深度学习新进展：Alpha因子的再挖掘

安宁宁 S0260512020003
邮箱：anningning@gf.com.cn
广发证券金融工程

2017年6月

01

02

03

04

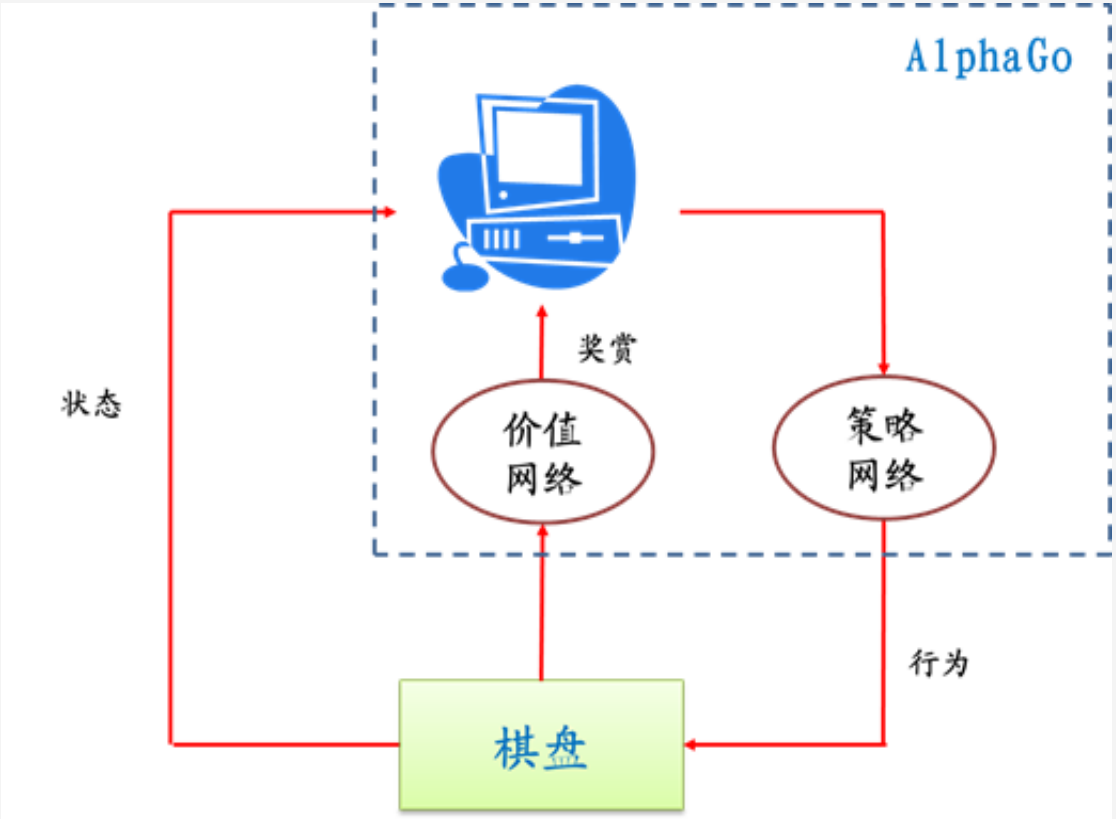
01

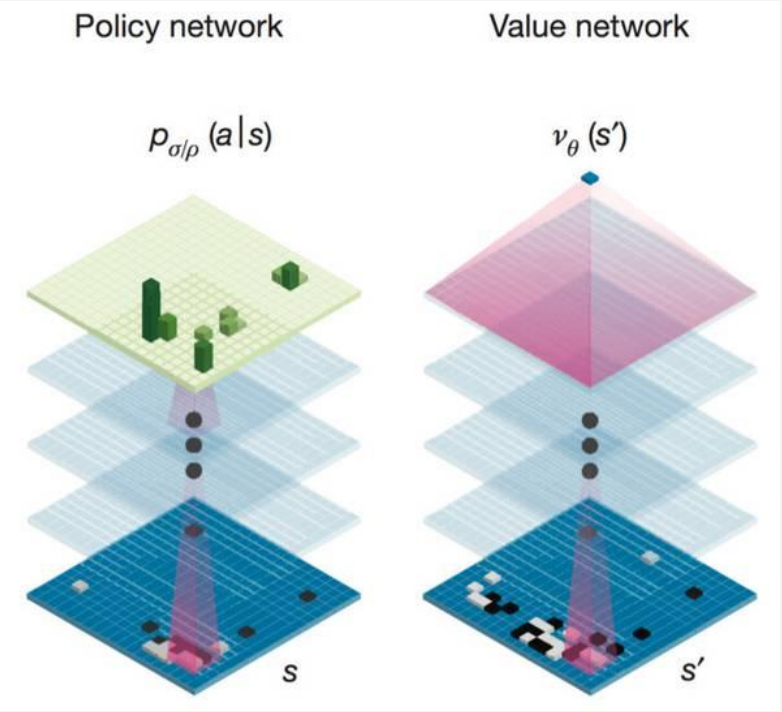
| 背景 |

>

- 图像和和语音识别
- 自然语言处理
- 医疗诊断
- 推荐系统
- 深度增强学习

近年来，深度学习应用上取得了重大的进展。





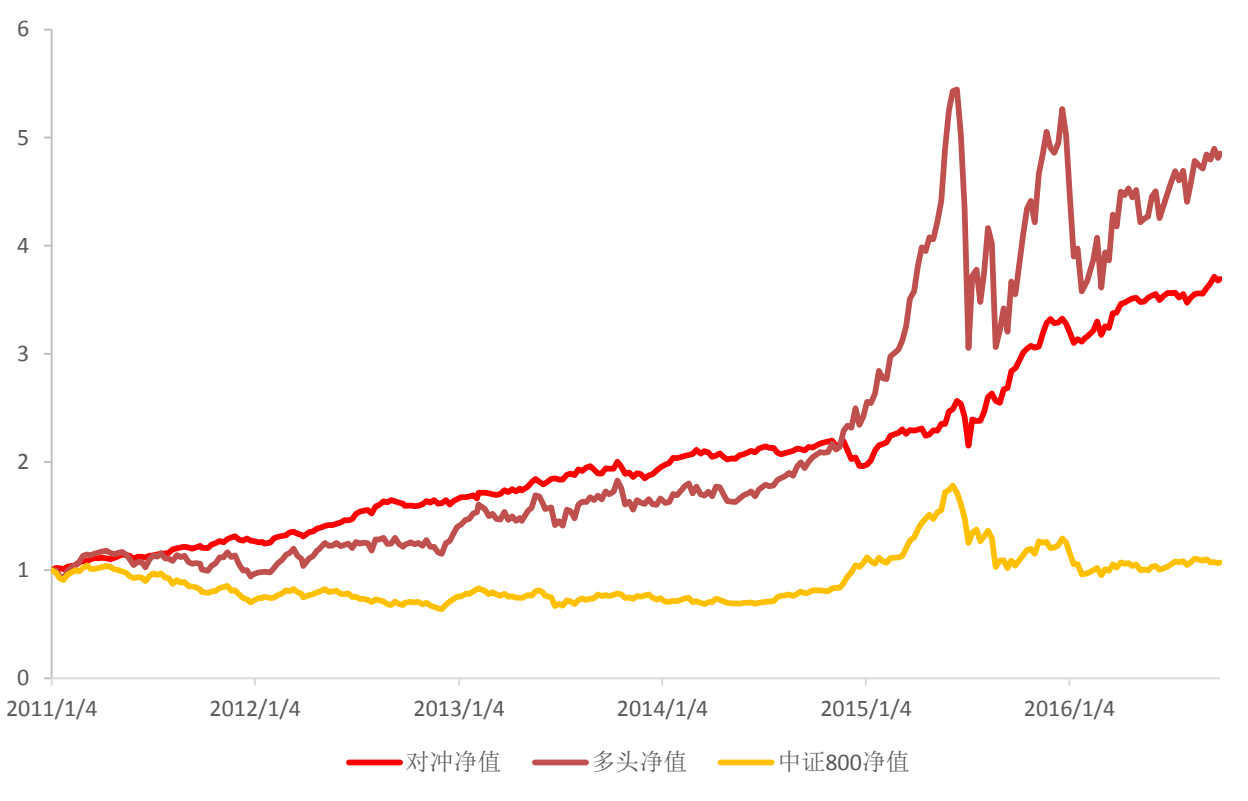
AlphaGo的核心就是深度增强学习。

对冲基金和投行开始在人工智能方向布局：

- **高盛**：Kensho系统
- **Bridgewater** ：2013年以来开始建立人工智能团队
- **Citadel**：聘请微软的人工智能首席科学家邓力为首席人工智能官
- 其他知名对冲基金如**Renaissance Technologies**和**Two Sigma**都在扩充自己的人工智能团队
- **Rebellion Research**：这是一家人工智能交易机构。它的交易系统通过机器学习，在全球44个国家的股票、债券、大宗商品和外汇上进行交易。类似的还有香港的**Aidyia**，旧金山的**Sentient Technologies**，伦敦的**Castilium**和**CommEq**，日本的**Alpaca**等机构

广发金工团队在2014年发表了两篇深度学习报告：

- 中证800选股策略从2011年以来，年化收益率为25.5%，夏普比率为1.75



数据来源：Wind，广发证券发展研究中心

新的思考

1. 调仓频率：周频修改为月频，降低换手和交易费用的损耗
2. 特征提取：在机器学习中融入金融行业知识，从选股因子中再挖掘
3. 深度学习技术更新：采用性能更好的模型结构

