uni 的量化日记

机器学习在量化交易中的应用（进阶课程）

Yiming Wang

2025 / 12 / 20~21

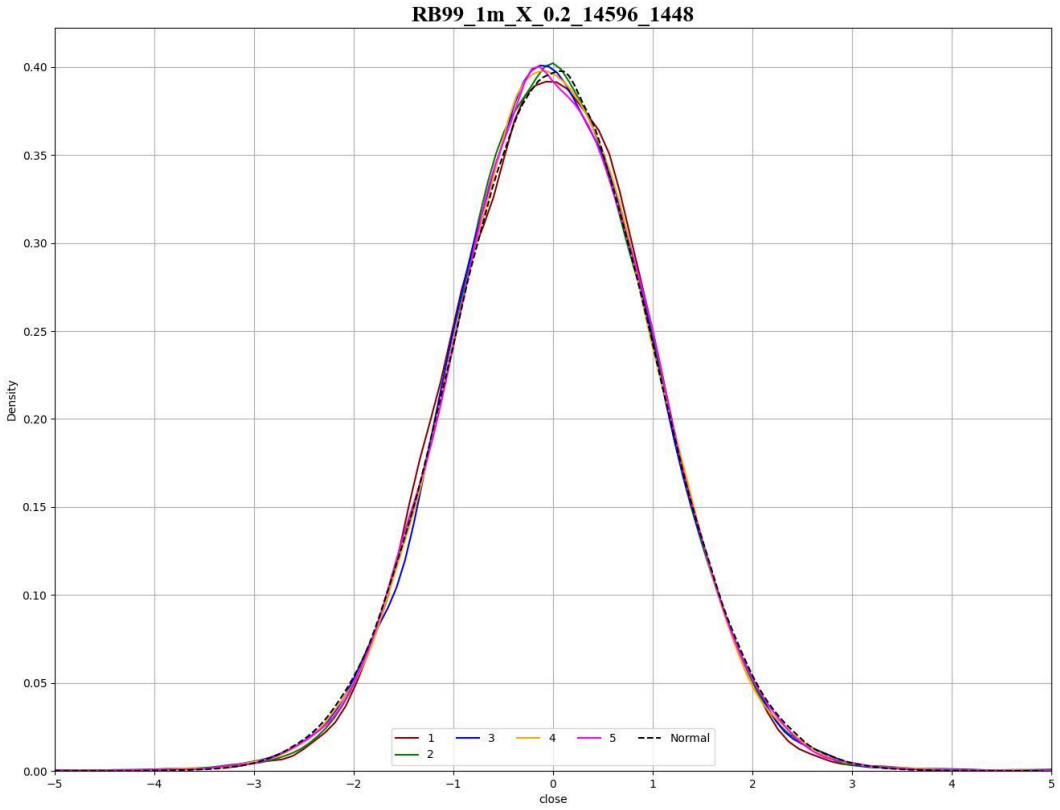
微信 uni

重采样与数据分析

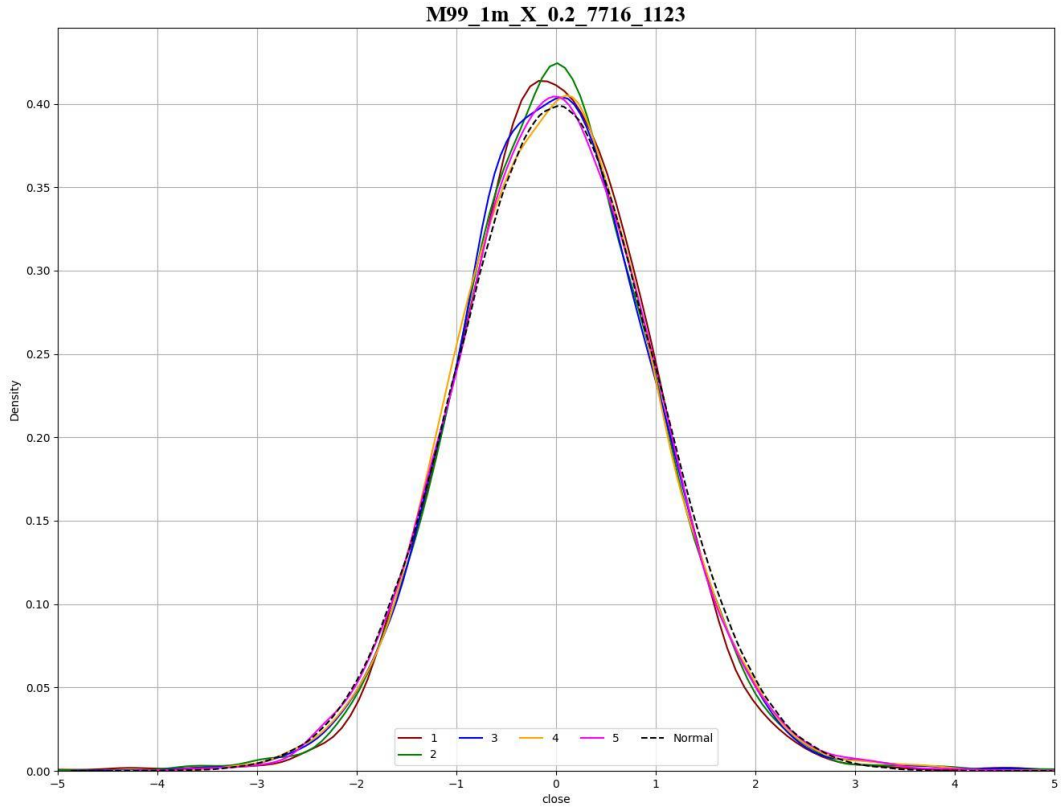
X轴

趋势分析

重采样（X轴）



RB



M

数据分析（趋势分析）

方法	解释	解读
Hurst 指数	衡量时间序列的长期记忆性和趋势持续性	<0.5：反持续性（均值回归） =0.5：随机游走（无趋势） >0.5：持续性（趋势） >0.6：强趋势潜力
ADF 检验	检验序列是否存在单位根（非平稳性）	p>0.05：非平稳（存在单位根） p≤0.05：平稳（不存在单位根）
KPSS 检验	检验序列是否以平稳为零假设（与 ADF 互补）	p<0.05：非平稳（拒绝平稳假设） p≥0.05：平稳（接受平稳假设）

