因子研究之机器学习基础. 第30讲

渔阳 2019-03-22 17:34

字数 5,586 阅读需 14分钟

“欢迎来到量化小学”

▲ 加入“量化小学”校友圈儿提问交流

下一篇:
进阶研究：集成学习和深度学习. 第31讲

详细内容请在wifi环境下观看视频

<本期课程5973字，视频21分58秒，请合理安排学习时间>

正文阅读

大家好，欢迎来到量化小学。

前面我们讲了因子研究的一些基本方法，包括线性回归、分组法等等。今天我们来谈一个有意思的话题，也是现在大家关注度非常高的一个领域——机器学习。

首先我们讲一讲，怎么样构建问题、以及在量化研究当中使用机器学习的一些要点，然后给大家介绍一些常用的基础方法。

机器学习的关键

首先我必须强调一点，机器学习并不是从石头缝里冒出来的，它是传统的统计研究方法的延伸。另外机器学习是非常强大的，但是也不必把它神化，它确实能够帮助我们更有效的提炼和利用数据中包含的信息，但是没有办法通过机器学习来凭空造出阿尔法。

我举一个例子，比如说我们可以把量化研究比喻成开采石油的过程，传统方法能够开采出一些油来，如果你掌握了页岩油的开采方法，你就可以采出更多的油。但如果这个地方根本就没有石油的话，你用什么样的方法也采不出来。

所以量化研究也是一样的，首先你要想明白你要研究什么样的问题，这个问题必须是可以被研究的，然后才谈得到用机器学习这样比较好的方法来使得研究效率提高，获取更多的阿尔法。

来自特辑

量化小学
解放你的投资动手能力

最近更新

【学业总结】量化学习的脉络梳理，以及继续学习提高的路径
2019-04-12更新

进阶研究：集成学习和深度学习. 第31讲
2019-03-28更新

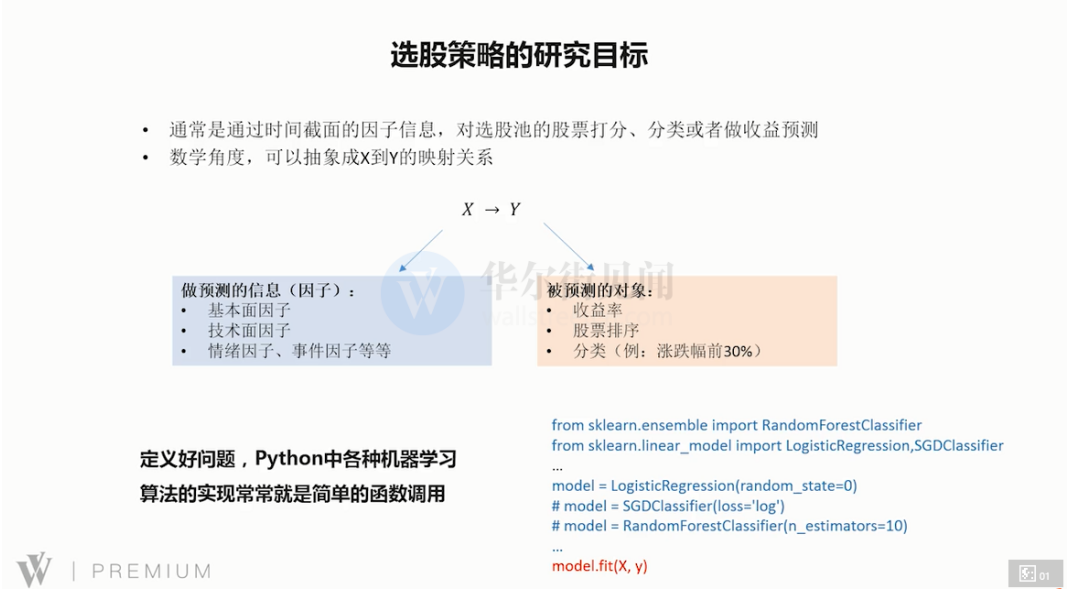
所以我们前面讲过几次了，选股策略通常就是通过一个时间截面的因子信息，对选股池的股票进行打分分类，或者来做收益预测。因此从数学的角度来说，我们可以把这种研究抽象成X到Y的一个映射关系。