01

02

03

04

02

深度学习的进展

>

输入X

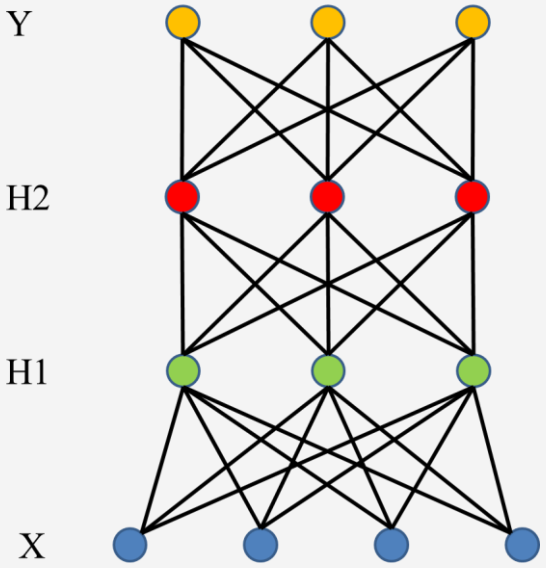
第1层隐含层H1

第2层隐含层H2

⋮

第n层隐含层Hn

输出Y

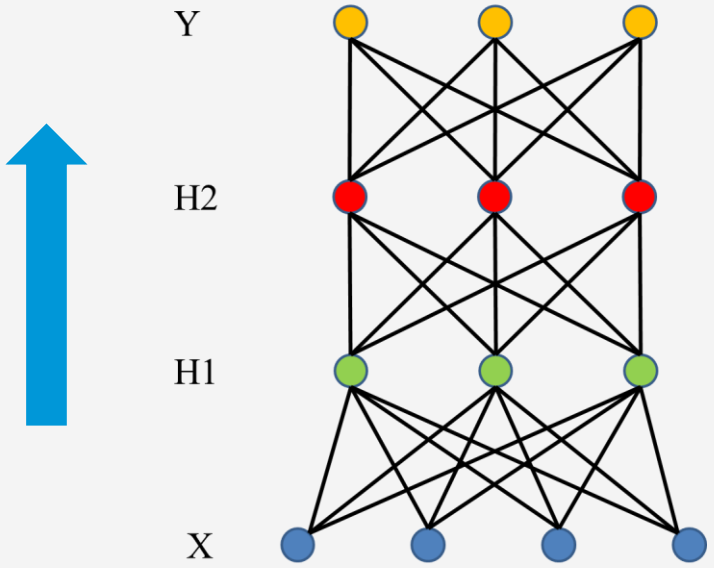


隐含层层数n≥2

多层的隐层结构具有更好的“表达能力”

更丰富的“表达”有望挖掘出大数据中更多的知识

问题：深层模型训练困难



$$y_k = \sigma_o\left\{\sum_{j=1}^{N_2}(w_{kj}^{(2)}h_j^{(2)} + w_{k0}^{(2)})\right\}$$

$$h_j^{(2)} = \sigma_h\left\{\sum_{i=1}^{N_1}(w_{ji}^{(1)}h_i^{(1)} + w_{j0}^{(1)})\right\}$$

$$h_j^{(1)} = \sigma_h\left\{\sum_{i=1}^{N_x}(w_{ji}^{(0)}x_i + w_{j0}^{(0)})\right\}$$

隐层激活函数： σ_h

输出层激活函数：σ_o ，对于分类问题，一般用Sigmoid函数或者Softmax函数

二分类问题 $\sigma(\theta^T x) = \dfrac{e^{\theta^T x}}{1+e^{\theta^T x}}$

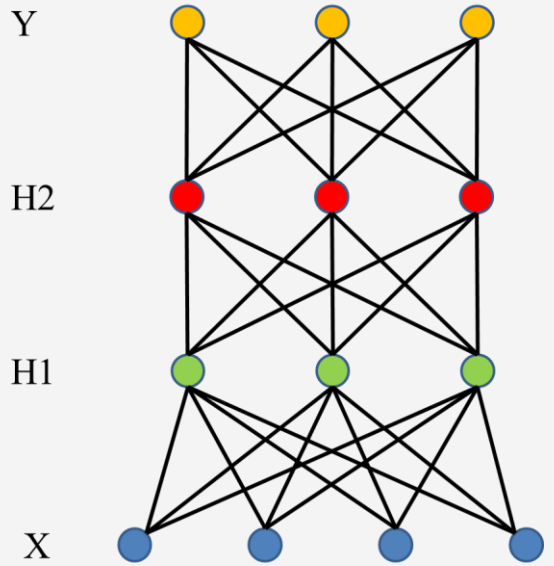
多分类问题 $\sigma(\theta_i^T x) = \dfrac{e^{\theta_i^T x}}{\sum e^{\theta_i^T x}}$

待优化网络参数：**w**

优化目标：最小化均方误差（MSE）

$$E(\mathbf{w}) = \frac{1}{NK}\sum_{n=1}^{N}\sum_{k=1}^{K}\left(y_{nk} - t_{nk}\right)^2$$

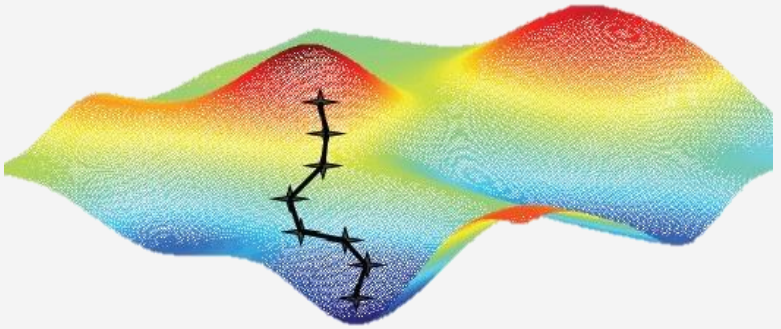
y_{nk}：预测输出　t_{nk}：实际标签

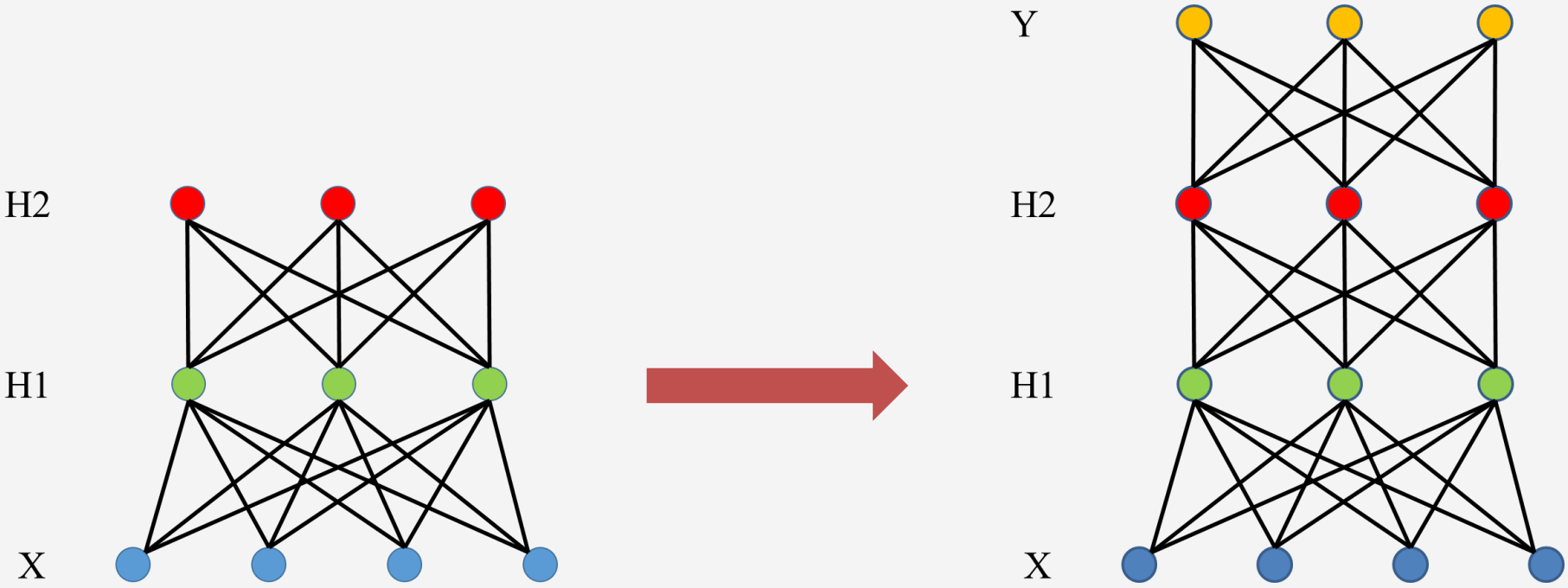


参数优化方法：BP算法，迷你批量梯度下降算法

- 集成了梯度下降法和随机梯度下降法的特点

$$w_{ij}^{(n)} := w_{ij}^{(n-1)} - \alpha'' \frac{\partial}{\partial w_{ij}} \sum_{n_k \in Batch(n)} E_{n_k}(\mathbf{w}^{(n-1)})$$





