

小学期演讲

目的

- 帮助大家储备进组所需知识
- 用血泪的教训告诉大家 如果不会这些，去面试暑研会死得很惨
- 因本人能力有限，暂时只能帮到AI的一部分:)

流程

- Topic Reading形式
- 选择一个Topic，寻找代表性论文，讲解并回应核心问题（由助教给出）
- 20分钟左右（难点topic可延长），PPT是可选的
- 可能会有专业评委。
- 获得评委意见后，上传最终版视频（可录屏剪辑）
- 视频可能会被上传至网上

评分标准

- 以最终视频为基础，从以下方面评分：
- 演讲详略、节奏是否正确，能否一直吸引听众兴趣？
- 演讲是否回应了核心问题？
- 演讲是否涵盖了恰当的论文及知识点？
- 演讲对于论文的理解是否准确？
- 荣获Outstanding awards的同学将获得3分附加分

核心问题参考-Transformers

- Transformers的提出旨在解决什么问题?
- Word Embedding是如何产生的? 为什么使用token而不是UTF-8?
- Positional Embedding为何要如此设计?
- Multi-head Self-Attention的作用是什么?
- Q、K、V是如何计算的, 各自有何意义?
- 大模型的输出是怎样变换回token的?
- 为什么当今大模型主要使用decoder-only结构?
- 另外参考<http://xhslink.com/a/qSIuDPU1mapdb>

LLM模块选题（拟定）

- Transformers
- Attention Optimization
- Pretraining、BERT
- SFT、LoRA、DeepSpeed
- RLHF、DPO、PEFT
- Llama3、Qwen3
- DeepSeek V3&R1、PPO、DAPO
- RAG、Multi-agent

多模态模块选题（拟定）

- DDPM、DDIM
- ViT、CLIP、Stable Diffusion XL
- Flow Matching、Flux、SD3.5
- BLIP、Llava
- Qwen-VL、Intern-VL
- Autoregressive Image Generation
- Image Editing
- RL&CoT for Image

时间序列预测

(人满就开 🚫)

我们的目标

- 预测股市！
- 在任意指数上打败随机猜测就好 ($p < 0.05$)

我们的数据

● 各种新闻

[illegible]

	1	2	
1	PUBDATE	REPORTTITLE	ABSTRACT
2	2016-10-20 00:00:00.000	关于全柴动力(600218)融资交易风险提示的公告	"上证公告 (监察) [20
3	2016-10-20 00:00:00.000	关于浙江五洲新春集团股份有限公司人民币普通股股票上市交易的公告	"上证公告 (股票) [20
4	2016-10-20 00:00:00.000	关于2016年内蒙古自治区政府一般债券(十六期)上市交易的通知	"各市场参与人: 根据
5	2016-10-20 00:00:00.000	关于2016年内蒙古自治区政府专项债券 (六期) 上市交易的通知	"各市场参与人: 根据
6	2016-10-20 00:00:00.000	关于2016年内蒙古自治区政府一般债券(十三期)上市交易的通知	"各市场参与人: 根据
7	2016-10-20 00:00:00.000	关于2016年内蒙古自治区政府一般债券(十四期)上市交易的通知	"各市场参与人: 根据
8	2016-10-20 00:00:00.000	关于2016年内蒙古自治区政府一般债券 (十五期) 上市交易的通知	"各市场参与人: 根据
9	2016-10-20 00:00:00.000	关于2016年内蒙古自治区政府专项债券 (五期) 上市交易的通知	"各市场参与人: 根据
10	2016-10-21 00:00:00.000	关于为阳光城集团股份有限公司2016年非公开发行公司债券 (第二期) 等七只债券提供转让服务有关事项的通知	"各会员单位: 阳光城
11	2016-10-21 00:00:00.000	关于红利ETF(510880)期权合约调整的通知	" 根据期权全真模拟交:
12	2016-10-21 00:00:00.000	关于IF1610、IH1610和IC1610合约交割的通知	"各会员单位: 股指
13	2016-10-21 00:00:00.000	外媒:马来西亚政府料2017年经济增长加快	" FX168财经报社(香港
14	2016-10-21 00:00:00.000	外媒调查:日本9月出口和家庭支出料下滑	" FX168财经报社(香港
15	2016-10-21 00:00:00.000	多重利空因素打压 油价受挫下滑	" 环球外汇10月21日
16	2016-10-21 00:00:00.000	股指期货新合约上市通知	"各会员单位: 沪深
17	2016-10-21 00:00:00.000	"下破1.09欧元多头颤抖吧 ""十大""理由看空欧元"	" 环球外汇10月21日
18	2016-10-21 00:00:00.000	离岸人民币跌破6.77关口!美指强势加剧贬值预期	" 环球外汇10月21日
19	2016-10-21 00:00:00.000	关于深圳市和科达精密清洗设备股份有限公司股票上市交易的公告	" 深圳市和科达精密清
20	2016-10-21 00:00:00.000	关于安徽黄山胶囊股份有限公司股票上市交易的公告	" 安徽黄山胶囊股份有
21	2016-10-21 00:00:00.000	"沪深市跌幅前二十名(2016.10.21) "	" 沪市跌幅前二十名