X是什么？一般来说就是我们用来做预测的信息，通常也把它叫做因子。股票研究，我们可以有基本面因子、技术面因子、情绪因子、事件因子等等，前面都讲过不少了。

那么被预测的对象呢？你要研究什么呢？根据实际情况我们可以去直接预测收益率，也可以对选股池的股票当中进行排序，也可以来做分类。

例如说我们希望知道我们把涨得多的股票分成一类，把跌得多的股票分成一类，让计算机去研究一种算法，能够正确的对股票的未来收益进行分类。

那么当你成功的把这个问题构建好了，然后又准备好了，“X”——也就是能来做预测的信息，也知道、想明白了被预测对象到底什么时候之后，其实机器学习这件事本身就没有想象中的那么复杂。在python当中各种机器学习算法的实现常常就是简单的函数调用。



比如右面（见上图），这是一段非常简单的核心的样例代码，我们应该感谢这个伟大的时代、开源的时代，有很多的世界上很多的聪明人把这些机器学习的算法都给我们实现了，哭着喊着让我们用还不要钱，真的是就做梦也想不到的好事情！但是在现实世界当中真实就在发生，最好的软件往往都是不要钱的。

那么我们看到，Python当中有一个很基本的机器学习的包叫“sklearn”，在“sklearn”里面还有各种各样的模型都有实现。比如说在样例当中有三个机器学习常用的算法，我们马上会涉及到。

第一个是随机森林“RandomForestClassifier”做分类的，你还可以用逻辑回归或者SGD来做这个分类器。那么实践当中你只需要告诉Python你的模型，你准备用哪个分类器。如果是“LogisticRegression”，也就是逻辑回归，主要就是SGD或者是随机森林，也就是一行代码的事情。

反正不管用什么样的方法，总是在研究从X到Y的这样一个映射关系。所以最后一行当你在fit这个模型当中代码都是一样的，所以实践当中是挺方便的.

所以我再次强调,关键的问题还是你要想明白你到底要研究什么？你用哪些因子，然后你被预测的对象、也就是学习当中常说的标签到底是什么？

机器学习方法：正则回归

机器学习有哪些基础的方法呢？我们先来讲一下**正则回归，可以认为是线性回归的一种改进。**

那么传统的线性回归，我们先来看一看它有哪些问题。第一个就是可能导致过拟合，这也是我们反复强调的，是量化研究当中特别容易掉进去的一个坑。因为过拟合的模型是没有意义的，它可以对过去有一个很好的完美的解读，但是放到未来常常就没有什么用处。

在量化研究当中，什么样的情况就容易出现过拟合？比如在线性回归当中，如果你的自变量或者因子太多的情况下，就可能导致过你和。因为一共就那么若干年的数据，就那么几千只股票，引入足够多的变量可以解释一切事情，但这个是要不得的。

另外就是你做线性回归的时候，最后你看某些回归系数会变得很大，也就是说某些因子的权重太大了，因此你整个模型就会失衡。比如说可能在某一段时间以内市值因子起到非常大的作用，大票表现特别好，或者小票表现特别好。如果你不加控制的话，可能这一个因子就在你的模型当中占据的比例太大了，其他因子就变得没有什么用处。这样的模型也是不好的。

其实不止线性回归会有这样的问题，其他的方法，比如决策树、逻辑回归等等，也可能出现类似的问题，自变量太多了，或者是某些系数过大。

因此这一类的问题也有一些比较通用的改进方法，最早适用于线性回归的改进，比如我们可能听说过的叫“岭回归（Ridge regression）”，或者是“**套索回归（Lasso regression）**”，以及把两者综合在一起的叫“**弹性网络（Elasticnet）**”。

这些是干什么的呢？数学公式大家可以自己去看。在本课的学习材料当中，我们也会提供给大家券商的总结性的研报，写的都挺好的，也比较详实。

我反复强调，在量化研究当中，我们首先要理解他的思想，他到底要干什么？你弄明白了这一点，再去看数学公式，往往是事半功倍的效果。岭回归也好，Lasso回归也好，它到底在干什么？**其实就是在线性回归的基础上加入一些惩罚项。**

让我们回顾一下。线性回归是最小二乘法，它就是要最小化二阶的误差；那么岭回归和Lasso回归，就是在这个基础上又引入了对自变量个数和回归系数的惩罚项。你自变量系数多了，它就会惩罚，回归系数太大了也会惩罚。

