机器学习进阶

Xin Tao

日程

I. **算法部分：**

核方法

II. **实操部分：**

机器学习应用于交易的可能模式

核方法与核方程

相对于直接映射方程、间接表达的内积形式是核方法的精髓

基本上对输入变量的映射(变形)都可以看做是<u>核方法</u>

$$\mathbf{x} \longrightarrow \phi(\mathbf{x})$$

<u>核方程</u>: 核方法还有一种间接表达方式，这种表达方式叫做核方程，只关注两个输入数据的内积关系，一般情况下会选用这种形式，记为（字母k代表kernel核的意思）

$$\kappa(\mathbf{x}, \mathbf{x}')$$

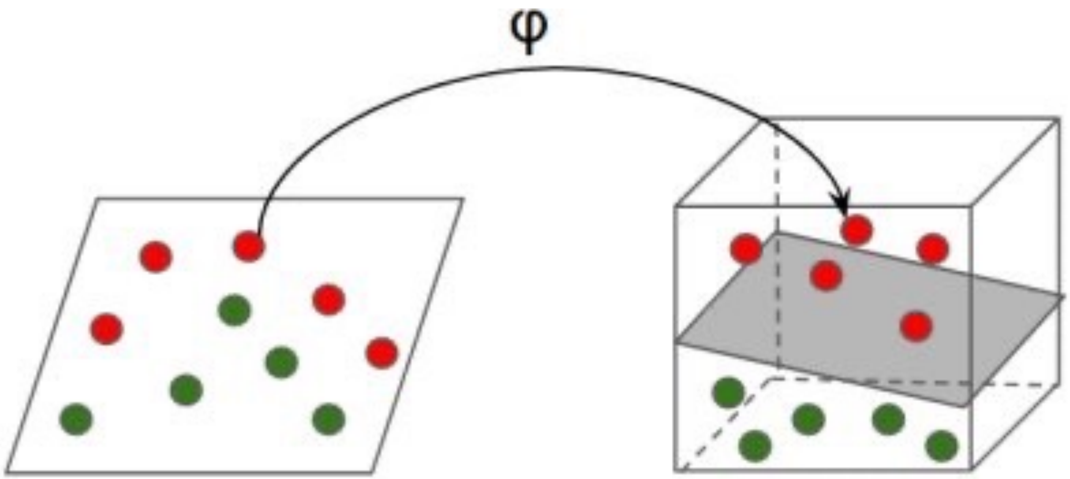
$$\boxed{k(x, x') = \boldsymbol{\phi}(x)^{\mathrm{T}} \boldsymbol{\phi}(x') = \sum_{i=1}^{M} \phi_i(x)\phi_i(x')}$$

常见核方程

Laplacian	$K(x, y) = \exp\left(-\frac{\|x-y\|}{\sigma}\right)$
Rational Quadratic	$K(x, y) = 1 - \frac{\|x-y\|^2}{\|x-y\|^2 + c}$
Multiquadratic	$k(x, y) = \sqrt{\|x-y\|^2 + c}$
Wave	$K(x, y_j) = \frac{\theta}{\|x-y\|} \sin \frac{\|x-y\|}{\theta}$
Power	$K(x, y) = -\|x-y\|^d$
Log	$K(x, y) = -\log \|x-y\|^d + 1)$
Bessel	$K(x, y) = \frac{J_{v+1}(\sigma\|x-y\|)}{\|x-y\|^{-n(v+1)}}$
Cauchy	$K(x, y) = \frac{1}{1 + \frac{\|x-y\|^d}{d}}$

核方法解决机器学习问题的基本方式

核方法的方式是将低维度无法线性解决的问题升维后解决



原输入空间　　　　　z空间

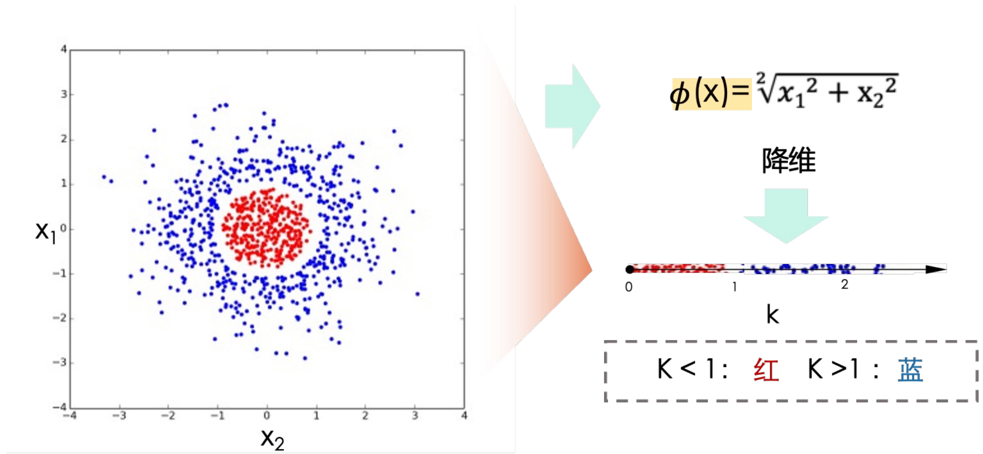
- 将原来线性不可分的空间通过升维的方式转换成更高维度、线性可分的Z空间，实现使用线性方法分类和回归。
- Z空间的维度可能比原空间高很多，甚至Z空间可以是无限高维的。

什么时候使用间接（核方程）表达式?