深层神经网络的训练方法：无监督学习预训练 => 监督学习模型训练

随着深度学习技术的发展，在样本足够多的情况下，通过选取更好的激活函数，可以不进行无监督学习，直接训练网络。

用于提高深层神经网络性能的方法主要有：

1、调整激活函数

2、调整优化目标函数

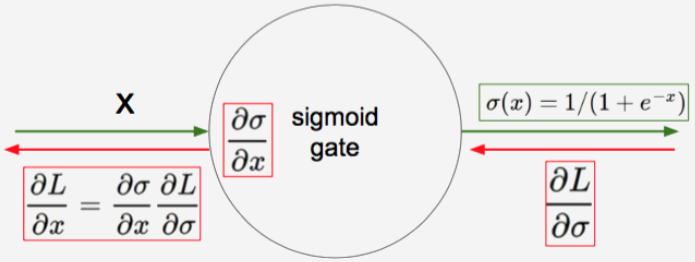
3、Batch normalization技术

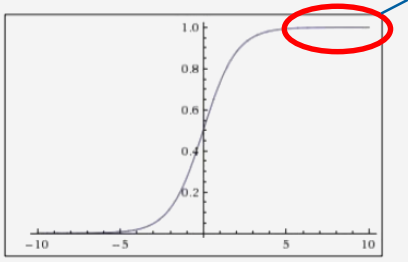
4、Dropout技术

5、优化网络结构

……

激活函数：非线性激活函数使得神经网络具有捕捉非线性特征的能力

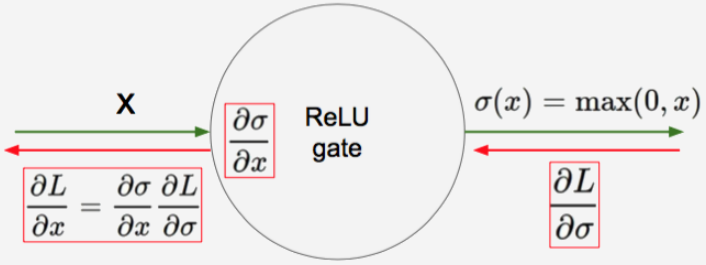


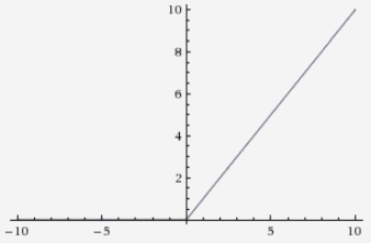


导数接近0

- 梯度弥散
- 指数函数计算复杂
- 输出值恒正，不是以0为中心

Sigmoid: 主流激活函数





- 大于0时导数恒定
- 计算十分简单高效
- 收敛速度大大加快
- 输出值恒正

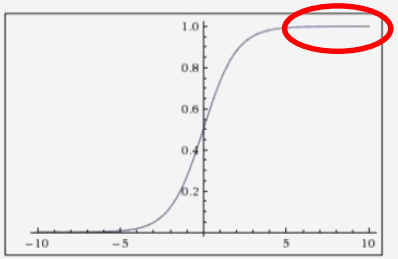
ReLU及其变种: 目前流行

目标函数的改进

优化目标：最小化均方误差（MSE）

$$E(\mathbf{w}) = \frac{1}{NK}\sum_{n=1}^{N}\sum_{k=1}^{K}\left(y_{nk} - t_{nk}\right)^2$$

预测输出　实际标签

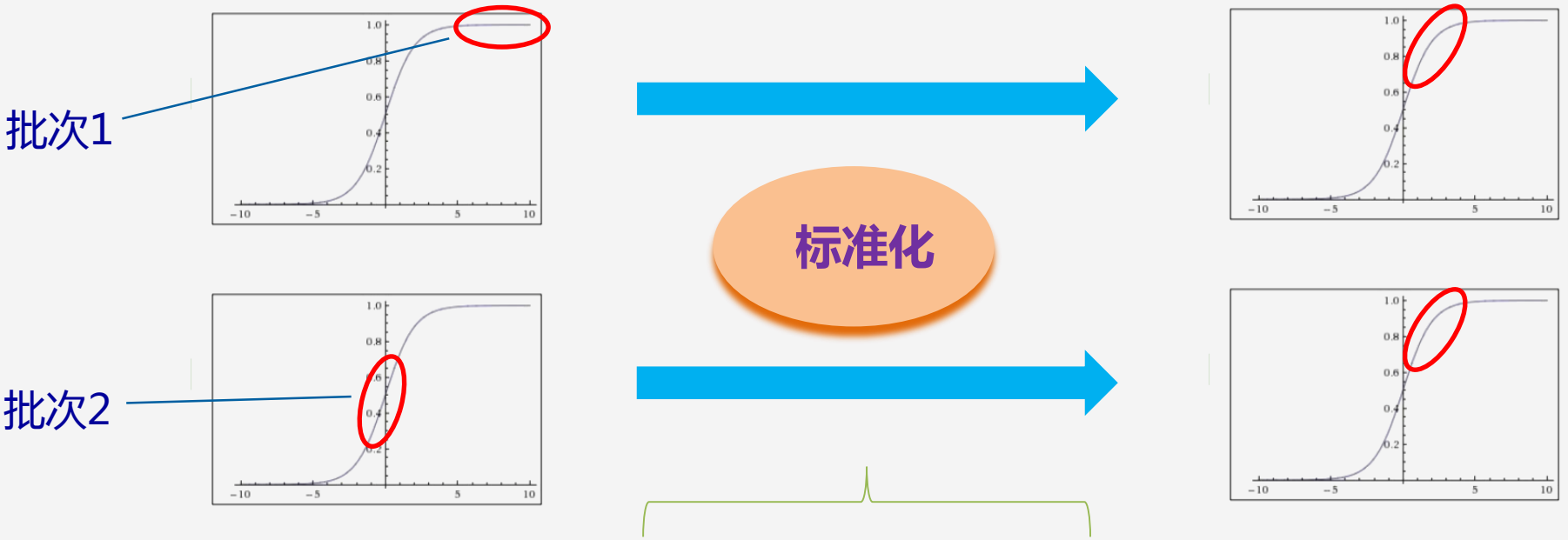


对于用Sigmoid或者Softmax激活函数的分类问题：交叉熵（Cross Entropy）更合适

$$J(\mathbf{w}) = -\frac{1}{NK}\sum_{n=1}^{N}\sum_{k=1}^{K}\left(t_{nk}\ln y_{nk} + (1-t_{nk})\ln(1-y_{nk})\right)$$

注：因为E(**w**)对**w**求梯度的时候，梯度值正比于Sigmoid函数的导数，Sigmoid导数可能接近于0，导致梯度为0

Batch Normalization: 使不同批次样本在神经网络隐层的输入保持相似的分布



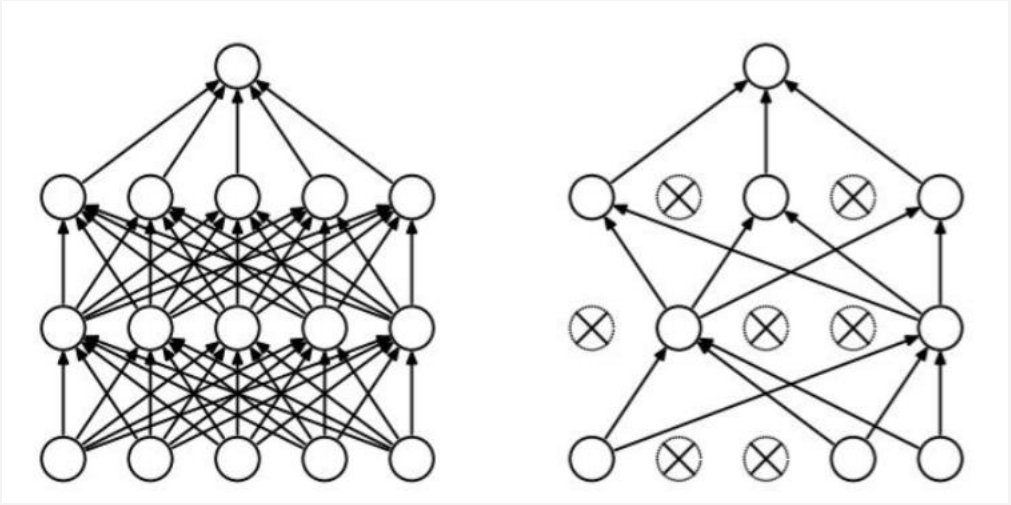
$$\hat{x}^{(k)} = \frac{x^{(k)} - \mathrm{E}[x^{(k)}]}{\sqrt{\mathrm{Var}[x^{(k)}]}}$$