因此模型就必须要试图去达到一个平衡，如果我引入一个新的自变量，对于这个模型的解释度只有一点点提高的话，就可能会得不偿失，因为它有惩罚项。同样回归系数太大了，也是一样的问题。因此这两种方法都能够有助于帮助防止过拟合。

我们刚才讲到了过，拟合的问题不仅仅出现在线性回归的方法里面，也可能出现在决策树等其他方法里面。因此类似于岭回归、套索回归这样用正则项来防止过拟合的方法也广泛地应用于机器学习的其他的领域。所以这几种方法老师讲是属于前机器学习时代，较老的方法，几十年前就有了。

现在实践当中往往会用一些更新的更好的方法，但是他们正则化的思想还是非常重要的。我们刚刚讲过，在其他的机器学习方法当中，你常常还会看到类似的手段。

基础防范：逻辑回归

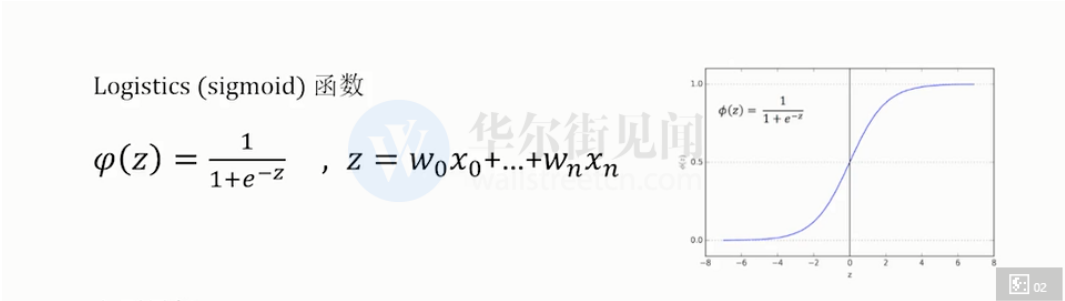
好，那么现在常用的方法是什么呢？

首先是逻辑回归。其实逻辑回归它是一个分类的方法，它要解决的问题，比如最经典的是二分类，就是一个股票到底是好股票还是坏股票，你给我一个直接的分类，或者说你给我一个概率。

那么逻辑回归是怎么做到这一点？我们先来看一下他的数学表达式。这个数学表达式里面是有两大块，左边的逻辑回归函数也叫sigmoid函数，它是一个指数化的方程。但是指数话e的肩膀上扛着一个Z，它倒是一个线性函数。

所以逻辑回归它有线性，所以也被称之为一种线性方法，因为这个Z本身是线性的，这跟传统的线性方法一样，它是有若干个因子，然后每个因子配一个权重，这样算出一个Z来。

那么为什么他要给你这样一个指数的形式，其实这就是做分类用的。我们先来看一下函数右边这张图（见下图），



logistics（sigmoid）函数

这是逻辑回归的函数的一个形状，我们就可以看到它在取值比较小的时候，是非常接近于0的；取值比较大的时候，是非常接近于1的，然后在中间、在0附近，随着Z的变化，这个函数快速就上去了。

所以这实际上是非常类似于“0,1”的一个电信号的概念，就是从0“啪”一下就变成1了。它这个变化的速度取决于这个Z的本身可以变得慢，也可以变得快。

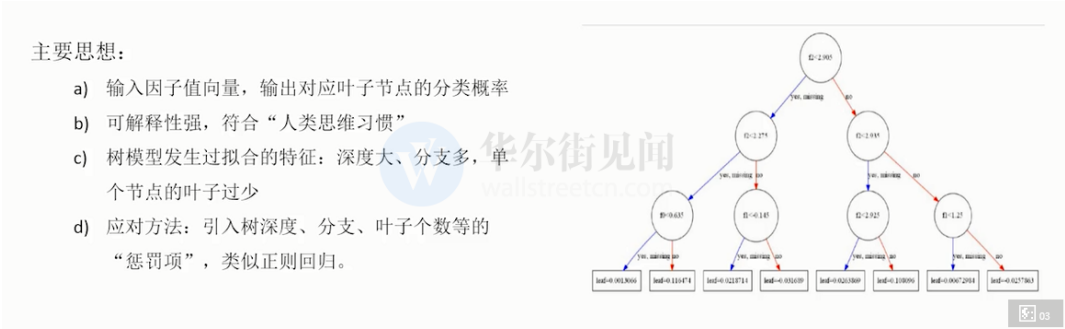
为什么要这样做呢？因为指数函数在数学上有一些良好的性质，它是连续的，也是可以求导数的，求几阶导数都可以。这个在机器学习后续求参的优化算法当中就更容易实施，你直接给他一个“0,1”的函数，数学上是无法处理的。