能使用间接核方程形式的机器学习方法是有限的。

- 如果能够直接用投射方程 ϕ（x）解决问题，不妨直接这样解决。

比如



- 能使用核方程表达式型核方法的机器学习算法是有限的，例如支持向量机(SVM), 高斯过程(GP), 主成分分析(PCA), 典型相关分析(CCA), 岭回归(ridge regression), 谱类聚 (spectral clustering)⋯.

- 关键条件：原算法的<u>损失方程</u>和<u>预测方程</u>可以达为某种x内积型式（对偶型式）来解或优化。

例子：线性回归与岭回归（Ridge Regression）

岭回归是L2正则化的线性回归

- 回顾，线性回归的损失方程为：

$$J(w) = \frac{1}{2}\sum_{n=1}^{N}(w^T x_n - y_n)^2$$

- 使用 ϕ（x）变形后我们得到核化 (kernelized) 线性回归：

$$J(w) = \frac{1}{2}\sum_{n=1}^{N}(w^T \phi(x_n) - y_n)^2$$

- 但目前的损失方程会使用高维数据容易造成过度优化问题，需<u>正则化</u>：

$$J(w) = \frac{1}{2}\sum_{n=1}^{N}(w^T \phi(x_n) - y_n)^2 \boxed{- \frac{\lambda}{2} w^T w}$$

正则化（regularization）部分，本质上是对使用过多系数的惩罚项

- 使用上述损失方程的线性回归即称为岭回归 (ridge regression)，是一种能一定程度上防止过度优化的线性回归（因为使用 ϕ(x)，这里准确地说是<u>核化</u>岭回归的损失方程）。因其正则化项对于w是二次的，因此也叫做L2正则，使用|w|作为正则项的叫做L1正则。

例子：核化岭回归损失的对偶表达

对偶表达就是把损失方程和预测方程中的w替换出来，核方程替换进去

已知核化岭回归损失方程 $$J(w) = \frac{1}{2}\sum_{n=1}^{N}(w^T\phi(x_n) - y_n)^2 - \frac{\lambda}{2}w^T w \quad (1)$$

若定义 $$a_n = -\frac{1}{\lambda}(w^T\phi(x_n) - y_n) \quad (2)$$

从（1），（2）可得 $$w = -\frac{1}{\lambda}\sum_{n=1}^{N}(w^T\phi(x_n) - y_n)\phi(x_n) = \sum_{n=1}^{N} a_n\phi(x_n) = \Phi^T a \quad (3)$$

则损失方程（1）可变为 $$J(a) = \frac{1}{2}a^T\Phi\Phi^T\Phi\Phi^T a - a^T\Phi\Phi^T y + \frac{1}{2}y^T y + \frac{\lambda}{2}a^T\Phi\Phi^T a \quad (4)$$

若定义 $$K = \Phi\Phi^T = \left\{\begin{matrix} \phi(x_1)\phi(x_1) & \phi(x_1)\phi(x_2) & \phi(x_1)\phi(x_3) & \\ \phi(x_2)\phi(x_1) & \phi(x_2)\phi(x_2) & \phi(x_2)\phi(x_3) & \cdots \\ \phi(x_3)\phi(x_1) & \phi(x_3)\phi(x_2) & \phi(x_3)\phi(x_3) & \\ \cdots & & & \end{matrix}\right\} = \boxed{\left\{\begin{matrix} k(x_1,x_1) & k(x_1,x_2) & k(x_1,x_3) & \\ k(x_2,x_1) & k(x_2,x_2) & k(x_2,x_3) & \cdots \\ k(x_3,x_1) & k(x_3,x_2) & k(x_3,x_3) & \\ \cdots & & & \end{matrix}\right\}} \quad (5)$$

Gram 矩阵

将（4）带回（5） $$\boxed{J(a) = \frac{1}{2}a^T K K a - a^T K y + \frac{1}{2}y^T y + \frac{\lambda}{2}a^T K a} \quad (6)$$

（6）的最优解为 $a = (K + \lambda I_N)^{-1}y$ ，预测方程为 $\boxed{y(x) = k(x)^T(K + \lambda I_N)^{-1}y}$

对偶表达

核方程和对偶表达的意义

核方程表达损失函数可以在升维的同时减少付出相应的高计算代价，也不需要理解和思考映射的具体形式

$$J(a) = \frac{1}{2}a^T KKa - a^T Ky + \frac{1}{2}y^T y + \frac{\lambda}{2}a^T Ka \qquad y(x) = k(x)^T (K + \lambda I_N)^{-1} y$$