结果	Hurst 指数	ADF 检验	KPSS 检验	结论
强趋势（适合趋势策略）	>0.55	>0.05	<0.05	非平稳且具有趋势持续性
短期趋势可能性	>0.55	≤0.05	≥0.05	平稳但具有一定趋势性
震荡（不适合趋势策略）	<0.5	≤0.05	≥0.05	平稳且无明显趋势
矛盾（需进一步验证）	>0.55	>0.05	≥0.05	ADF 和 KPSS 结论冲突， 可能存在结构突变或非线性趋势

数据分析（趋势分析）

RB120m

	窗口大小	Hurst指数	ADF p值	KPSS p值	趋势适配性评分	趋势类型	min_lag	实际max_lag	KPSS警告
0	20	0.778750	1.000000	0.031823	5	强趋势且非平稳（适合趋势策略）	5	10	
1	40	0.758137	0.148578	0.031764	5	强趋势且非平稳（适合趋势策略）	6	13	
2	60	0.616384	0.909414	0.047859	5	强趋势且非平稳（适合趋势策略）	10	20	

M120m

	窗口大小	Hurst指数	ADF p值	KPSS p值	趋势适配性评分	趋势类型	min_lag	实际max_lag	KPSS警告
0	20	0.516935	0.995244	0.014676	2	弱趋势或反趋势	5	10	
1	40	0.457586	0.672524	0.055655	1	弱趋势或反趋势	6	13	
2	60	0.848464	0.767465	0.058080	3	矛盾（需进一步验证）	10	20	

数据分析（趋势分析）

RB X轴 0.2

	窗口大小	Hurst指数	ADF p值	KPSS p值	趋势适配性评分	趋势类型	min_lag	实际max_lag	KPSS警告
0	20	0.864555	3.909326e-20	0.022232	3	弱趋势或反趋势	5	10	
1	40	1.037698	6.046234e-01	0.045600	5	强趋势且非平稳（适合趋势策略）	6	13	
2	60	0.972240	6.425344e-01	0.011341	5	强趋势且非平稳（适合趋势策略）	10	20	

M X轴 0.2

	窗口大小	Hurst指数	ADF p值	KPSS p值	趋势适配性评分	趋势类型	min_lag	实际max_lag	KPSS警告
0	20	0.836720	1.133349e-14	0.046676	3	弱趋势或反趋势	5	10	
1	40	0.888904	6.651547e-01	0.080098	3	矛盾（需进一步验证）	6	13	
2	60	0.699088	8.302010e-01	0.090625	3	矛盾（需进一步验证）	10	20	