对每个批次的样本分别执行

$$y^{(k)} = \gamma^{(k)}\hat{x}^{(k)} + \beta^{(k)}$$

Beta和gamma用于调整 x 的分布

Dropout: 减少过拟合

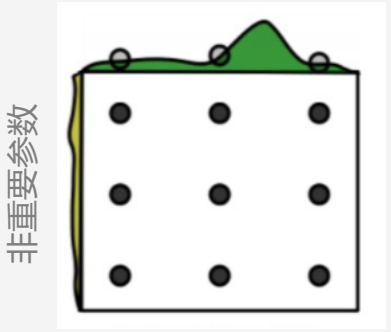


标准神经网络　　使用Dropout

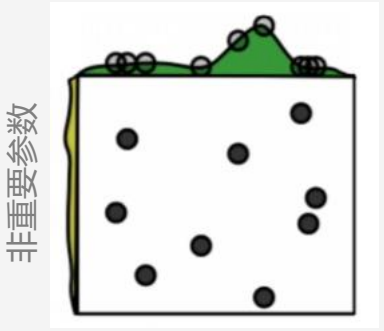
随机地选择一部分神经元，设定其输出为0

对每次输入的数据，对应的网络结构都不相同，可以理解成多个模型的集成学习，能够抵消一些相反的拟合

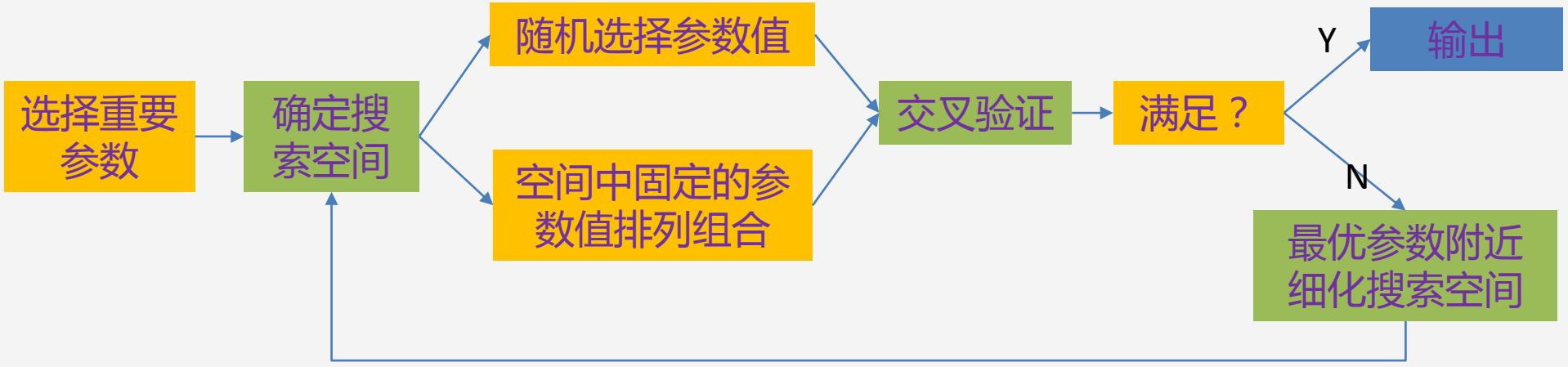
通过网格搜索或者随机搜索，对神经网络的模型结构（层数、节点个数、……）进行选择



网格搜索



随机搜索



01

02

03

04

03

策略与实证分析

>

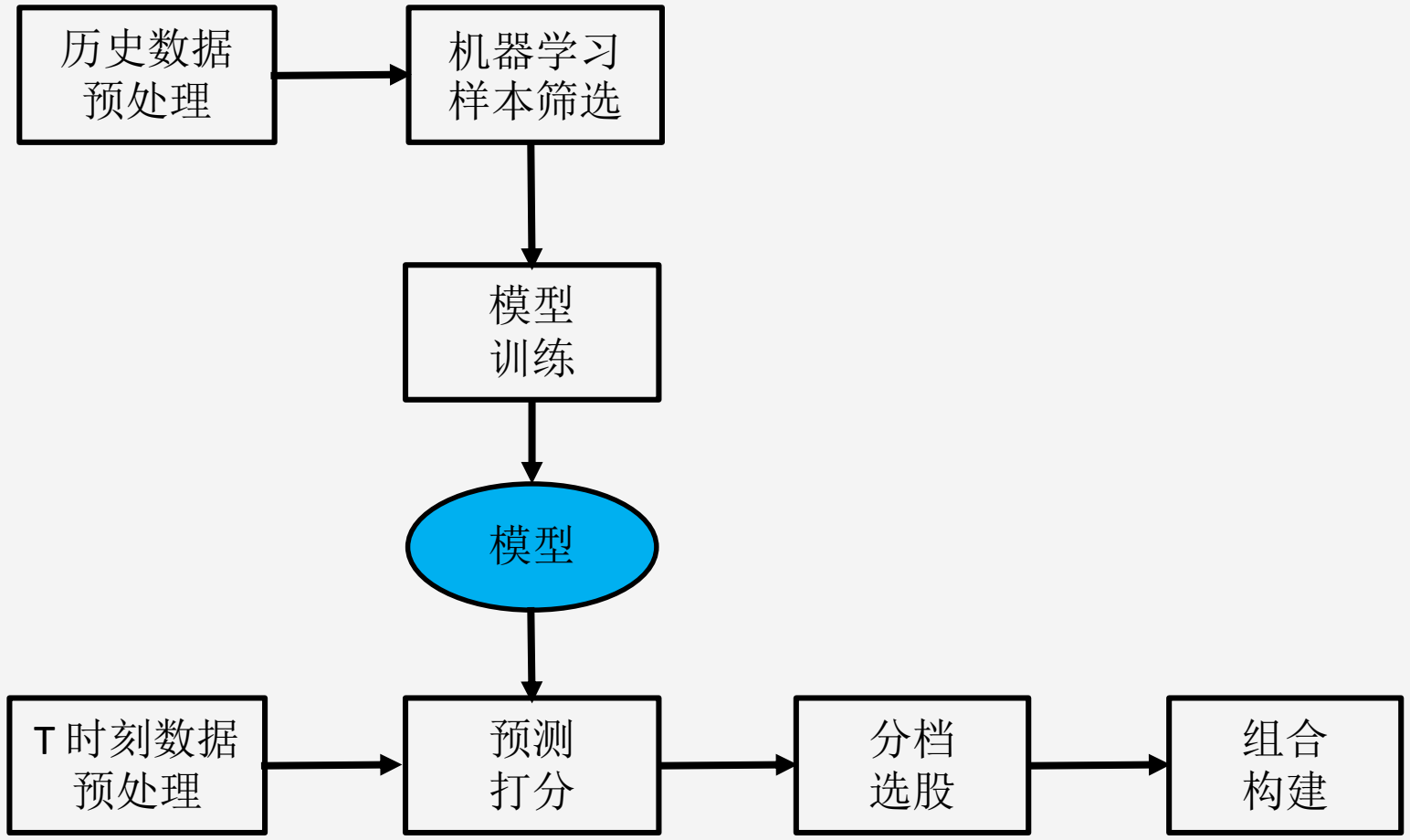
回测参数设置

- **调仓周期**：20个交易日
- **股票池**：全市场选股，剔除上市交易时间不满一年的股票，剔除ST股票，剔除交易日停牌和涨停、跌停的股票
- **超配组合**：调仓时分十档，等权买入深度学习模型打分最高的一档
- **对冲基准**：中证500指数
- **原始因子数据**：估值因子、规模因子、反转因子、流动性因子、波动性因子、技术指标，共计128个因子
- **深度学习模型训练期**：2007年1月-2010年12月
- **策略回测**：2011年1月-2017年4月
- **交易成本**：千分之三

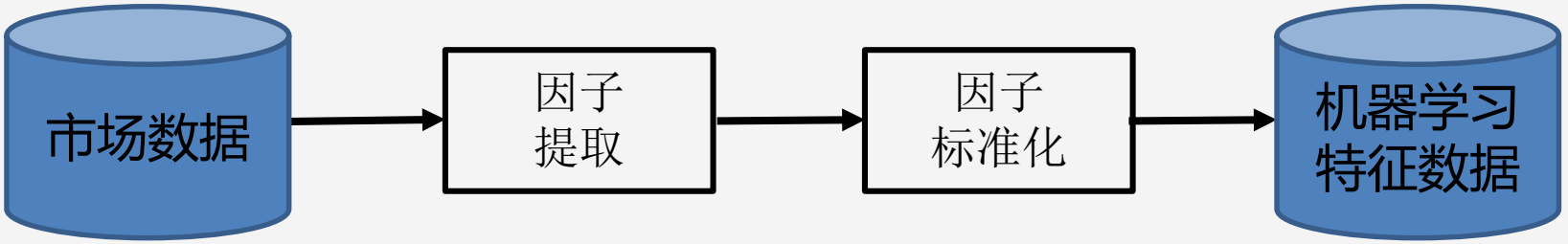
策略流程

模型训练

选股交易



数据预处理

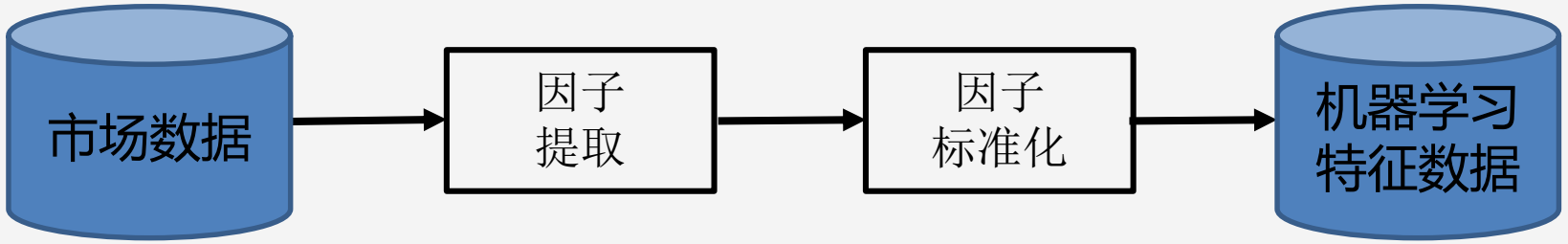


因子提取：从Wind终端提取数据，计算因子

一般而言，结合专业领域知识，提取合适的特征，有利于提高机器学习模型的性能

- 常用的选股因子，如规模因子、反转因子、PB、PE等本身就是比较有效的选股因子
- 技术指标选股也可以获取一定的超额收益（参见广发金工报告《Alpha因子何处寻 掘金海量技术指标》）