损失方程和预测方程都不需要使用和计算 w， $\phi(x)$， $\phi(x_i)^T\phi(x_j)$，因此不需要 $\phi(x)$ 的具体表达式，只需要k（x_i， x_j）的表达式。

- 为什么这样做有意义？主要是降低高维变化后与$\phi(x_a)^T\phi(x_b)$相关计算的计算难度。

考虑一个2项式升维映射 $\phi(x) = (1, x_1, x_2, x_1x_2, x_1^2, x_2^2)$，其与之相对应的核函数为

$$k(x_a, x_b) = \phi(x_a)^T\phi(x_b)$$

$$k(x_a, x_b) = 1 + x_{a1}x_{b1} + x_{a2}x_{b2} + x_{a1}x_{b1}x_{a2}x_{b2} + x_{a1}^2x_{b1}^2 + x_{a2}^2x_{b2}^2 \qquad (1)$$

$$k(x_a, x_b) = (1 + x_a^{\mathrm{T}}x_b)^2 \qquad (2)$$

直接计算（1）比展开计算 $\phi(x_a)\,\phi(x_b)$更简单！

如果是20项式升维映射，计算量差别将更大，很多常见的核函数对应着高维升维计算

如 $k(x_a, x_b) = \exp(-\gamma||x_a - x_b||^2)$ 对应的是无限多项式升维映射！

核方程与协方差矩阵

协方差矩阵是线性统计方法的重要核心

为什么$\phi(x_a)^T\phi(x_b)$的计算被难以避免？因为线性统计学方法的核心是协方差矩阵（covariance matrix）和相关性矩阵（correlation matrix），对于n维输入数据X，都可以看做描述不同维度变量"一起动"或"反向动"联动性的矩阵。

$$\text{协方差矩阵 } \mathbf{\Sigma} = \begin{bmatrix} \mathrm{E}[(X_1-\mathrm{E}[X_1])(X_1-\mathrm{E}[X_1])] & \mathrm{E}[(X_1-\mathrm{E}[X_1])(X_2-\mathrm{E}[X_2])] & \cdots & \mathrm{E}[(X_1-\mathrm{E}[X_1])(X_n-\mathrm{E}[X_n])] \\ \mathrm{E}[(X_2-\mathrm{E}[X_2])(X_1-\mathrm{E}[X_1])] & \mathrm{E}[(X_2-\mathrm{E}[X_2])(X_2-\mathrm{E}[X_2])] & \cdots & \mathrm{E}[(X_2-\mathrm{E}[X_2])(X_n-\mathrm{E}[X_n])] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n-\mathrm{E}[X_n])(X_1-\mathrm{E}[X_1])] & \mathrm{E}[(X_n-\mathrm{E}[X_n])(X_2-\mathrm{E}[X_2])] & \cdots & \mathrm{E}[(X_n-\mathrm{E}[X_n])(X_n-\mathrm{E}[X_n])] \end{bmatrix}$$

$$\text{相关性矩阵 } \operatorname{corr}(\mathbf{X}) = \begin{bmatrix} 1 & \frac{\mathrm{E}[(X_1-\mu_1)(X_2-\mu_2)]}{\sigma(X_1)\sigma(X_2)} & \cdots & \frac{\mathrm{E}[(X_1-\mu_1)(X_n-\mu_n)]}{\sigma(X_1)\sigma(X_n)} \\ \frac{\mathrm{E}[(X_2-\mu_2)(X_1-\mu_1)]}{\sigma(X_2)\sigma(X_1)} & 1 & \cdots & \frac{\mathrm{E}[(X_2-\mu_2)(X_n-\mu_n)]}{\sigma(X_2)\sigma(X_n)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\mathrm{E}[(X_n-\mu_n)(X_1-\mu_1)]}{\sigma(X_n)\sigma(X_1)} & \frac{\mathrm{E}[(X_n-\mu_n)(X_2-\mu_2)]}{\sigma(X_n)\sigma(X_2)} & \cdots & 1 \end{bmatrix}.$$

*相当于标准化数据（所有数据除以其标准差）后的协方差矩阵

核方程与协方差矩阵

核方程对于计算核化后的协方差矩阵的意义重大

- 协方差矩阵和相关性矩阵在计算重要的统计量中无可避免，也往往是计算量最大的部分。
- 如一位线性回归的回归系数（斜率）就可由协方差除以自变量方差算得，PCA的核心是对协方差矩阵进行特征值分解…… 。
- 如果我们对X事先中心化，中心化后对与任何维度n都有 E（X'_n）= 0，协方差矩阵变为：

$$\Sigma = \frac{XX^T}{N} = \begin{Bmatrix} x_1 \times x_1 & x_1 \times x_2 & x_1 \times x_3 \ldots \\ x_2 \times x_1 & x_2 \times x_2 & x_2 \times x_3 \ldots \\ x_3 \times x_1 & x_3 \times x_2 & x_3 \times x_3 \ldots \\ \ldots & & \end{Bmatrix} / N$$