标签

标签类型

标签组合

硬标签的扩展

单步标签与多步标签

机器学习

任务场景：

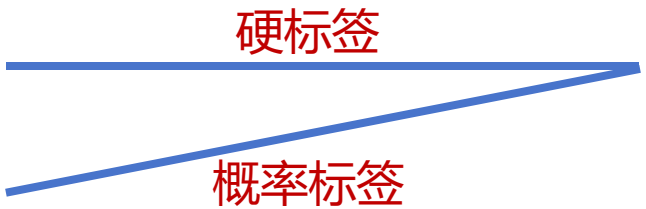
分类

回归

降维

聚类

任务目标：



有监督

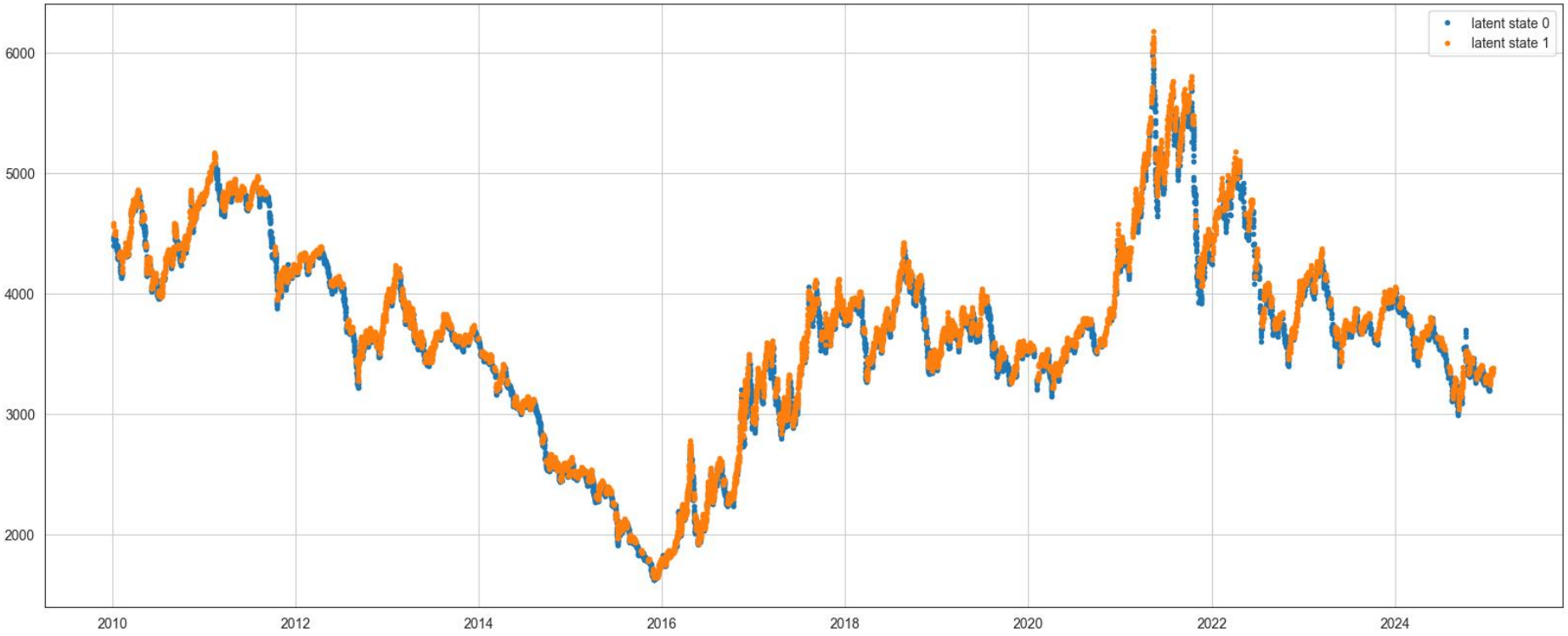
无监督

半监督

强化学习

重点：不同的应用场景，不同的目标任务，需要做不同的选择与组合

用概率模型来为数据做趋势性标签



使用高斯混合模型（GMM）来对数据进行自动标注，
通过不同参数的设置来获取不同任务的趋势性标签

重点：趋势性标签具有通用性，通过GMM可以获得标签的全局最优解

GMM模型的优势

1. GMM可以为每个隐藏状态定义观测值的概率分布

2. n_mix=3 表示每个隐藏状态的观测数据由3个独立的高斯分布混合而成，即每个状态下的观测概率密度函数是3个高斯分布加权组合的结果

3. 由于数据通常是非线性和多模态的，所以单一高斯分布只能描述对称的单峰分布，而3个高斯分量组合可有效拟合双峰、偏态或更复杂的分布形态

例如：“温和”（分量1）、“快速”（分量2）、“回调”（分量3）

标签类型

标签类型	基于上涨概率 P (1)、下跌概率 P (0)	核心信息	任务类型
硬标签	若 P (1) > P (0) 则为 1，否则为 0	仅保留涨跌方向，无概率信息	二分类
双概率标签	同时保留 P (1) 和 P (0)，P (1)+P (0)=1	完整概率分布，含不确定性	多标签回归
带方向的单概率标签	上涨时为 + P (1)，下跌时为 - P (0)	方向 + 概率大小（非对称）	回归
方向强度标签	P (1) - P (0)，等价于 2P (1)-1	方向 + 相对强度（对称）	回归

标签类型

硬标签　　双概率标签　　带方向的单概率标签　　方向强度标签

datelist	logreturn	state	state_0_return	state_1_return	state_0_prob	state_1_prob	state_p1	state_p2
2010/1/6 10:33	0.008750615	1	0	0	2.084606145206		1	1
2010/1/6 14:49	-0.0019361	1	-0.001936142	0	0.001161586785	0.998838413200783	0.9988384132	0.9976768264157128
2010/1/7 9:57	0.006896480	1	0.00689648069582	0	0.006603469166	0.993396530841767	0.9933965308	0.9867930616749884
2010/1/7 10:44	-0.0271297	0	-0.027129748	0	1	1.865910290822842	-1	-1
2010/1/7 13:55	-0.0014196	0	0	-0.001419555	0.999528587710	0.000471412286949	-0.999528588	-0.99905718
2010/1/8 9:22	-0.0023	0	0	-0.002300019	0.999933404591	6.65953969940386e	-0.999933405	-0.99986681
2010/1/8 10:54	-0.0102505	0	0	-0.010250526	0.999997838693	2.161295772340758	-0.999997839	-0.99999568
2010/1/8 14:40	0.005132575	0	0	0.00513257555421	0.999096901857	0.000903098131099	-0.999096902	-0.9981938
2010/1/11 9:53	0.013922773	1	0	0.013922773332033	0.148537646535	0.851462353476430	0.8514623534	0.7029247069411381
2010/1/11 14:50	0.001506410	1	0.0015064108120 9	0	0.003104587293	0.996895412711916	0.9968954127	0.9937908254183083
2010/1/12 11:13	-0.0034821	1	-0.003482123	0	0.004133222923	0.995866777086295	0.9958667770	0.9917335541631469
2010/1/12 14:50	0.005101333	1	0.00510133326158	0	0.003354828895	0.996645171103825	0.9966451711	0.9932903422085816
2010/1/13 9:51	-0.0172761	0	-0.017276087	0	0.999831163463	0.000168836533773	-0.999831163	-0.99966233
2010/1/13 13:48	-0.0031806	0	0	-0.003180628	0.999389580888	0.000610419106014	-0.999389581	-0.99877916
2010/1/14 9:35	0.010293394	0	0	0.010293394666870	0.982937712273	0.017062287731474	-0.982937712	-0.96587542
2010/1/14 14:32	-0.0046278	0	0	-0.004627769	0.998116628358	0.001883371642108	-0.998116628	-0.99623326
2010/1/15 11:19	-0.0074782	0	0	-0.007478244	0.999910861879	8.913812842134692	-0.999910862	-0.99982172
2010/1/15 14:43	-0.0059666	0	0	-0.005966558	0.999365281317	0.000634718687952	-0.999365281	-0.99873056
2010/1/18 10:40	0.002157567	0	0	0.00215756772331	0.987250277898	0.012749722106612	-0.987250278	-0.97450056
2010/1/18 14:55	0.003353415	0	0	0.0033534153836 0	0.994795675170	0.005204324838852	-0.994795675	-0.98959135
2010/1/19 10:55	-0.0016092	0	0	-0.00160916	0.987462678083	0.012537321930619	-0.987462678	-0.97492536
2010/1/20 9:02	0.002035996	0	0	0.00203599625087	0.904955537521	0.095044462466293	-0.904955538	-0.80991108
2010/1/20 10:57	-0.0052486	0	0	-0.005248583	0.950074348417	0.049925651575280	-0.950074348	-0.9001487
2010/1/21 9:06	-0.0039326	0	0	-0.003932579	0.992077902127	0.007922097879208	-0.992077902	-0.9841558
2010/1/21 13:52	-0.0009337	0	0	-0.000933747	0.996560506727	0.003439493264808	-0.996560507	-0.99312101
2010/1/22 10:15	-0.006317	0	0	-0.006317048	0.999370675037	0.000629324959737	-0.999370675	-0.99874135
2010/1/22 14:24	-0.0063073	0	0	-0.006307317	0.999927543254	7.245674407488096	-0.999927543	-0.99985509
2010/1/25 10:37	-0.0020043	0	0	-0.002004329	0.999879695494	0.000120304514308	-0.999879695	-0.99975939
2010/1/26 9:36	-0.0016086	0	0	-0.001608606	0.999585767674	0.000414232315584	-0.999585768	-0.99917154
2010/1/26 14:42	-0.0010169	0	0	-0.001016922	0.992043202784	0.007956797213318	-0.992043203	-0.98408641
2010/1/27 11:20	0.005631159	0	0	0.00563159358 83	0.624775928928	0.375224071065628	-0.624775929	-0.24955186
2010/1/27 14:37	0.004194626	1	0	0.00419462612624	0.053162969979	0.946837030017352	0.9468370300	0.8936740600375004
2010/1/28 9:51	-0.0019688	1	-0.001968832	0	0.037546781543	0.962453218451665	0.9624532184	0.9249064369077737
2010/1/28 11:23	0.000223056	1	0.00022305641894	0	0.004599901607	0.995400098387706	0.9954000983	0.9908001967802698
2010/1/28 14:59	-0.0025566	1	-0.002556578	0	0.004270230802	0.995729769184235	0.9957297691	0.9914595383819149
2010/1/29 10:50	0.003749460	1	0.00374946030384	0	0.016339302742	0.983660697271811	0.9836606972	0.9673213945291897
2010/2/1 9:05	0.004439462	1	0.00443946203540	0	0.083276578746	0.91672342125779	0.9167234212	0.8334468425114412