数据预处理



因子标准化：

1、异常值、缺失值处理

2、极值压边界处理

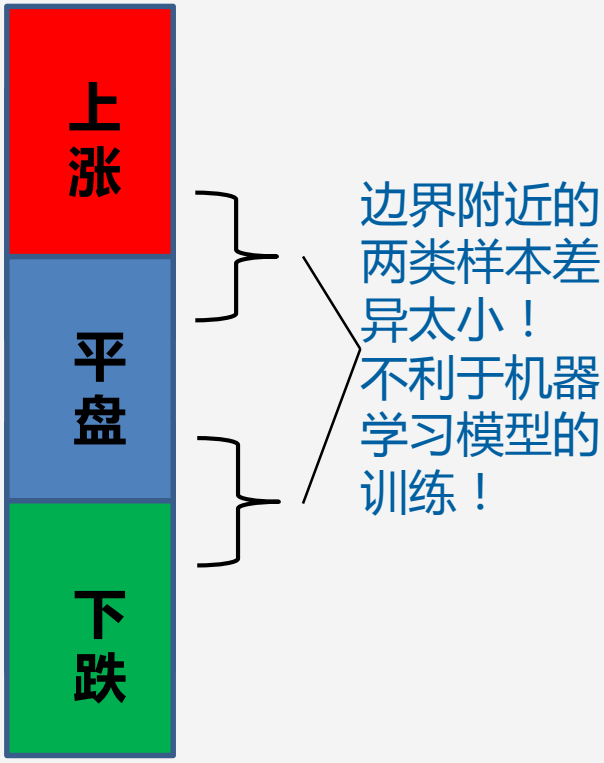
3、沿时间方向的因子标准化

4、沿截面的因子标准化

5、按照机器学习模型来调整因子分布（应该服从正态分布，还是[0,1]区间分布？）

样本筛选：减少噪声影响（仅用于训练机器学习模型）

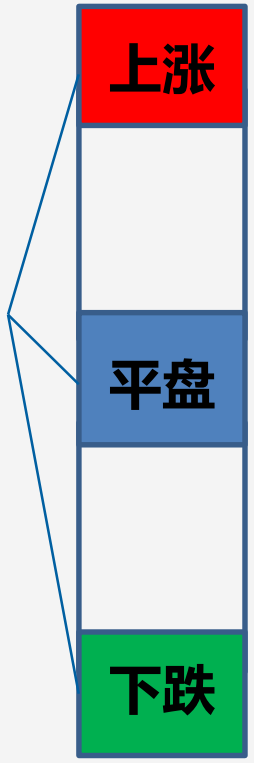
不筛选样本



按涨跌幅排序

边界附近的两类样本差异太小！不利于机器学习模型的训练！

筛选样本



不同类别样本差异大！

训练时，根据未来20日后的股票涨跌幅来给样本贴“标签”：上涨、下跌、平盘

同一时刻，按照涨跌幅进行排序，确定样本的输出标签

模型训练：通过训练样本，确定模型结构，优化模型参数

预测输出 Y 的维度：3

输入特征 X 的维度：156（128个因子+28个行业）

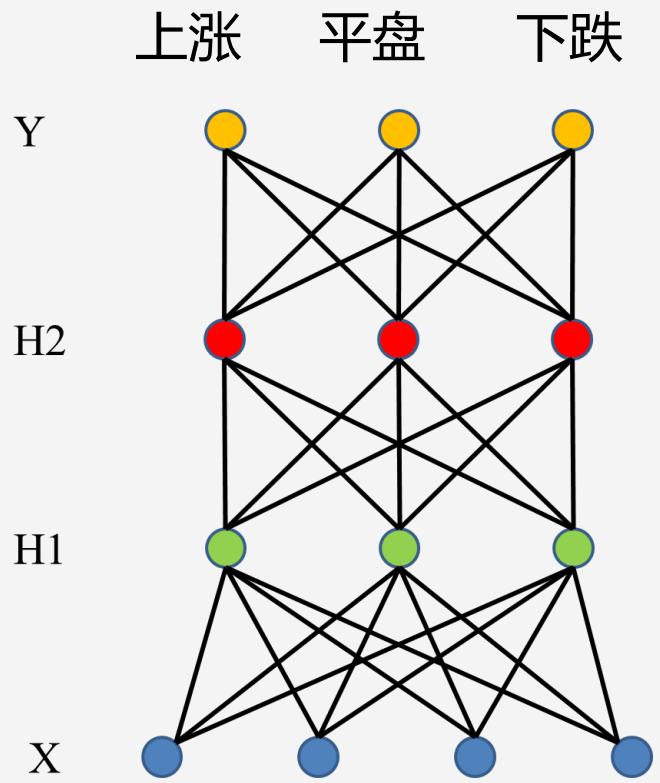
通过网格搜索获取最优的模型结构

选取模型结构为：

156（输入层）-512-200-200-200-128-3（输出层）

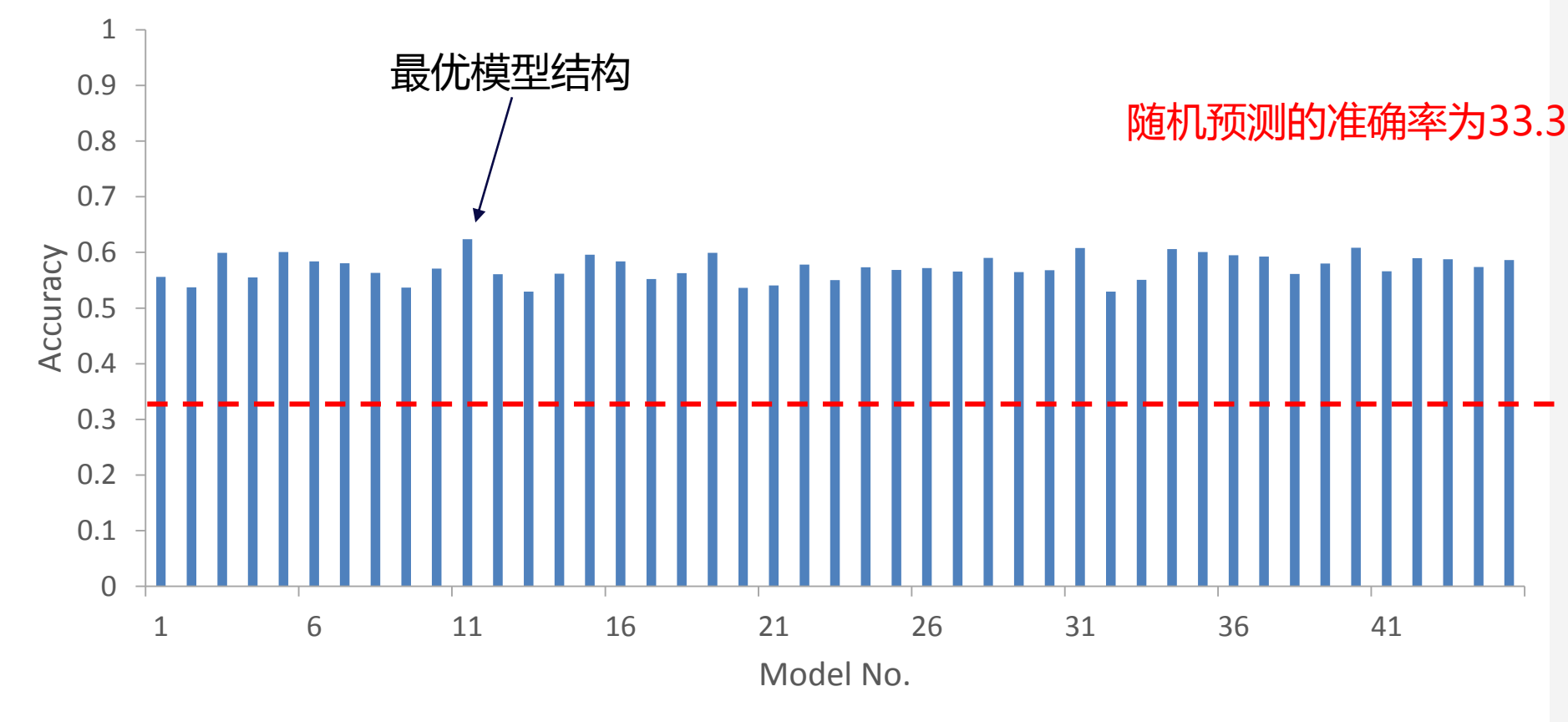
即一共包含5个隐层

隐层节点数依次为：512（隐层1）、 200（隐层2）、200（隐层3）、 200（隐层4）、 128（隐层5）



模型训练：

通过网格搜索获取最优的模型结构（5个隐层的结构下，不同模型的预测准确率基本上都在50%以上）



数据来源：Wind，广发证券发展研究中心

模型性能

训练样本数量：37万　　验证集样本数量：5万