- 对比Gram矩阵的公式，很容易看出核方程k（x_i， x_j）对计算使用核方法后的协方差矩阵的意义：

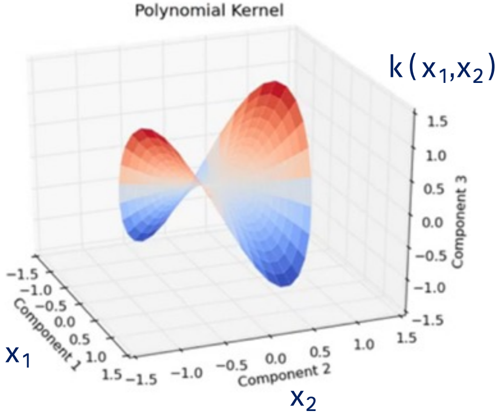
$$K = \Phi\Phi^T = \begin{Bmatrix} \phi(x_1)\phi(x_1) & \phi(x_1)\phi(x_2) & \phi(x_1)\phi(x_3) & \\ \phi(x_2)\phi(x_1) & \phi(x_2)\phi(x_2) & \phi(x_2)\phi(x_3) & \ldots \\ \phi(x_3)\phi(x_1) & \phi(x_3)\phi(x_2) & \phi(x_3)\phi(x_3) & \\ \ldots & & & \end{Bmatrix} = \begin{Bmatrix} k(x_1, x_1) & k(x_1, x_2) & k(x_1, x_3) & \\ k(x_2, x_1) & k(x_2, x_2) & k(x_2, x_3) & \ldots \\ k(x_3, x_1) & k(x_3, x_2) & k(x_3, x_3) & \\ \ldots & & & \end{Bmatrix}$$

Stationary核方程的映射

Stationary核方程映射的意义是将空间关系转化成相关性 *Inference*

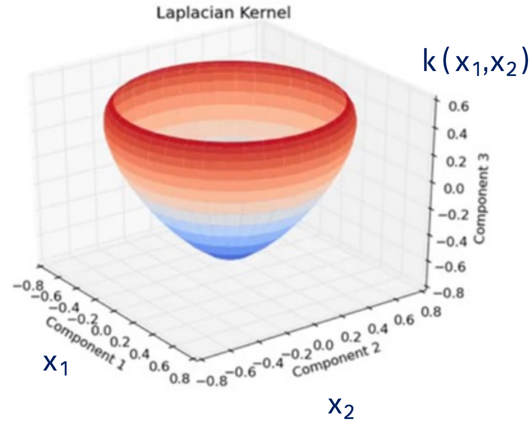
- 一类重要的核方程类型为 stationary kernels， k（x_i， x_j）= k（x_i - x_j），其特殊形式为距离核 k（x_i， x_j）= k（|x_i - x_j| $_n$），这类核的映射取决于x_i，x_j的<u>相对位置和距离</u>。

多项式核



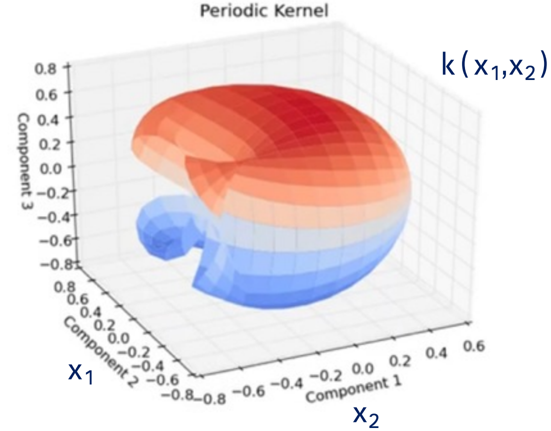
$$k(x,y) = (\alpha x^T y + c)^d$$

拉普拉斯核



$$k(x,y) = \exp\left(-\frac{\|x-y\|}{\sigma}\right)$$

周期性核



$$k_{\mathrm{Per}}(x,x') = \sigma^2 \exp\left(-\frac{2\sin^2(\pi|x-x'|/p)}{\ell^2}\right)$$