标签组合

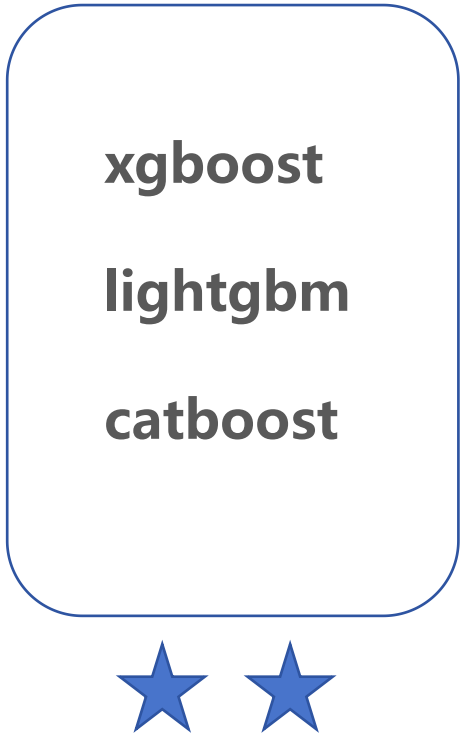
标签组合	组合逻辑	训练方式
硬标签 + 方向强度标签	硬标签提供明确方向， 方向强度标签补充概率信息， 平衡分类准确性与强度预测	分别训练后期组合 多任务集成训练
双概率标签 + 方向强度概率标签	双概率保留完整概率分布， 方向强度提供简化的单变量指标， 适合精细化决策	
硬标签 + 带方向单概率标签	硬标签用于分类决策， 带方向单概率用于量化确定性， 适合阈值敏感的策略	

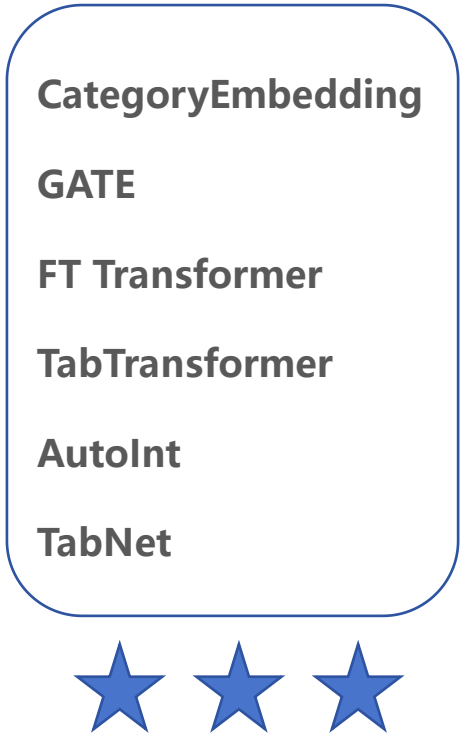
硬标签的扩展

硬标签的扩展	具体操作	训练方式
二分类转三分类	两个或多个标签的不同标记为2 例如：513、713、913不同的标签为2	按照三分类进行模型训练 将2删除后按照二分类训练

