落(0或1，0代表降低，1代表升值)的差来对θ进行更新，给出单个变量的损失函数，我们所要求的总的代价函数为，为了求得代价函数J的最值，我们对其进行求导，，通过推导会发现代价函数极值点与我们所要给出的函数f形式有关。

**四、算法实现**

重写logistic回归，我所选择的函数形式为反比例函数，考虑数据的提取，由于我们的数据量比较多，内存限制无法一次将所有数据全部提取，因此我们采用耗时比较长的方法来进行数据提取，即通过股票代码提取同一组数据，对其进行运行后得到我们所要的估计值。然后再次提取下一组数据，这样减弱了数据量的限制，但是由于对数据库的操作次数较多，使得运行时间增长了很多,导致运行单次循环时间非常长。

考虑到个人电脑的运算能力以及多次读取数据库所带来的时间延迟，我们并不会使用数据库中的所有数据来进行拟合，我们选择随机从所有数据中选取多组数据来进行一个拟合。

由于时间消耗依然很长，因此我们改变了策略，即改为了一次性从提取少量样本的数据，我们选择了提取10组，20组数据来进行拟合，然后提取出来之后直接对数据进行操作，减少对数据库的操作次数，从而得到较快的响应。

首先选择10组数据来进行拟合，得到返回的θ=0.9807256736374502，正确率为60%，接下来选择20组数据来进行拟合，得到返回的θ=0.97751905447499183，正确率依然为60%，接下来选择50组数据来进行拟合，返回θ=0.9954719470712639，正确率降低为44%，得到的结果与我们的学习率与判断范围以及选取的实例相关性很大，但是基本可以确定一次反比例函数是相对比较好的拟合结果。

**五、结论**

通过计算，可以发现我们函数的预测准确率平均达到了50%以上，这个结果还是比较好的，说明我们的思想是比较符合实际的，而且可以观察到随着样本数的增加，其正确率也会降低，从实际意义上来说，这是符合实际的，因为大部分股票的变化其实是与现实相关更多，历史数据的影响最多只能规划其总体变化趋势，而不能预测每一天的变化，该算法其本质思想是通过历史数据来进行一个预测，而我们所知道的是，股票除了通过往期的数据来判断趋势之外，更多的依赖于现时的政策等其他因素的影响，因此为了而这些变量不容易数字化，因此可以说股票的准确预测预测是非常困难的，在某种意义上说，股市的规律是长期可预测，短期不可预测，而我们这个算法只是做到了一个简单的预测，只能提供参考，炒股需谨慎。