- 如用在时序模型的相关性矩阵上，可以理解为stationary 核函数刻画了数据间时间差与其联动性变化的关系，这是核方法模型在金融时序模型应用的基石之一。

核方程的构建

简单的核方程可以构建复杂的核方程

对于合法的基础核方程k_1、k_2（其gram矩阵为正定矩阵），以下复合核方程依然为合法的核方程

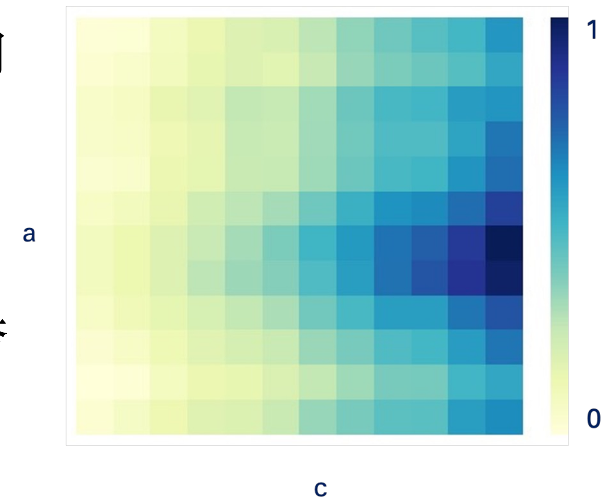
$$
\begin{aligned}
k(\mathbf{x}, \mathbf{x}') &= c k_1(\mathbf{x}, \mathbf{x}') \\
k(\mathbf{x}, \mathbf{x}') &= f(\mathbf{x}) k_1(\mathbf{x}, \mathbf{x}') f(\mathbf{x}') \\
k(\mathbf{x}, \mathbf{x}') &= q\left(k_1(\mathbf{x}, \mathbf{x}')\right) \\
k(\mathbf{x}, \mathbf{x}') &= \exp\left(k_1(\mathbf{x}, \mathbf{x}')\right) \\
k(\mathbf{x}, \mathbf{x}') &= k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}') \\
k(\mathbf{x}, \mathbf{x}') &= k_1(\mathbf{x}, \mathbf{x}') k_2(\mathbf{x}, \mathbf{x}') \\
k(\mathbf{x}, \mathbf{x}') &= k_3\left(\boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}')\right) \\
k(\mathbf{x}, \mathbf{x}') &= \mathbf{x}^{\mathrm{T}} \mathbf{A} \mathbf{x}' \\
k(\mathbf{x}, \mathbf{x}') &= k_a(\mathbf{x}_a, \mathbf{x}_a') + k_b(\mathbf{x}_b, \mathbf{x}_b') \\
k(\mathbf{x}, \mathbf{x}') &= k_a(\mathbf{x}_a, \mathbf{x}_a') k_b(\mathbf{x}_b, \mathbf{x}_b')
\end{aligned}
$$

谱方法：可以利用离散拉普拉斯变换和傅里叶变换的N级近似的思路构建高度灵活且复杂度可控的核方程。

核化模型的训练

核化模型训练的是超参数，超参很难解释。

- 核化模型并不直接训练变量的参数，而是训练核函数k（x_1,x_2）相关的超参数。非参模型（non-parametric models）往往不是“没有参数”，而是使用超参数。
- 例如，对于10项式核函数k （x_1,x_2） = （$\alpha\, x_1x_2+c)^{10}$ 来说，需要优化的超参数只有α，c， 若直接优化参数则非常困难。
- 核化后超参的可解释性相比于直接参数很低。不利理解问题，因此学术研究上很少用核化模型。
- 对超参数优化的可以采用贝叶斯的思路，完整勾勒出超参数与模型损失的关系。



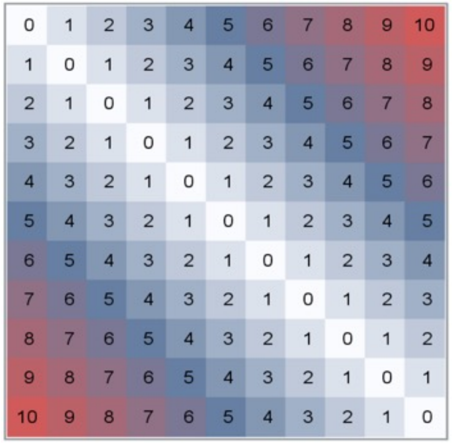
超参数与模型准确度的可能后验关系示例

金融核化模型的训练

金融时序核化模型训练的计算捷径

对于使用时间自变量距离类核函数金融类时序模型的训练而言，因为k线的时间差是固定的，因 $|x_t - x_{t-n}| = n$ 对任何t，n都成立，其基本形式如下。

$$K = \left\{ \begin{matrix} k(0) & k(1) & k(2) & k(3) & k(4)\ldots & k(N) \\ k(1) & k(0) & k(1) & k(2) & k(3)\ldots & k(N) \\ k(2) & k(1) & k(0) & k(1) & k(2)\ldots & k(N) \\ \ldots & & & & & \\ k(N) & k(N-1) & k(N-2) & k(N-3) & k(N-4)\ldots & k(0) \end{matrix} \right\}$$