注意：硬标签的扩展用法是双刃剑，易产生极端结果

多步标签的模型选择





对比	树模型	结构化神经网络
多步标签处理方式	需通过多输出分类（同时预测多个标签）或分步预测（前一步输出作为后一步输入）实现，间接建模多步依赖	可直接建模序列输出（如通过时序层或 Transformer 捕捉时间步关联），原生支持多步标签的联合优化
时间依赖捕捉能力	难以显式建模时间步之间的关联（如 t+1 与 t+2 标签的依赖），增加了特征工程的难度	通过注意力机制、时序编码等显式捕捉时间步间的依赖关系，更贴合多步预测的序列特性

特征工程

分数阶微分特征

特征工程核心逻辑分类

波动率在不同核心逻辑特征工程中的表现形式

线性高斯状态空间模型特征

深度状态空间模型特征

深度状态空间模型特征与自编码器的区别

分数阶微分计算方法对比

方法	numpy-fracdiff	fracdiff
计算方法	GL + 二项式展开	FFD + Sowell 递归
权重生成逻辑	直接通过二项式系数递归公式生成所有权重（无窗口限制），再通过threshold截断小权重	基于 Sowell 递归，仅在window范围内计算有效权重，天然截断超出窗口的部分
边界处理	丢弃无法完整计算的开头部分数据	提供可选的边界填充策略（如均值、零填充）

特征工程核心逻辑分类

分类	核心	特征
数据驱动	数据分布与统计规律	波动率、样本熵、分布矩
逻辑驱动	领域规则与先验知识	技术指标（MACD、RSI）
状态驱动	系统结构与隐状态	SSM 趋势、HMM 市场状态

状态空间模型（State Space Model, SSM）

SSM在生成金融特征时的核心优势，源于其对动态系统的结构化建模能力和对潜在信息的挖掘能力，这与传统金融特征（如技术指标、统计量）的设计逻辑有本质区别，它是将金融时间序列的隐性信息、动态变化、不确定性转化为可量化的特征

对比	传统金融特征	SSM 生成的特征
底层逻辑	基于固定规则或统计量（如价格变化率的标准差）	基于动态系统建模，从数据中学习状态与观测的关联（数据驱动的自适应规则）
状态感知	仅捕捉显式观测（如价格、成交量） 忽略潜在状态（如市场情绪、风险偏好）	聚焦潜在状态提取，将隐性信息转化为可量化特征
时变性	依赖固定窗口或参数（如 100个窗口期） 无法适应市场结构变化	状态随新数据实时更新，自动适配 regime 切换
不确定性	单点估计（如RSI=65），无置信度信息	概率分布输出（如趋势强度为 0.7±0.1），支持可量化的置信度
多源融合	单变量或简单组合（如价格变动与成交量的比值） 忽略变量间动态依赖	整合多模态数据（基础数据、另类数据、宏观数据）到统一状态空间，建模复杂交互关系
噪声鲁棒性	对异常值敏感	通过滤波算法平滑噪声

重点：状态方程 + 观测方程 + 滤波算法 = 状态的最优估计

波动率在三类特征工程中的表现形式

分类	表示方法	计算逻辑
数据驱动（分布）	历史波动率（滚动标准差）	基于固定窗口期内收益率的标准差 （例如：100个窗口期滚动 σ）
逻辑驱动（规则）	平均真实波幅（ATR）	基于最高价、最低价、收盘价的滚动均值 （例如：N 日 ATR = 平均真实波幅 / 收盘价）
状态驱动（结构）	深度状态空间模型（DSSM）波动率	用NN学习状态转移函数，输出时变波动率 （例如：LSTM-SSM 耦合模型）

重点：三类特征可以混合使用，因为彼此具有互补性，如果追求解释性，需要分别使用

线性高斯状态空间模型（LGSSM）参数

参数	含义	解释
A	状态转移矩阵	潜在状态如何从 t-1 时刻演化到 t 时刻 例如：A 的对角线元素接近 1，说明状态具有持续性
C	观测矩阵	潜在状态如何映射到高维观测数据 例如：将 5 维状态映射到 80 维原始特征， 本质是状态对观测的解释权重
Q	状态噪声协方差矩阵	状态转移过程中的随机扰动（不确定性） 例如：市场突发消息导致状态异常波动， Q 越大，模型允许状态的随机性越强
R	观测噪声协方差矩阵	观测数据中的测量噪声 例如：原始 80 维特征中的噪声， R 越大，模型对观测噪声的容忍度越高

深度状态空间模型（DeepSSM）参数

参数	解释
state_dim	潜在状态向量的维度
lstm_hidden	LSTM 隐藏层维度，控制对时序依赖关系的提取能力
transition网络	状态转移的非线性映射，实现状态演化
observation网络	状态到观测的非线性映射，将潜在状态映射回高维观测数据
initial_state_log_var	初始状态方差的对数，控制初始状态的不确定性
obs_log_var（模型输出）	观测噪声方差的对数，控制模型对观测数据的信任度
transition_log_var（模型输出）	状态转移噪声方差的对数，控制状态演化的随机性，允许状态突变
lr（学习率）	优化器更新步长，控制参数收敛速度和稳定性
eps（雅可比计算）	数值法求导的扰动值，控制雅可比矩阵的计算精度

DeepSSM（LSTM）与 循环自编码器（LSTM）的区别

对比	Deep SSM（LSTM-SSM）	循环自编码器（LSTM-AE）
核心目标	捕捉时序背后的潜在动态规律， 输出可解释的状态特征	对时序进行压缩与重构， 输出用于还原输入的抽象特征
输出性质	输出带概率分布的潜在状态变量	输出确定性，抽象压缩特征
核心逻辑	聚焦状态的转移规律， 通过状态方程建模时序依赖	聚焦重构误差最小化， 通过编码与解码过程压缩冗余信息