所以，这个逻辑回归sigmoid函数是在机器学习当中相当基础的一个数学函数。它可以用来做分类，给出学习标签的“0,1”的概率，其实也可以用来模拟一个电信号的刺激，这个在我们后续将会讲到的神经网络当中是有广泛的应用的。

所以逻辑回归是一个非常重要的**“基学习器”**。它**是更复杂的机器学习方法的一个基础零件**，下一次我们会讲到集成学习，大家可以看到集成方法的“基学习器”，就是逻辑回归或者是决策树这样的方法。

基础方法：决策树

那么第二个基础方法就是决策树。右面这张图（见下图）是实际上从quant<os>的因子研究当中直接截取的一张决策树的图，希望大家能够理解清楚。



决策树（Decision Tree）

这个数有三层，每一层是根据一个节点来作出判断，F0、F1、F2就是一些因子或者所谓的特征。比如第一层就是F2，如果他一个判断条件是不是小于2.9，如果“是”就往左边走，“不是”就往右边走。F2就是某一个量化的特征，比如说可以是PE值、可以是成长性等等。

决策术有什么好处呢？

它可解释性很强，非常符合人类的思维习惯，比如我们选股的时候也是按照一些类似的方式。比如我就先只看价值股，不是价值股我就不看了，在价值股里我又只选大股票，小的我不看，那么在大股票里我要选近期表现还不错的。所以根据这三个条件、一棵决策树我就选到了茅台、招商银行这种股票，人类就是这么决策的。

所以在机器学习当中，我们可以通过决策树来模拟这个过程，把历史数据输入给这个模型，看看他能够学习到一些什么样的逻辑。

比如刚才我说的先看价值、再看市值、最后看动量的一个简单的决策过程，如果说比较多的投资者都使用了这样的一个过程，反映在这些股票表现比较好，那么机器学习的方法就应该能够从实际的数据当中学到这个规律。

最后产生的模型呢，你只要输入因子值的向量，它就能够给出对应叶子节点的分类概率，到底是好股票的概率大，还是坏股票的概率大。

我们反复讲过，不管是传统模型还是机器学习模型，做的时候一定要注意防止过拟合，因为过拟合的模型是没有意义的，它可以完美地解释过去，但是却不能预测未来。

那么在树模型当中，我们要注意什么时候会发生过拟合？它有什么样的特征？比如树的深度太大、分支太多、单个节点的数据点数过少，都是过拟合的特征。简单来说，如果我们跟线性回归做一个对比的话，就是你这里面变量太多了、规则太多了，最后你就是过拟合了，你去完美的解释过去了。

比如说树的深度，你用一个10层的树来分析A股市场合适不合适呢？很可能就是不合适的，因为一共就那么两三千只股票，一个10层的数节点可以有上千了，你2000多只股票，最后一个叶子上就两三只股票，这就不是什么可以泛化的规律，就是一些非常具体的在凑数了，所以这么深的树来研究A股是不太合适的。

所以我们怎么应对呢？就是要引入一些惩罚项，这跟前面讲的正则回归是类似的。当树的深度分支和叶子节点数增加的时候，都是有惩罚的。

就好比买东西你要付钱一样，你必须要做到物有所值。如果我增加这一个分支，我能够获得足够多的新的解释能力来覆盖我这个惩罚的成本的话，我就增加，否则的话我就不增加，最后可以把树的复杂度控制在一个合理的范围以内。

所以和刚才我们讲的逻辑回归一样，决策树模型也是分类问题，重要的“基学习器”。在下一次我们将会讲到集成学习，可以看到，把“基学习器”组合起来，就会获得更强大的学习器。

我们讲了几类基础的方法，当然了还有一些其他的方法，比如说K-邻近(KNN)算法、SDM算法等等，大家都可以有兴趣的同学都可以自己去阅读课外的材料。

那么从实践的角度而言，刚才我们讲到的几种方法是用的比较多的，作为基学习器，也是更复杂方法的构建基石，所以我们重点讲了这些。

学习目标和方法的分类

最后我们来总结一下机器学习的目标和分类的方法。可初学的时候可能对机器学习的名词感到比较困惑，不知道他们干什么的，其实内在都是很有逻辑的。

监督学习

比如第一个你经常听说的词叫“监督学习（supervised learning）”。

这个什么意思呢？就是我会为算法设定具体的学习目标，我告诉你什么是对的，什么是错的，让你去学习这其中的规律。

比如说以选股策略为例，X是很多个因子，有截面期的因子值。那么我们设定的学习目标是什么呢？要对未来一期的股票超额收益率做分类。比如说我们可以定义什么叫好股票，前30%的就是好股票；什么是差股票？后30%的。