Toeplitz矩阵示例

该矩阵是稳定的，在计算上可以复用，另外因这类矩阵是toeplit矩阵z形式，计算上难度可以下降一个数量级。

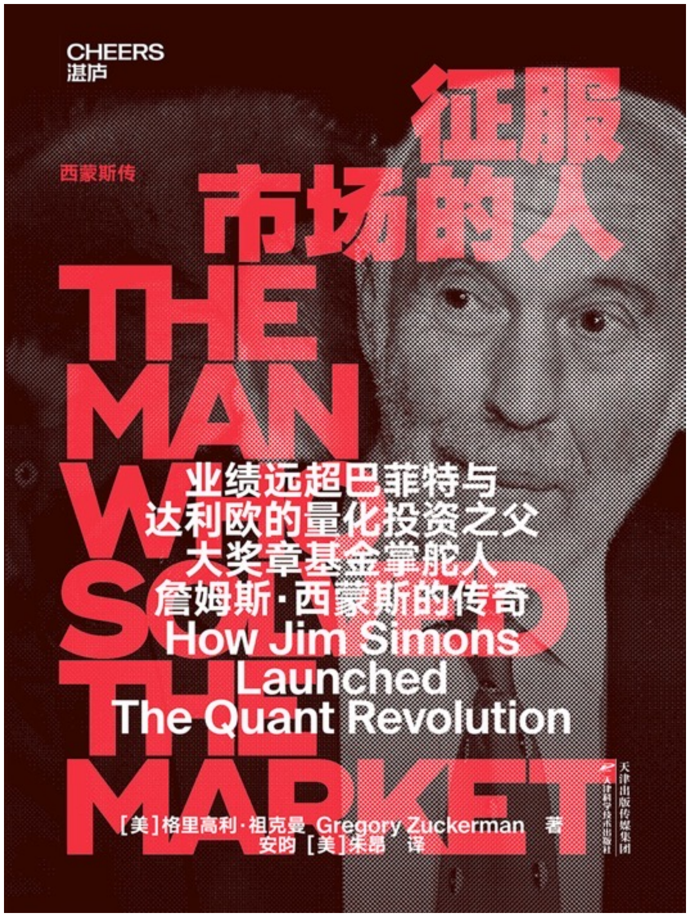

实操部分

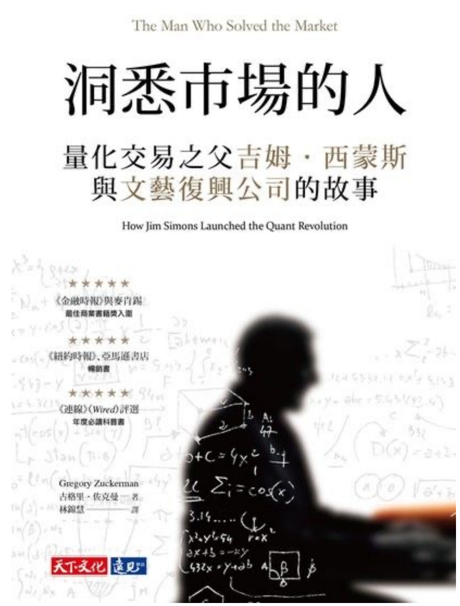

趋势预测
可以通过趋势来追踪商品或货币
可交易的、时间上稳定的
统计异常
你可以在足够长的时间段里 看到这些异常现象是长期存在的
机器
学习
某种意义上 我们做的就是机器学习
短期性
某种程度上来说 更短期的方法
Jim Simons
机器学习与交易
大数据
我们一天内接收兆兆字节的数据
统计学
主要是统计学