模型训练

二分类模型训练

三分类模型训练

单标签回归模型训练

双标签回归模型训练

组合标签分类回归模型集成训练

回归模型评估指标

均方根误差(RMSE)：计算预测值与真实值差值平方的平均数的平方根 $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$

均方误差(MSE)：计算预测值与真实值差值平方的平均数 $\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$

平均绝对误差(MAE)：求预测值与真实值绝对误差的平均数 $\frac{1}{n}\sum_{i=1}^{n}\|y_i-\hat{y}_i\|$

R^2 分数(决定系数)：反映模型对数据的拟合程度，其取值范围在 0 到 1 之间 $1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$

均方根对数误差(RMSLE)：适用于目标变量呈对数分布的情况 $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(1+y_i)-\log(1+\hat{y}_i))^2}$

平均绝对百分比误差(MAPE)：计算绝对误差占真实值的百分比的平均数 $\frac{100\%}{n}\sum_{i=1}^{n}\left\|\frac{y_i-\hat{y}_i}{y_i}\right\|$

注意：MAE 忽略误差方向仅关注偏离程度，RMSLE 对数函数无法处理负数，MAPE 负数百分比误差解释复杂

回归模型评估指标

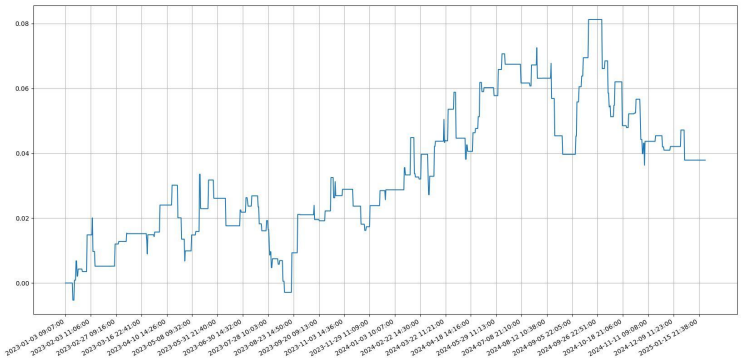
指标	标准	含义
RMSE（均方根误差）	越小越好	整体误差的平方根
MSE（均方误差）	越小越好	整体误差的平方均值
MAE（平均绝对误差）	越小越好	整体误差的绝对值均值
R^2 分数（决定系数）	越大越好	模型对数据的解释能力
RMSLE（均方根对数误差）	越小越好	对数尺度下的整体误差
MAPE（平均绝对百分比误差）	越小越好	相对误差的百分比均值