我整个的问题都定义好了，我把历史数据输入到这个模型当中，让它去学习这里面的规律。也就是每一期截面的因子值和后一期实际上哪些股票是好股票、哪些股票是坏股票，这样你数据足够多的时候，如果有规律的话，机器学习就能够学习出来。比如刚才我们讲到的，价值类的、有动量的大股票表现好的话，它是可以学习到这样的规律的。

那么最后我们怎么样来衡量监督学习的效果？一般就是使用跟正确率、召回率等等相关的指标，最常用的是AUC，大家自己可以去看。

无监督学习

除了监督学习以外，我们还常常听说一类机器学习叫做无监督学习。顾名思义就是你不预设学习目标，而是通过算法来挖掘数据中的内在规律。

比如说在金融领域，我们可以用它来做聚类分析，因为哪些股票有共性，是在看数据之前我们未必知道的，所以这是无监督学习。

我们前面讲过有一些传统的分类方法，按照行业按照市值按照风格进行分类，那么是不是在这个之外，或者说有更好的分类的方法，无监督学习和聚类分析就可以帮助我们解决这样的问题。

也许你就发现了有一类的，比如创业板的股票可能是一个整体，经常作为一个整体行动的，这就是一种数据中看到的聚类的行为；或者说是跟5G相关的股票，可能会不管他是什么行业的都会一起动，这也是一种聚类的无监督学习。

最后，在实践当中大家常常采取的是更新的方法，因为今天我们讲的这些基础方法，在实践当中解决问题的能力还是比较有限的，比传统的线性回归也未必能强多少。近些年来，特别是随着计算机技术的发展和计算能力的提高，更复杂的方法，对于数据、对于计算要求更高的方法纷纷的涌现，也被应用到量化研究的领域当中。

主要有集成学习，他的想法就是把多个学习系组合起来，三个臭皮匠顶个诸葛亮；还有深度学习，就是在大数据高计算能力支撑下的监督学习，他跟前面的监督学习是一个意思，

通常也是来解决分类问题或者是模式识别问题的。

最后大家还听说过强化学习，就是一种可以与环境交互的无监督学习。比如说文明的下围棋的AlphaGo，就是使用了深度学习，还使用了强化学习。从理论上来讲，强化学习在金融领域也是可以有应用场景的。当然从一个实践的角度而言，和环境交互，你奖惩的逻辑其实不是太容易设置，所以应该说强化学习还处在一个探索性的阶段。

那么今天我们简单的给大家介绍了机器学习的基础方法，和怎么样在量化研究的领域当中构建一个可以用机器学习来研究的问题。

下一次，我们就给大家来简要的介绍一下这些所谓更新的方法，也是在实践当中更有用的方法——集成学习和深度学习。

课后学习资料

最后，关于机器学习是一个很大的一个话题，我们也不可能在有限的一两讲当中来覆盖很多的细节。所以对量化真的有研究兴趣的同学，一定要仔细地阅读课后材料。

我在这也给大家推荐一篇民生证券写的《人工智能系列之一：机器学习量化投资实战指南》。其实它就是略微有点小百科的意思，把一些常见的方法都做了简要的介绍。希望大家要仔细的阅读一下。

好，今天的课程就到这里，下期再见。

-END-

加入“量化小学”的见识圈，关注动态

感谢您订阅本特辑，扫描下方二维码或点击圈子链接，即可加入专属见识圈子提问交流

量化小学

渔生

小学而大不遗，量化师生联谊会

感谢大家订阅《量化小学》，这里是学校见识社群，你可以随时提问、随时互动，我们一起投资，一起分享！

风险提示及免责条款

市场有风险，投资需谨慎。本文不构成个人投资建议，也未考虑到个别用户特殊的投资目标、财务状况或需要。用户应考虑本文中的任何意见、观点或结论是否符合其特定状况。据此投资，责任自负。

写评论

请发表您的评论

表情 图片 发布评论