训练集预测准确率 = 67.84%　　**验证集预测准确率 = 62.32%**

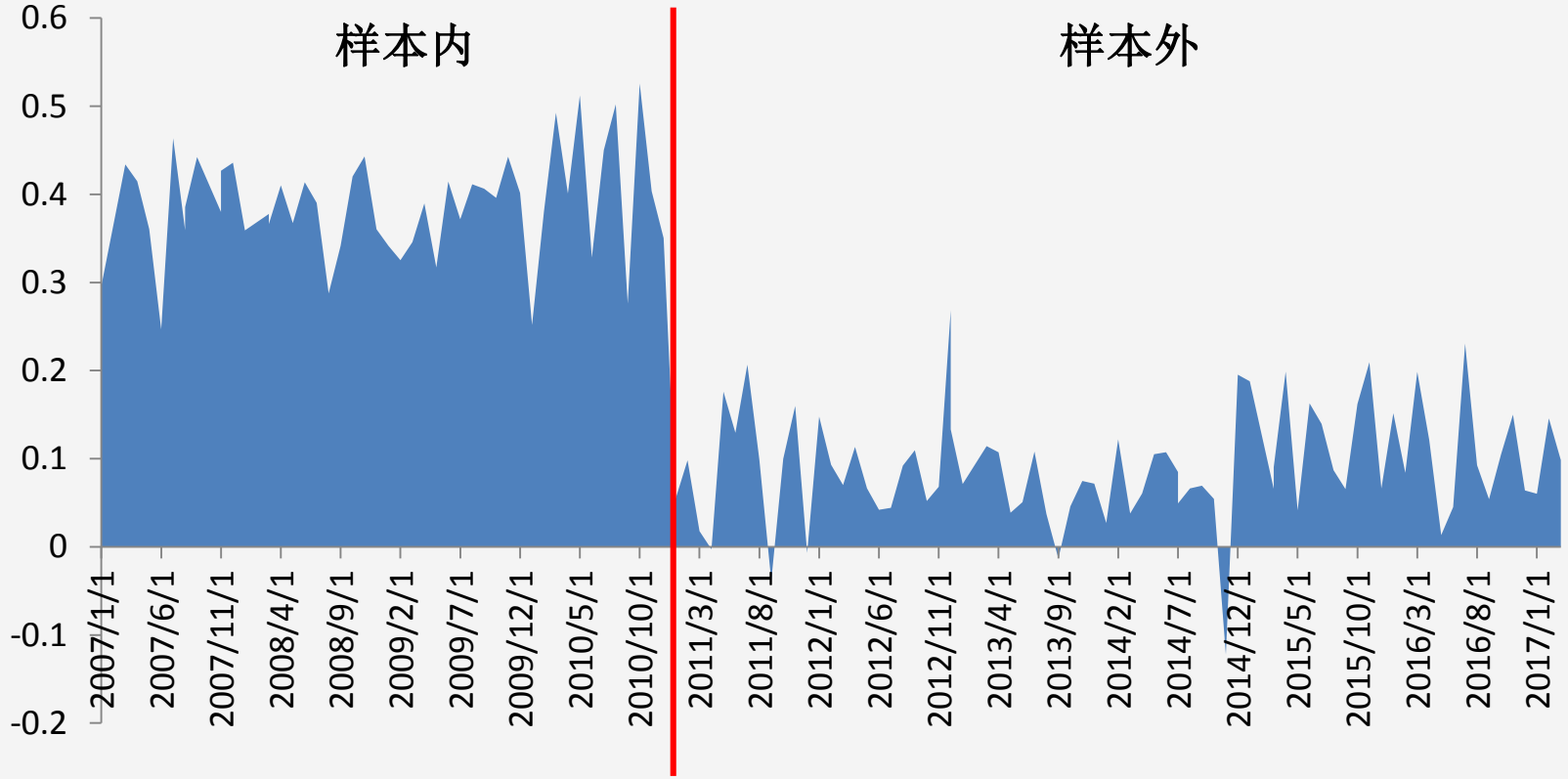
		预测			
		上涨	平盘	下跌	合计
实际	上涨	12403	3087	1182	16672
	平盘	6052	9103	2248	17403
	下跌	3762	2535	9700	15997
	合计	22217	14725	13130	50072

预测上涨的样本中，有55.8%属于上涨一类，仅有16.9%下跌。

数据来源：Wind，广发证券发展研究中心

因子预测能力：IC值

样本外 IC 的平均值为 0.092， 标准差为 0.065

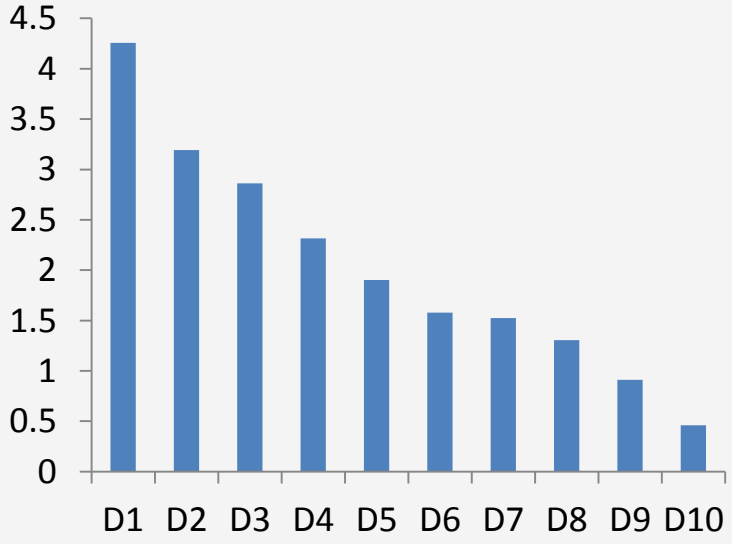


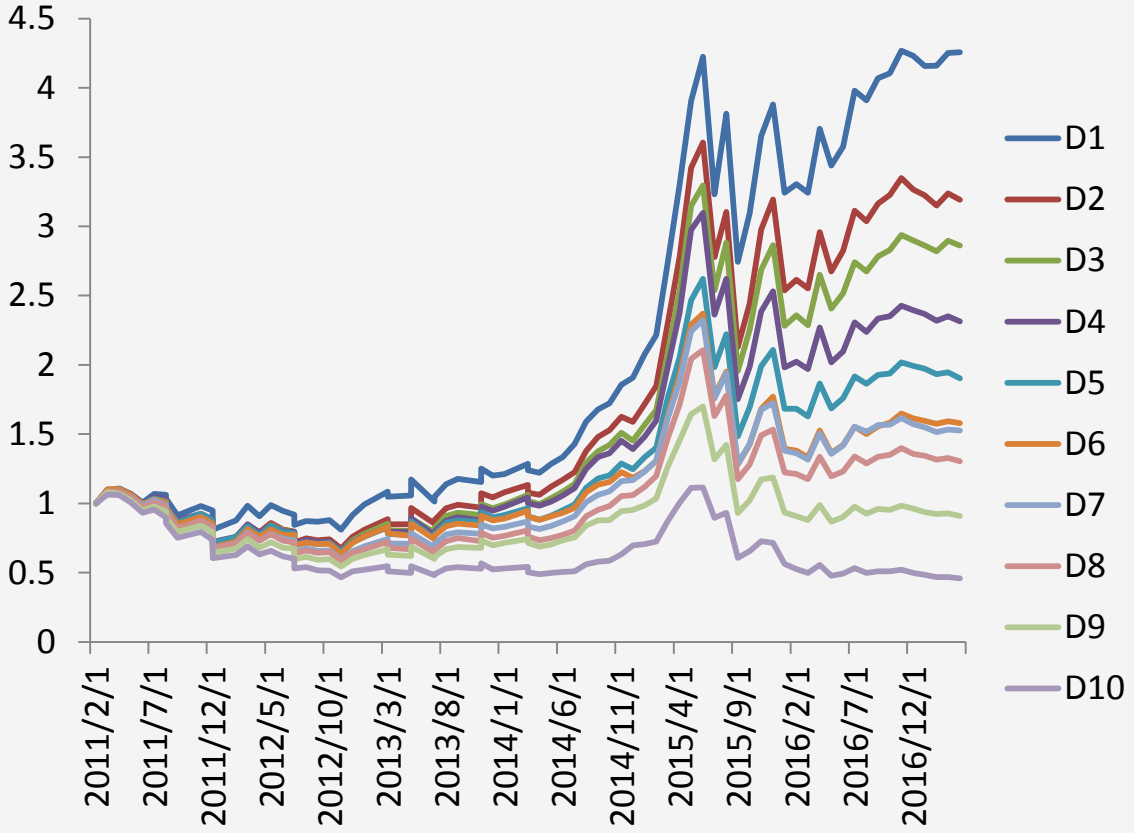
数据来源：Wind，广发证券发展研究中心

因子分档表现

因子表现的单调性好

深度学习打分越高，累积收益率越高

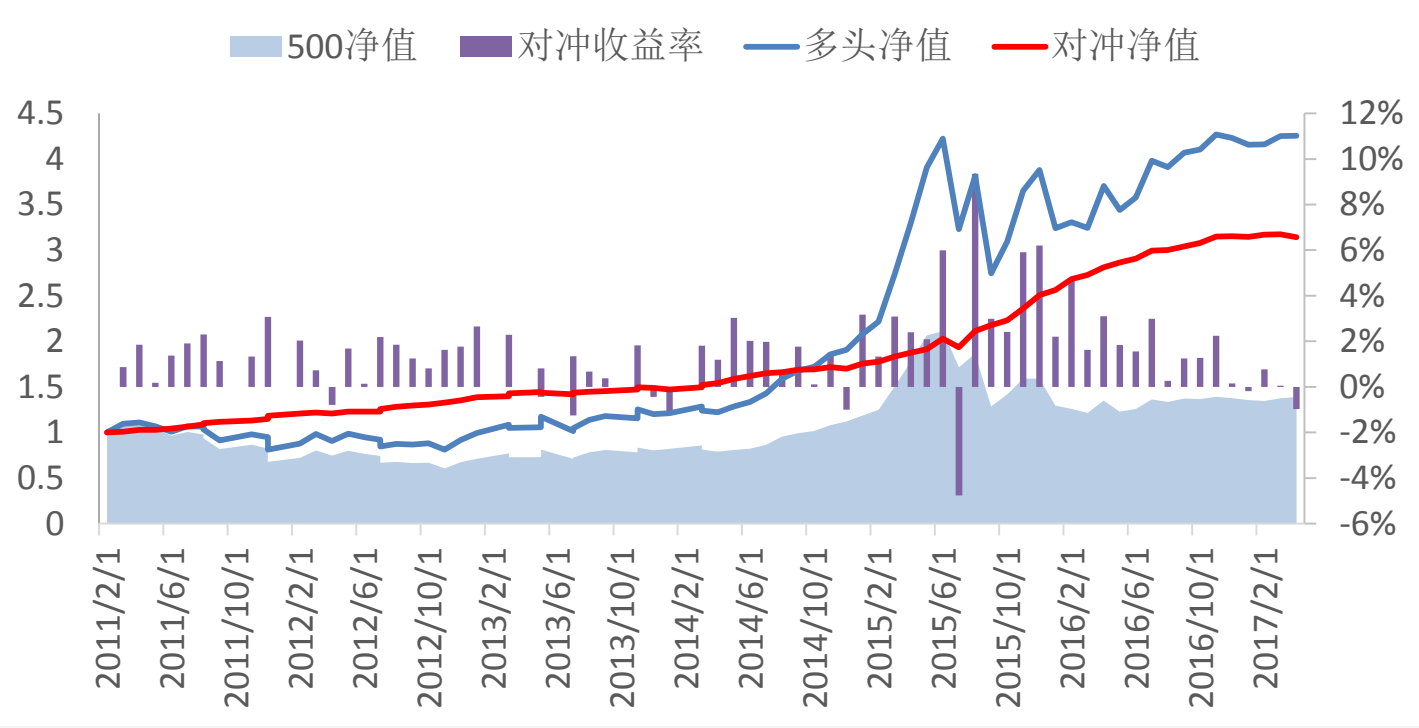




数据来源：Wind，广发证券发展研究中心

因子预测能力：与中证500指数对冲表现

2011年以来，年化收益率20.3%，最大回撤 -4.77%，月度胜率为 88.0%



年份	累积收益率	最大回撤
2011	21.59%	-1.93%
2012	17.98%	-1.35%
2013	13.00%	-2.53%
2014	18.72%	-3.40%
2015	52.48%	-4.77%
2016	26.43%	-1.79%
2017	0.83%	-1.17%