置信度切片分析

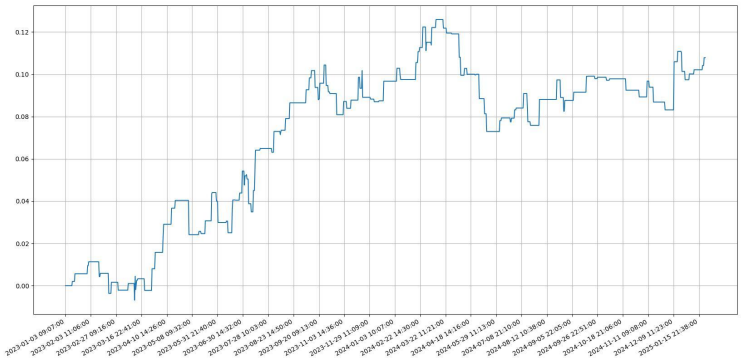
分类模型输出的置信度

回归模型输出的置信度

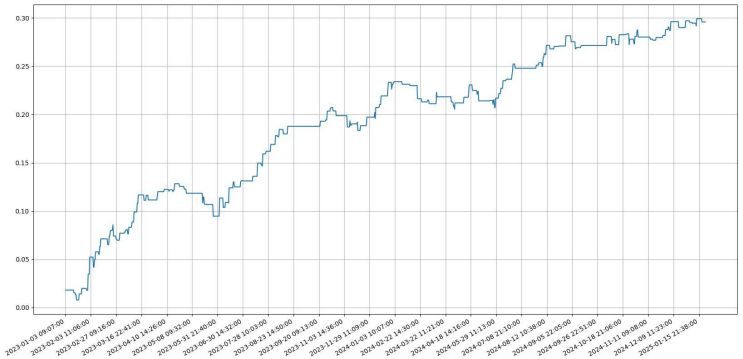
分类或回归模型输出的置信度切片分析



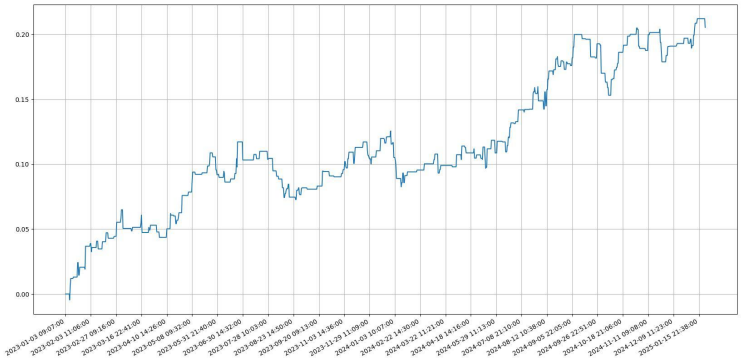
0.5_0.12



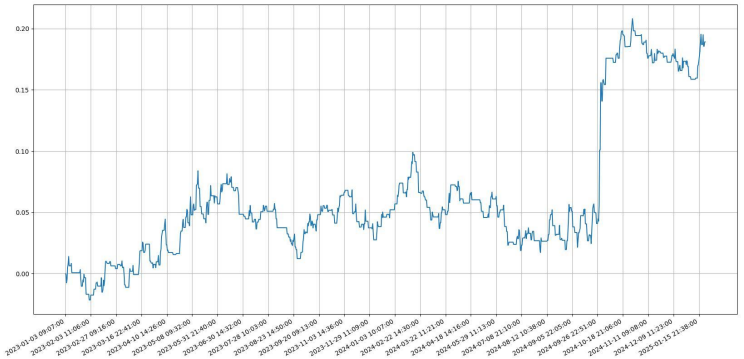
0.6_0.12



0.7_0.15



0.8_0.2

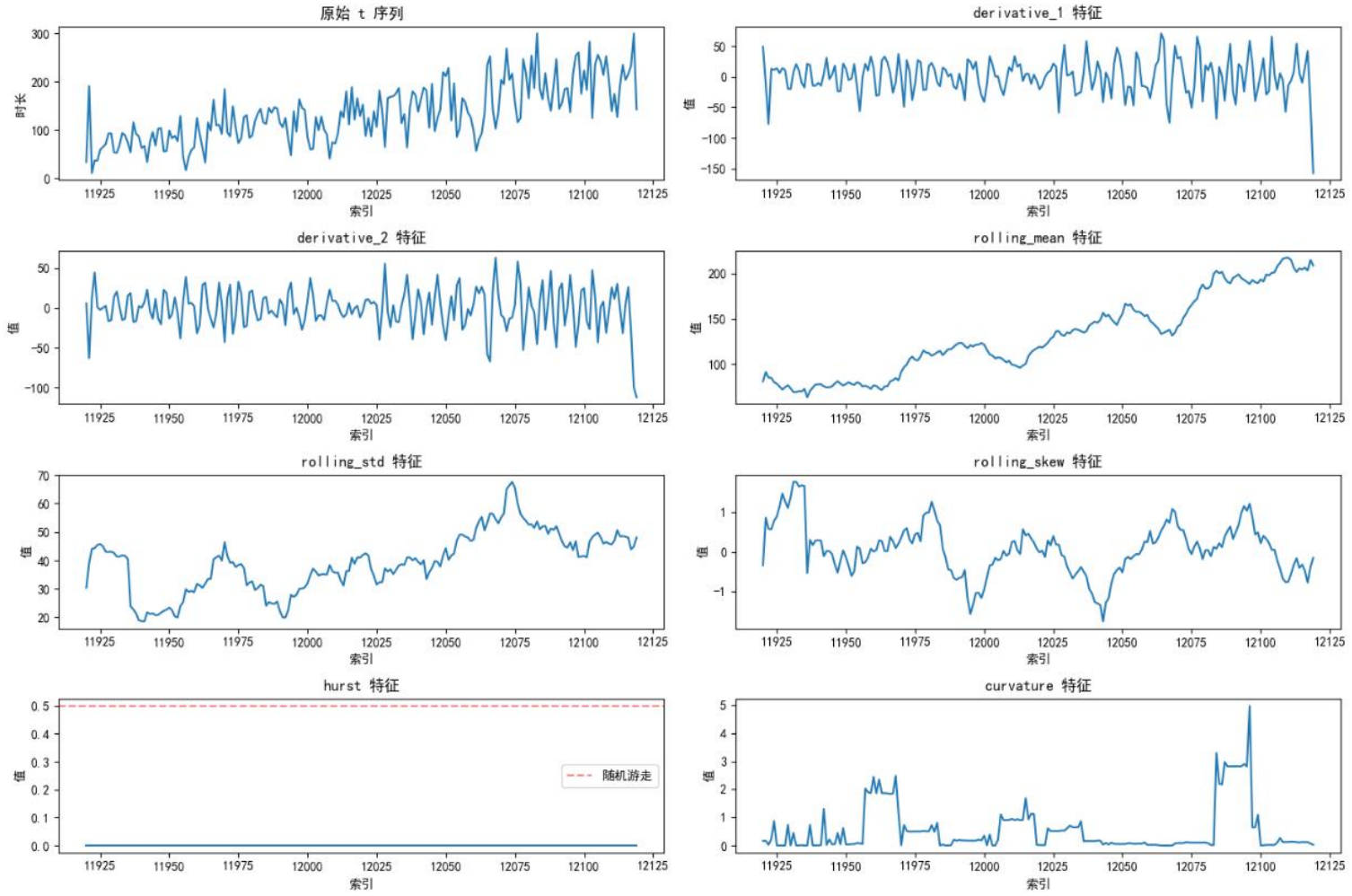


0.9_0.41

市场几何结构分析

Bar的时长

Bar的时长可以看作是动态微分流形，它的曲面特征本质上是市场活跃度与时间交互的几何映射，可以作为波动率预测与市场状态识别提供数学理论支撑，重点是需要跳出只单纯使用bar时长的绝对数值，而是提取它动态变化的拓扑或微分特征，比如：曲率、距离、分形维数，并通过实证验证这些特征与市场状态的映射关系



市场几何结构与统计特征含义

特征	解释	意义
derivative_1	时间序列的一阶导数，相邻数据点的变化率	反映数据的短期变化速度和趋势方向
derivative_2	时间序列的二阶导数，一阶导数的变化率	反映变化速度的快慢，用于判断趋势是加速、减速还是即将出现拐点
rolling_mean	窗口内数据的平均值	平滑短期随机波动，突出长期趋势，减少噪声对整体趋势的干扰
rolling_std	窗口内数据的标准差	衡量局部数据的离散程度，反映窗口内数据的波动性
rolling_skew	窗口内数据分布的偏度	描述数据分布形态，反映局部分布的不对称性
hurst	衡量时间序列的长期记忆性或自相似性	<0.5：反持久性，=0.5：随机游走，>0.5：持久性
curvature	时间序列曲线的弯曲程度	反映趋势的变化率，曲率大说明曲线弯曲明显可能预示趋势即将快速转变
persistence_entropy	衡量时间序列中持久性的不确定性	熵值低：持久性模式稳定，熵值高：模式复杂、不确定性大

