机器学习, 量化交易, 特征工程

[量化学堂-机器学习]量化投资中的特征工程

iQuant 5月11日 #1

> 导语：近年来，国内量化投资迎来了发展的黄金期，但涉及机器学习的量化投资还比较少。机器学习领域的大神Andrew Ng(吴恩达)老师曾经说过机器学习很大程度上就是特征工程，因此本文主要介绍下特征工程在量化投资领域的应用。

1.特征工程是什么？

有这么一句话在业界广泛流传： **数据和特征决定了机器学习的上限**。那特征工程到底是什么呢？顾名思义，其本质是一项工程活动，目的是最大限度地从原始数据中提取特征以供算法和模型使用。简单理解为：特征工程是使用专业背景知识和技巧处理数据，使得特征能在机器学习算法上发挥更好的作用的过程。

特征工程在量化投资领域有非常适宜的土壤，首先金融市场拥有海量数据，数据比较规整;其次，金融市场量化研究员开发优异策略离不开专业背景知识、行业经验和数据处理技巧；最后，金融市场的投资收益、风险可以直接检验机器学习算法性能。

特征工程之所以重要是因为它直接决定了机器学习算法的性能，对于量化交易员策略开发也是如此，**特征工程的相关工作将直接决定策略的盈利能力。**

2.开发策略就是特征工程

特征工程是一项工程活动，和量化交易有什么关系呢？量化交易员开发策略的过程本质就是特征工程。我们以一个量化领域比较经典的双均线模型（也称金叉死叉模型）来解释，该模型的策略核心是当短期均线上穿长期均线时，形成金叉，买入股票，当短期均线下穿长期均线时，形成死叉，卖出股票。在金融市场上，双均线模型可以实现长期盈利，那么量化交易员开发双均线模型的择时策略为什么就是特征工程呢。我们不局限在双均线模型的交易规则这一个层面，而是上升到K线数据的另一个特征层面，对于每一根均线而言，我们可以计算一个短期移动平均值与长期均线移动平均值之差这个特征，定义如下：

[Math Processing Error]

因此本质上双均线模型就是基于*[Math Processing Error]*这个特征进行交易。这个特征的构建先是通过每日收盘价计算短期和长期移动平均值，然后再做减法获得，如果你是涉足金融市场刚第一天的人，你很可能不会联想到这个特征，但是如果你是长期待在金融市场上的人那么你拥有了投资交易经验和金融背景知识，因此你很可能会开发出基于该特征的策略——双均线模型，你无须关心其他数据，只需知道每个*[Math Processing Error]*线上的*[Math Processing Error]*特征值即可。因此，量化交易员开发策略就是特征工程。

量化交易员发现股票收益和股票的某些因子之间存在线性关系，因此在开发策略时，更多的是关注具有超额收益的这些因子，选择符合因子条件的股票本质上也是特征工程。

3.特征工程的重要性

数据工程项目往往严格遵循着RIRO(rubbish in, rubbish out) 的原则，所以我们经常说数据预处理是数据工程师或者数据科学家80%的工作，它保证了数据原材料的质量。 如何从成百上千个特征中发现其中哪些对结果最具影响，进而利用它们构建可靠的机器学习算法是特征选择工作的中心内容。

也有人曾这样描述特征工程：特征工程就是研究我们应该输入什么数据。我们可以把量化交易员开发策略获取收益的过程看成以下映射：

[Math Processing Error]
其中*[Math Processing Error]*为输入的数据，*[Math Processing Error]*为策略，*[Math Processing*

5月11日

1 / 1
5月12日

5月11日

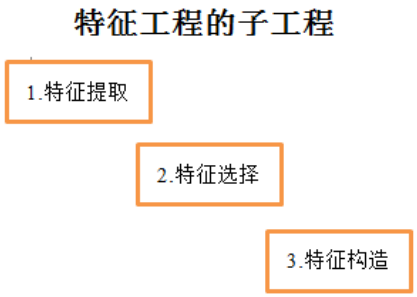
*Error]*为输出的策略收益。

金融市场上每天产生海量的数据，比如交易数据、行业数据、企业财务数据、宏观经济数据等，这些原始数据可以形成天量的特征，如何从这些特征中发现能够产生策略超额收益的好的特征对于量化交易员至关重要。

如果市场符合有效市场假说，那么量化交易员只需关心交易数据中的价格和成交量就可以，但是大多数的市场都不是有效市场。因此，量化交易员必须设计好选择什么作为输入。如果做不好特征工程，输入的数据有问题，那么输出的策略收益也不会高。因此在开发策略过程中，特征工程非常重要。

4.如何做好特征工程

要做好特征工程主要是解决以下几个特征工程子问题。（如图1）



[Math Processing Error]