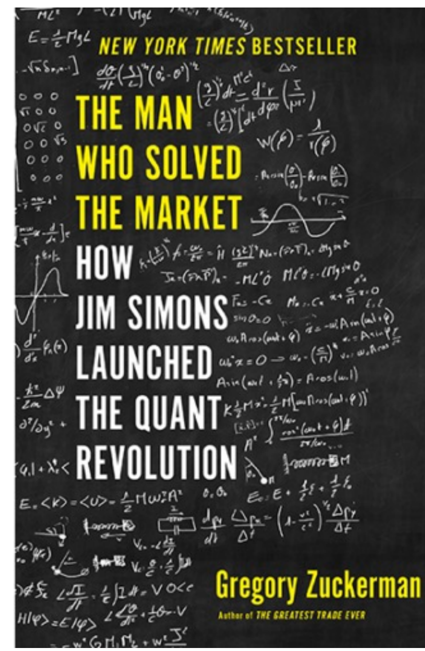
文艺复兴的机器学习交易

基本参考传记

华尔街日报专栏作者格里高利·祖克曼2019年11月出版的关于文艺复兴和西蒙斯的传记。







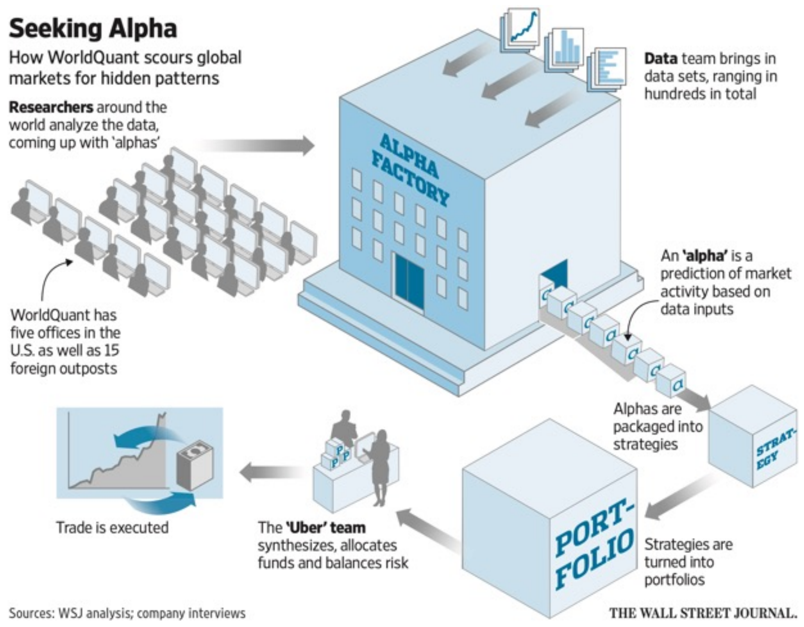
文艺复兴的预测型机器学习交易

核方法、时序相关性、策略分散的短小交易，日历效应、冲击成本控制、统计套利、高性能计算、另类大数据、高频

早期策略主要依靠预测货币、商品和债券市场的趋势获利，95年开始股票市场的统计套利系统才开始成功。

- 1987 René Carmona在文艺复兴的前身公司使用核化回归方法预测趋势的概率分布和不易发觉的时序相关性的特征。
- 1988年文艺复兴成立，Elwyn Berlekamp 基于信息论和凯利公式建议基金使用小额、策略分散和短期（高频）交易方式稳定收益和分散风险。
- 1990，Robert Frey 试图使用统计套利策略建立基金的股票交易系统（基于股价与几乎任何价格序列的统计关系），但由于其编程能力限制而未能成功。
- 1992, Henry Laufer 建议利用基于5分钟K线在特定时间点的日历效应获利， Nick Patterson利用统计方法预测和降低交易的冲击成本。
- 1995年，Robert Mercer 等前IBM语音识别专家完善了股票统计套利系统，股票部分开始盈利。公司的高性能计算系统建立。
- 2000年，基金完成大量收集舆情、市场交易单、和其它高频另类数据以及事件市场冲击数据，基于高效的计算机系统，基于这些数据公司开始高频交易，同年基金回报率升至接近100%。
- 2012年，成为CEO的Mercer投资剑桥分析，2014年剑桥分析对美国选举和脱欧等重大政治事件和金融市场产生重大影响。

常见的其它类型的机器学习交易和管理模式



世坤（World Quant）的Alpha工厂模式，据称已发现4千万个可交易的“市场异常”，未来计划发现一亿个。

其它策略类型

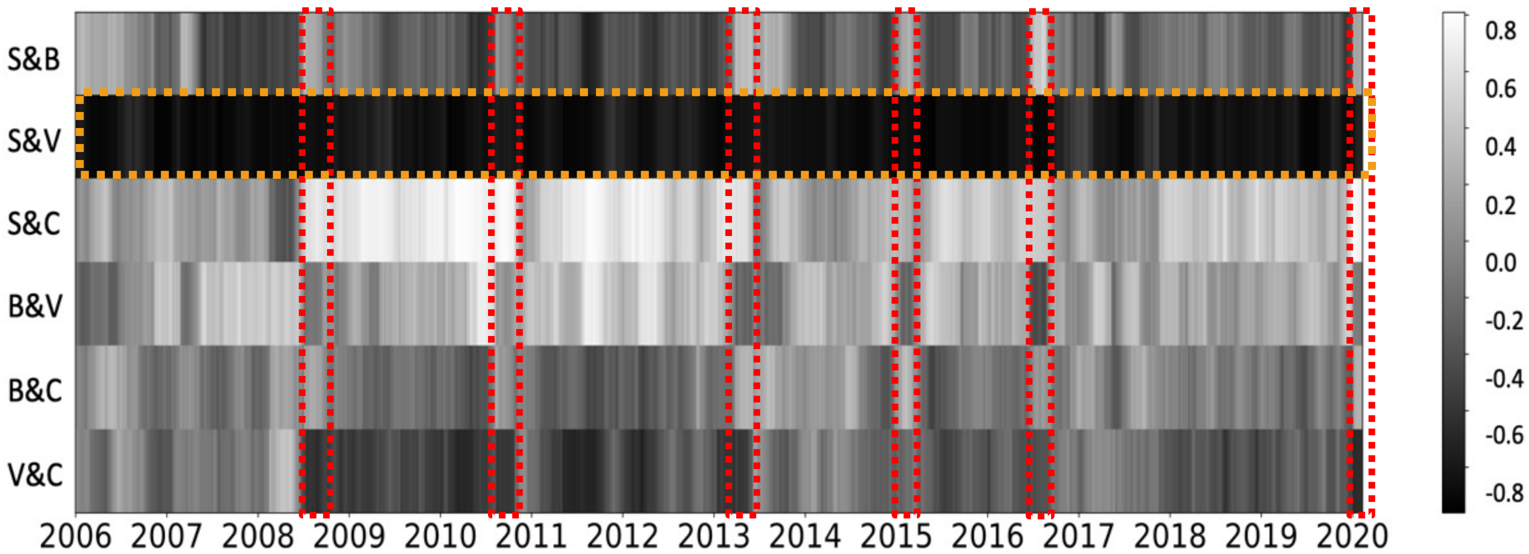
- 统计套利（DE Shaw）
- 高频（Citadel）
- 因子模型（AQR）
- 基于定价的套利
- 混合类型