趋势模型生产线总结

uni 的量化日记

轴	标签	单步标签 多步标签	特征	选择	降维	任务	模型	损失函数	交叉验证	置信度 切片	概率过滤 （回归）	组合	二次过滤	二级模型	Bar时长
时间轴 自定义轴	二分类硬标签 二分类双概率 二分类方向概率 （原始） 二分类方向概率 （强度） 二分类转三分类 二分类强化 （删除噪音标签）	不同位置的 单步标签可 与特征组合 多步标签可 进行多步预 测	数据驱动 （数值、差分、 熵、频域） 基于分布 逻辑驱动 （1/0） 基于规则 状态驱动 （SSM） 基于结构	SHAP LME Dalex	PCA AE	分类（单标签） 回归（单标签） 回归（多标签） 集成（多标签）	xgboost lightgbm catboost	数学逻辑 金融逻辑	时间 （回归/分类） 分层 （分类） KFold （回归）	0.1 0.05 0.01	>0.6 >0.7 >0.8	不同的轴 不同标签 不同特征 不同维度 不同任务 不同模型	平仓： 1个反向信号 开仓： 连续3个信号 相同	可用组合 方式代替	可作为i特征 （导数、曲率） 可用于过滤 （单一、组合） 可用于策略 （期权、回归）
自定义轴 的信息密 度更均匀， 同时降低 了震荡的 频次	概率标签不是未来 的收益率，而是未 来上涨或下跌的概 率	与标签的趋 势性联动， 趋势越大， 需要学习的 标签越远	最终的处理方 式最好保持一 致性， 注意特征含义 漂移	过滤 包装 嵌入	可与特 征选择 同时使 用	可适度的调参	看整体的平均 值是否有提升	必须可导， 或近似可导	可以做嵌套	在某一个 范围内， 会出现负 贡献	可设置综 合过滤方 式	单一组合最 好不要超过 3个模型	先组合再过滤	需留出更 多的数据 给到测试 集	波动率预测， 市场状态识别

推荐的库

Pyro：概率模型构建

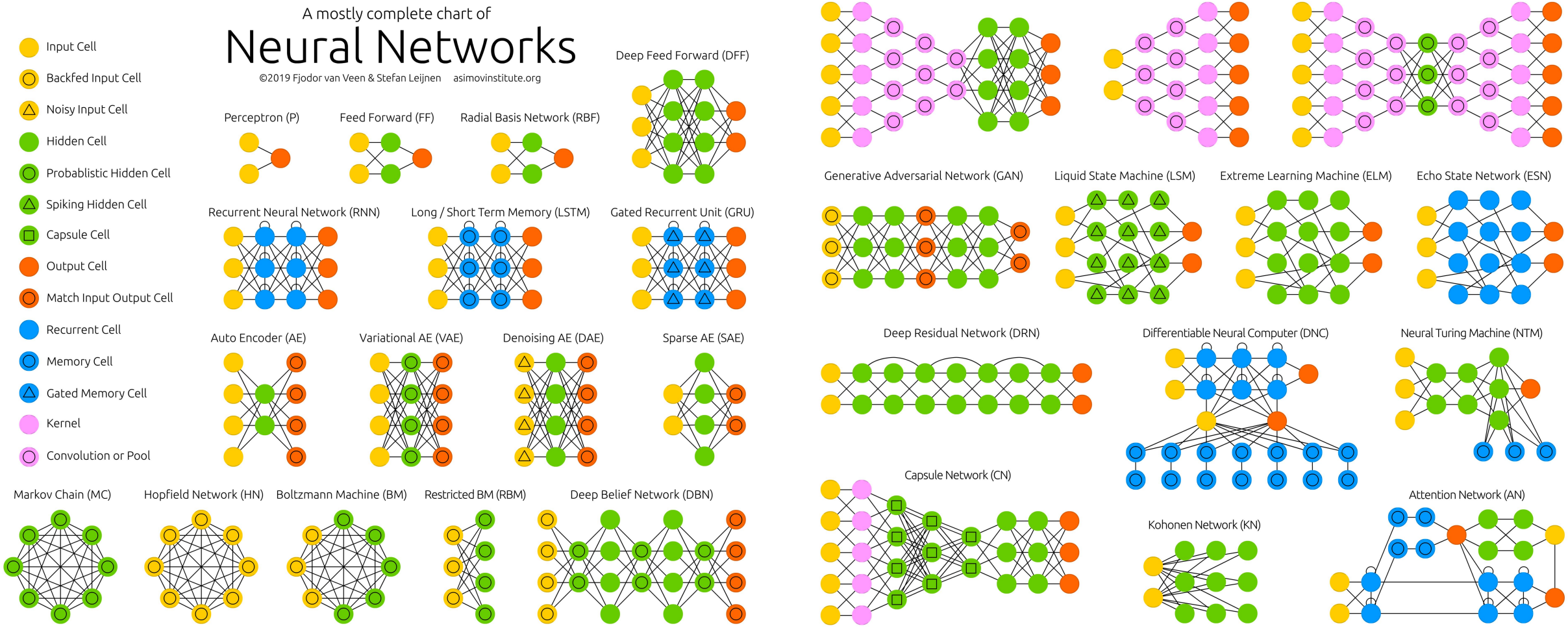
PyPOTS：时序模型构建

LightZero：强化模型构建

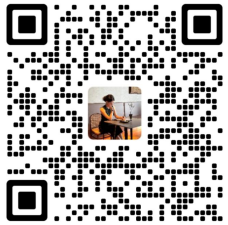
PyTorch Tabular：结构化神经网络模型构建

Neural Network Zoo

www.asimovinstitute.org/neural-network-zoo/



uni 的量化日记



微信 uni