4.1 特征提取

在数据挖掘领域，特征提取是将原始特征转换为一组具有明显物理意义（Gabor、几何特征[角点、不变量]、纹理[LBP HOG]）或者统计意义或核的特征。比如通过变换特征取值来减少原始数据中某个特征的取值个数等。对于表格数据，可以在设计的特征矩阵上使用主要成分分析（Principal Component Analysis，PCA)来进行特征提取从而创建新的特征。对于图像数据，可能还包括了线或边缘检测。常用的特征提取的方法有：主成分分析（PCA）和线性判别分析（LDA）。

金融领域也是如此。特征提取的对象是原始数据（raw data），它的目的是从原始数据中提取特征，比如我们获取股票的行情数据，行情数据里包含开盘价、最高价、最低价、收盘价、复权因子，我们不能直接使用这些股票价格作为特征，因为公司可能会有分红、派息等行为，因此股票价格不能反映真实的股价，所以要对其进行复权处理，进行处理以后，得到复权后的价格数据可以提取成新的特征了。

4.2 特征选择

当数据预处理完成后，我们需要选择有意义的特征输入机器学习的算法和模型进行训练。通常来说，从两个方面考虑来选择特征：

- 特征是否发散：如果一个特征不发散，例如方差接近于0，也就是说样本在这个特征上基本上没有差异，这个特征对于样本的区分并没有什么用。比如我们选取股票ST状态（ST:0,非ST:1）这个特征，这个特征较长时间会保持不变，因此该特征不发散，这样的特征我们尽量不选取。

- 特征与目标的相关性：这点比较显见，与目标相关性高的特征，应当优选选择。在有效市场或强有效市场中，类似于成交价格、成交量这类行情特征能够很充分地反映股票的大部分信息，因此这类特征应该优先选择。另外，量化交易员开发策略时，特征的选择与策略的模式也高度相关，比如，希望开发一个策略能够挖掘“长期低量盘整，价格突然持续拉高”的股票，如果这样的选股模式在金融市场上是可以盈利的，那么在选择特征的时候，应该选择长期平均成交量与短期成交量之比、长期移动平均值与短期移动平均值之比这类特征。这两个特征能够将具有“长期低量盘整，价格突然持续拉高”模式的股票选择出来。

4.3 特征构造

有时，原始数据集的特征具有必要的信息，但其形式不适合数据挖掘算法，在这种情况下， 由原特征构造的新特征可能比原特征更有用。

我们举一个例子，考虑一个包含人工制品信息的历史数据集，该数据集包含每个人工制品的体积和质量，以及其他信息。假设这些人工制品使用少量材料（木材、陶土、铜、黄金）制造，并且我们希望根据制造材料对它们分类。在此情况下，由质量和体积特征构造的密度特征（即密度=质量/体积）可以直接地产生准确的分类。尽管有些人试图通过考察已有特征的简单的数学组合来自动地进行特征构造，但是最常见的方法还是使用专家意见构造特征。

在金融领域也是如此，比如我们想区分股票价格的波动性，我们可以构造一个收盘价标准差的一个特征，这个特征能够反映顾及近期的波动情况；此外，我们也可以构造一个平均振幅的一个特征，该特征是每日最高价减每日最低价的差值的平均值，从数值的角度反映股票价格的近期波情况。

如果你是专业的量化交易员，那么面临众多的特征，你可以根据你的行业经验和投资心得在浩瀚的特征海洋里构造新的特征来开发策略。

> 小结：特征工程已经是很古老很常见的话题，引用几句大师的一些原话吧。
> “Coming up with features is difficult, time-consuming, requires expert knowledge. "Applied machine learning" is basically feature engineering.”— *Andrew Ng*
> 直译：构思特征很困难，很花时间，需要专家经验。机器学习的应用很大程度就是特征工程
> “some machine learning projects succeed and some fail. What makes the difference? Easily the
> most important factor is the features used.” —*Pedro Domingos*
> 直译：机器学习项目有些成功了，有些失败了，主要因为特征使用不一样
> “Actually the success of all machine learning algorithms depend on how you present the data.” — *Mohammad Pezeshki*
> 直译：事实上所有机器学习算法上面的成功都在于你怎么样去展示这些数据

参考资料

- **Discover Feature Engineering, How to Engineer Features and How to Get Good at It**
- **Feature selection**
- **Feature learning**
- **使用sklearn进行数据挖掘**
- **机器学习中，有哪些特征选择的工程方法？**
- 数据挖掘导论（完整版）.Introduction.To.Data.Mining.（美）Pang－Ning_Tan.范明译.人民邮电.2011

BigQuant AI策略详解

社区干货与精选整理（持续更新中...）

特征选择的实践重要性