……

市场的变与不变：大类资产相关性为例

“日光之下无新事”与“唯一的不变是变化”都是对的。

股市（S）、波动性（V）、债市（B）、商品（C）相关性的历史变化



- 统计相关关系长期随着时间而改变，同时也有短期稳定性
- 存在时间上稳定的统计关系
- 也存在重复出现的短期、资产间特定的“互动”模式
- 交易策略的成功概率一定程度上取决于判断不同的市场状态

机器学习交易工程

金融时序核化模型训练的计算捷径

- 分类问题：
  - 买、卖、等的决策变界。
  - 买什么、卖什么、等什么？
- 回归问题：
  - 未来的回报、波动性、定价、流动性。
  - 仓位的大小、投资组合的构建。
  - 风险的衡量与对冲。
- 聚类问题：
  - 市场的基本形态。
  - 资产的聚类和套利。

- 基于机器学习的交易是个复合问题，可用方法、数据的组合几乎无穷。
- 每种交易策略组合的构建、完善都需要大量投入。
- 而事实上每一类的组合仅在非常有限的市场和市场状况（异常情况）能稳定获利，很多时候使用模型的前提比模型更重要。
- 不同交易策略组合本身需要分散化，不然很容易在极端情况出现关联性增大的问题。成功稳定的交易需要大量独立的策略-资产-择时组合，和持续研发投入来寻找新的组合。
- 研发投入（数据密集、人才密集、设备密集）的规模效应很重要，长期很少有单策略可以一招鲜吃遍天。

基于预测的一种单项资产交易模式

金融时序核化模型训练的计算捷径

1. 数据清洗、特征工程。

2. 选择输入长度、预测长度、K线级别、基本模型和损失函数等基本模型设定。

3. 基于全数据抽样训练和交叉验证损失产出多个基础预测器（同模型、不同超参组合）。

4. 计算不同条件下（时间、技术指标、流动性等等）的不同超参组合预测器的后验概率分布。

5. 找到现有模型都很弱的条件范围，基于符合这些条件的数据抽样训练和交叉验证产出补充模型。

6. 创建基于基础模型和补充模型在条件空间的独立增强集成模型（独立交易子模型），能生成集成走势预测与和考虑当前市场条件的的预测信心（概率）。

核心模型 + 修正算法　　**条件空间的增强集成**

$$P(r\,|\,c,D)=\int \boxed{P(r\,|\,\theta,c,D)}\,\boxed{P(\theta\,|\,c,D)}\,d\theta$$

独立交易子模型的预测公式

基于预测的一种单项资产交易模式

金融时序核化模型训练的计算捷径

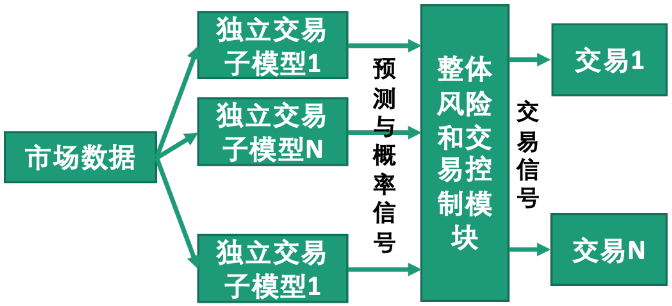
7. 使用不同的输入、预测长度、K线级别、模型形式、损失函数等重复2-6步，产生多个的可独立交易子模型。

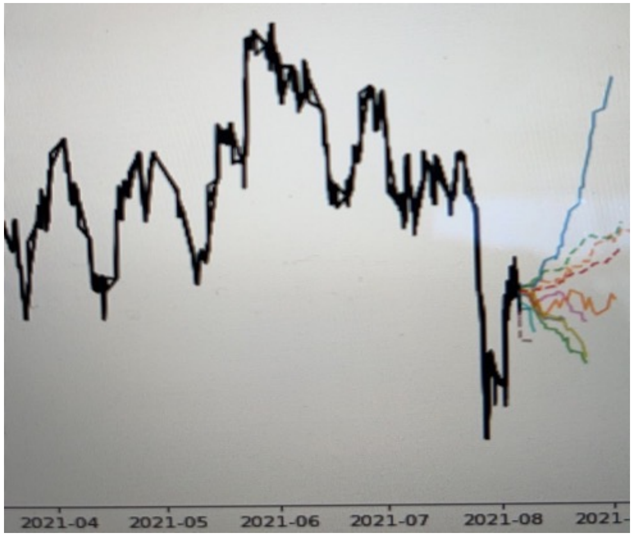
8.从独立交易子模型的预测信号中转化和过滤出交易信号。设计不同市场条件下信号启用、概率修正、资本分配，交易方式和风控规则，交易权重，权重上限等规则。

9. 回测，删除相关过高的模型和交易成本侵占利润过高的模型。再回测，结果作为回报-风险参考。

10. 一段时间后根据新数据更新模型，删除无效的独立模型，训练新独立模型。

整体交易流程





2021年8月6日对沪深300走势的预测

实线为有效的独立集成预测，交易

虚线为无效的独立集成预测，不交易

过于极端的有效预测，不交易