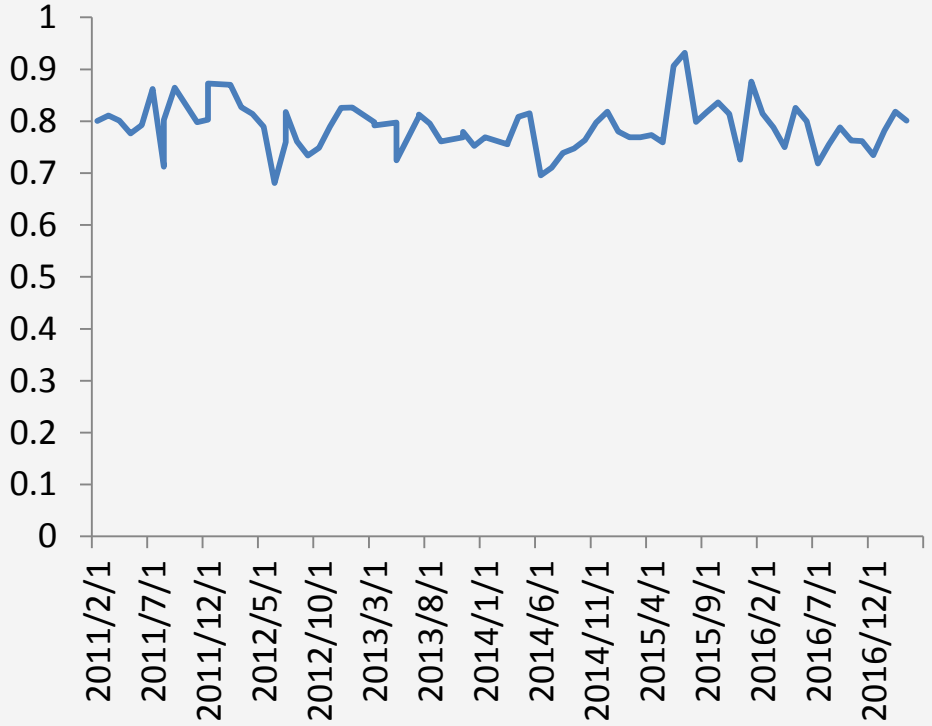
数据来源：Wind，广发证券发展研究中心

换手率和交易成本的影响

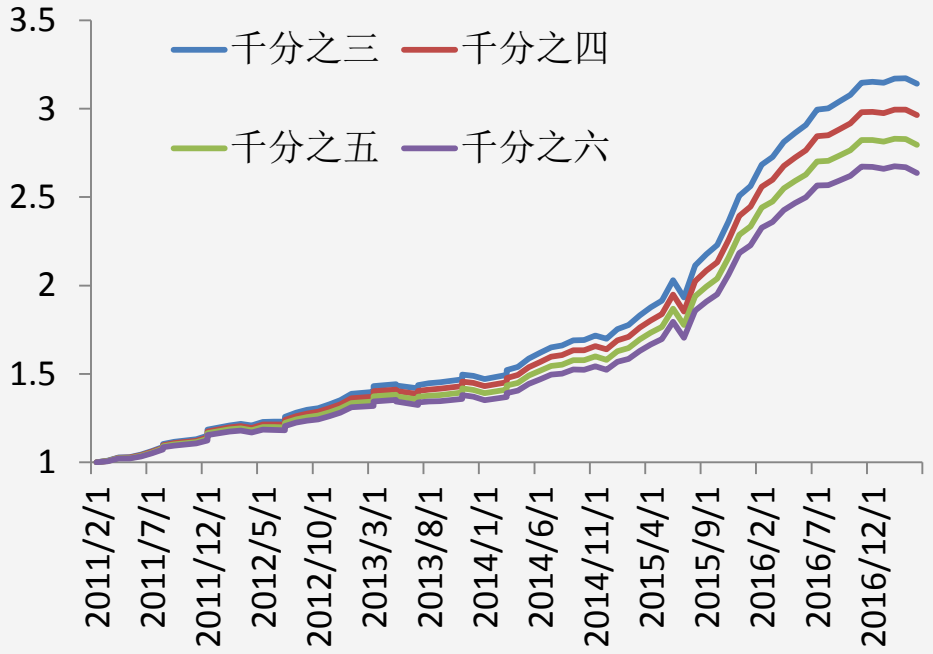
由于技术因子比较多，模型的换手率比较高

交易成本	0.3%	0.4%	0.5%	0.6%
年化收益率	20.28%	19.15%	18.04%	16.93%
最大回撤	-4.77%	-4.86%	-4.95%	-5.04%

超配组合换手率



不同交易成本下策略表现



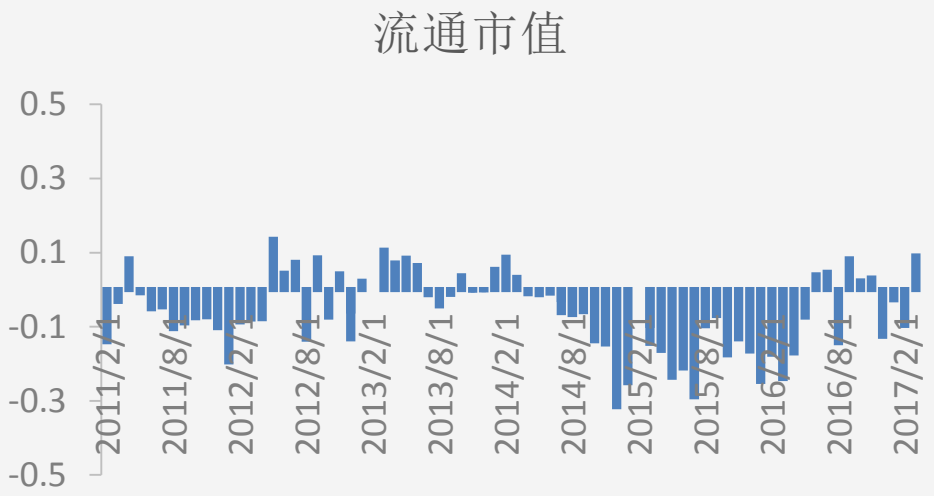
数据来源：Wind，广发证券发展研究中心

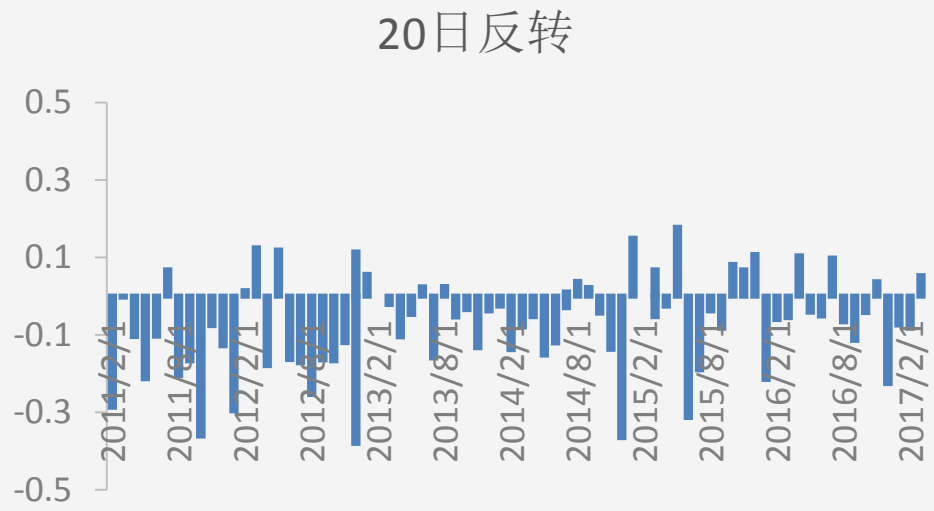
深度学习因子与风格因子的相关性

与常见风格因子的相关性不高

样本外因子的相关性

	深度学习因子
流通市值	-0.060
20日反转	-0.073
20日换手率	-0.119
盈市率	0.017





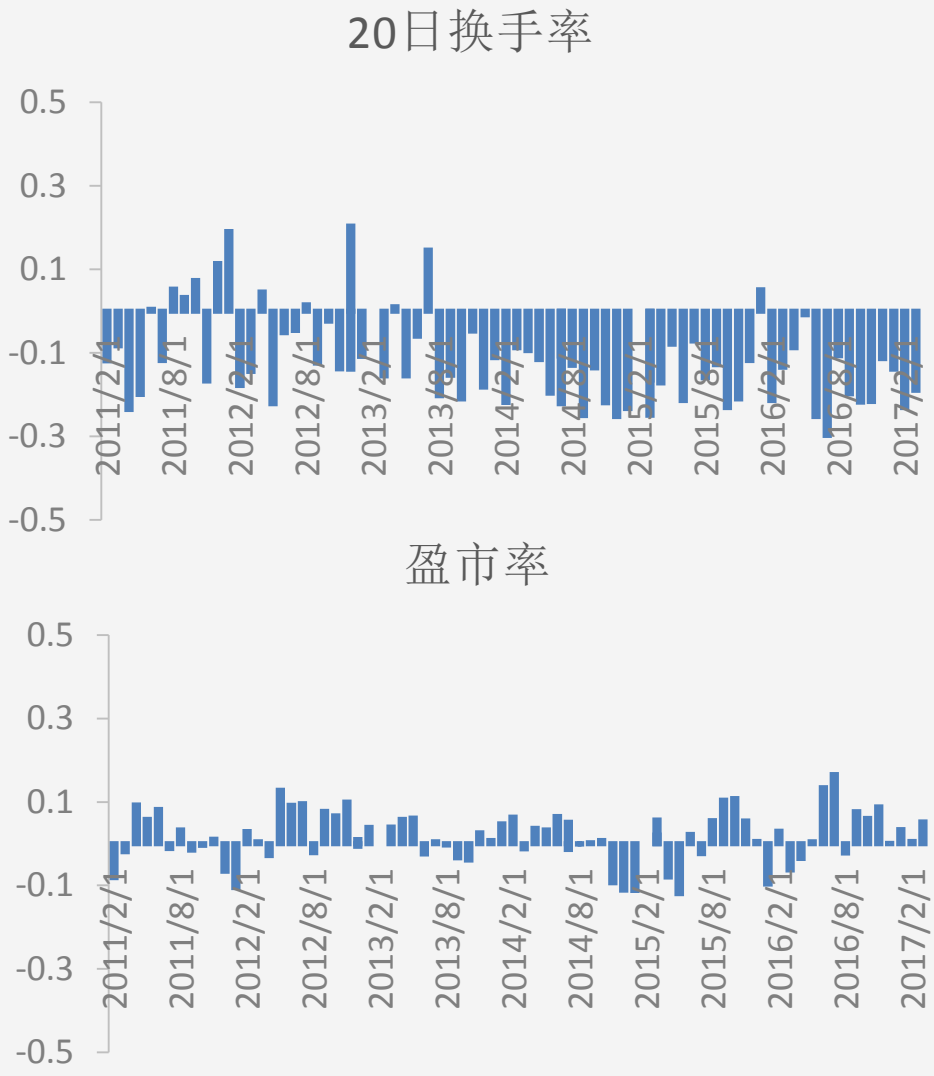
数据来源：Wind，广发证券发展研究中心

深度学习因子与风格因子的相关性

与常见风格因子的相关性不高

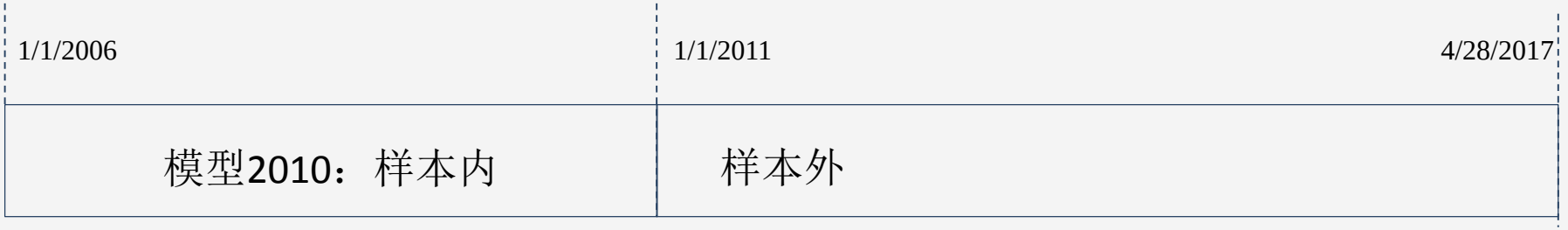
样本外因子的相关性

	深度学习因子
流通市值	-0.060
20日反转	-0.073
20日换手率	-0.119
盈市率	0.017



数据来源：Wind，广发证券发展研究中心

拓展讨论：模型更新



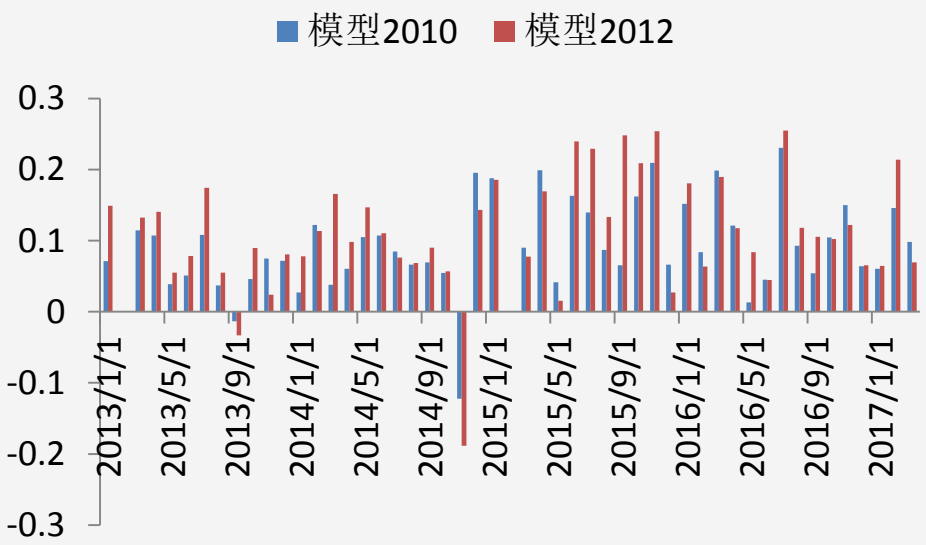
滚动更新模型：训练初始模型之后，固定时间（如，每年）更新一次模型

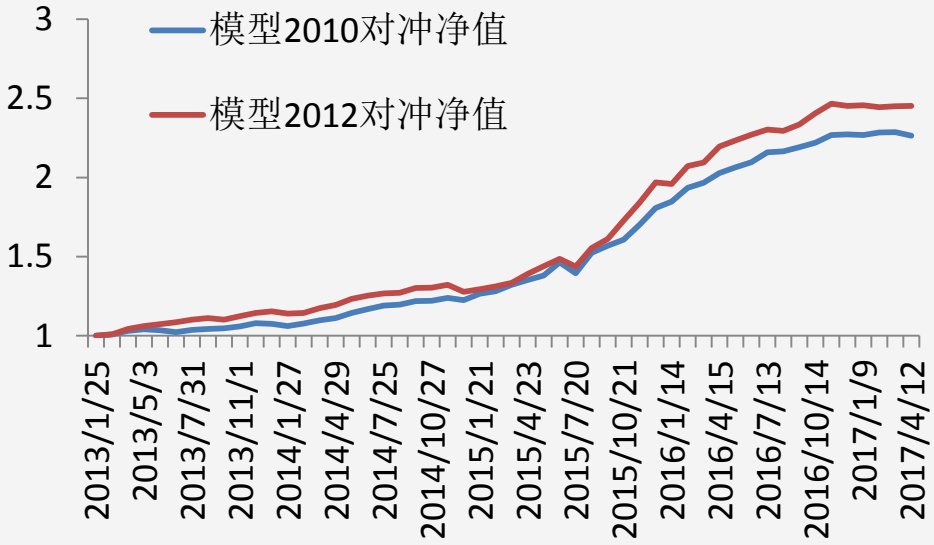
拓展讨论：模型更新

模型2010和模型2012在2013年之后的回测对比

更新后的模型年化收益有明显增强，最大回撤有明显降低

	模型2010	模型2012
年化收益率	22.2%	25.7%
最大回撤	-4.77%	-3.34%





数据来源：Wind，广发证券发展研究中心

01

02

03

04

04

总结

>

- 通过实证分析，证实了深度学习交易策略可以用于月频的选股交易上
- 全市场选股，选取十分之一数量的股票进行配置，用中证500指数对冲，从2011年以来，年化收益率为 20.3%，最大回撤为 -4.77%，月度胜率为88.0%
- 选股因子与常见风格因子（规模、反转、流动性、估值）的相关性不高

本文旨在对所研究问题的主要关注点进行分析，因此对市场及相关交易做了一些合理假设，但这样会导致建立的模型以及基于模型所得出的结论并不能完全准确地刻画现实环境。而且由于分析时采用的相关数据都是过去的时间序列，因此可能会与未来真实的情况出现偏差。本文内容并不是适合所有的投资者，客户在制定投资策略时，必须结合自身的环境和投资理念。

广发证券股份有限公司具备证券投资咨询业务资格。本报告只发送给广发证券重点客户，不对外公开发布。

本报告所载资料的来源及观点的出处皆被广发证券股份有限公司认为可靠，但广发证券不对其准确性或完整性做出任何保证。报告内容仅供参考，报告中的信息或所表达观点不构成所涉证券买卖的出价或询价。广发证券不对因使用本报告的内容而引致的损失承担任何责任，除非法律法规有明确规定。客户不应以本报告取代其独立判断或仅根据本报告做出决策。

广发证券可发出其它与本报告所载信息不一致及有不同结论的报告。本报告反映研究人员的不同观点、见解及分析方法，并不代表广发证券或其附属机构的立场。报告所载资料、意见及推测仅反映研究人员于发出本报告当日的判断，可随时更改且不予通告。

本报告旨在发送给广发证券的特定客户及其它专业人士。未经广发证券事先书面许可，任何机构或个人不得以任何形式翻版、复制、刊登、转载和引用，否则由此造成的一切不良后果及法律责任由私自翻版、复制、刊登、转载和引用者承担。

Thanks！

谢谢

地址: 广州市天河北路183号大都会广场 P.C.510075 电话: 020-87555888 传真: 020-87553600 WWW.GF.COM.CN