我们的数据

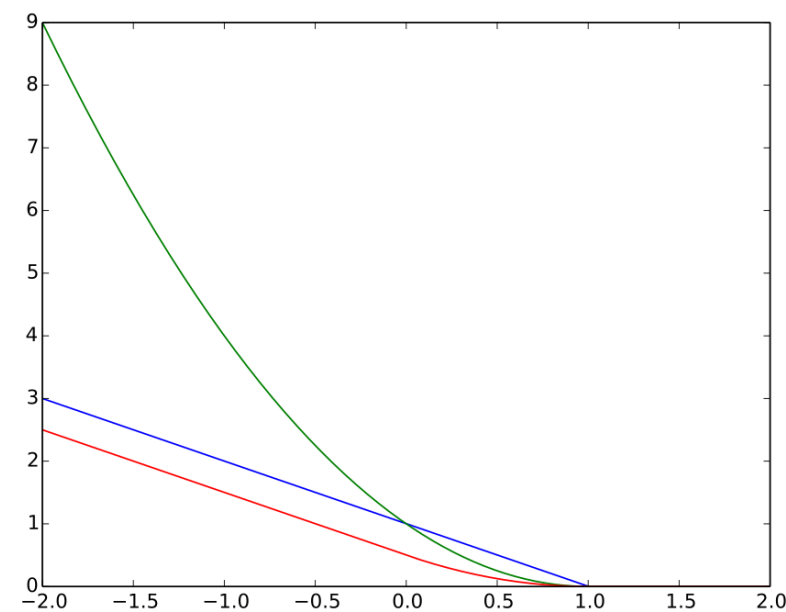
- 各种股价指标（个股&指数，更细粒度的也有）

	A	B	C	D	E	F	G	H	I	
1	trade_date	open	pre_close	high	low	volume	amount	turnover_ratio	close	
2	2008/1/2	5349.76	5338.27	5404.93	5283.45	2.72E+08	9.89E+10	0.14	5385.1	
3	2008/1/3	5381.15	5385.1	5422.67	5315.95	-2.1E+09	1.36E+11	0.2	5422.03	
4	2008/1/4	5430.63	5422.03	5499.08	5422.46	8.8E+08	1.19E+11	0.16	5483.65	
5	2008/1/7	5480.44	5483.65	5569.15	5455.5	1.05E+09	1.23E+11	0.17	5556.59	
6	2008/1/8	5575.95	5556.59	5630.62	5485.23	-2.1E+09	1.48E+11	0.2	5528.05	
7	2008/1/9	5507.12	5528.05	5614.68	5490.31	3.85E+08	1.08E+11	0.15	5613.76	
8	2008/1/10	5631.65	5613.76	5707.67	5602.59	2.13E+09	1.44E+11	0.2	5672.15	
9	2008/1/11	5688.56	5672.15	5702.57	5629.64	1.32E+09	1.22E+11	0.18	5699.15	
10	2008/1/14	5717.52	5699.15	5756.92	5679.29	6.9E+08	1.09E+11	0.16	5731.76	
11	2008/1/15	5743.43	5731.76	5754.11	5652.04	7.94E+08	1.20E+11	0.16	5696.45	
12	2008/1/16	5642.57	5696.45	5642.57	5498.39	1.28E+09	1.27E+11	0.17	5505.72	
13	2008/1/17	5445.8	5505.72	5533.65	5226.64	1.9E+09	1.36E+11	0.19	5365.62	
14	2008/1/18	5358.18	5365.62	5419.57	5307.5	96333204	9.51E+10	0.14	5414.47	
15	2008/1/21	5424.64	5414.47	5434.35	5122.46	4.88E+08	1.02E+11	0.15	5145.73	
16	2008/1/22	5015.43	5145.73	5032.25	4708.39	1.78E+09	1.26E+11	0.19	4753.87	

评价指标

- ACC: 涨跌准确率
- MAE:

$$\frac{1}{n} \sum \|True - Predicted\|$$



- RMSE:

$$\sqrt{\frac{1}{n} \sum \|True - Predicted\|^2}$$

- Cumulative Return: 预测涨就买, 预测跌就卖, 得到的总回报率
- Others, maybe hinge loss

太难了！ 有没有更简单的版本

- 有的兄弟有的
- 不含新闻的纯时间序列预测：公交到站！
- <https://www.kaggle.com/c/paris-city-brain-traffic-flow-forecasting/data>
- 指标已经帮你计算好了:)

How to solve

- 调包并理解！
- <https://github.com/thuml/Time-Series-Library>
- 根据数据特征，提出自己的思路
- 模仿xx、成为xx、超越xx

How to 调包

- 首先速成Python
- 观察代码需要的输入格式
- 写一个程序，把输入数据转换成这种格式
- numpy, pandas :)

Problem 1

- 熟悉Python和数据格式。
- 使用以下naive的方法完成两道题，并计算相应指标： 10pts
 - 用上一刻数据预测下一刻
 - 用前一段时间平均数据预测下一刻
 - 用前一段时间的数据，搭一个全连接神经网络，预测下一刻

Problem 2

- 了解TimeXer、TimesNet、PatchTST、iTransformer和Dlinear的原理, 撰写论文阅读报告（讲清楚每个方法的做法和novelty即可） 阅读报告共15pts
- 使用这些方法完成两道题 10pts
- 想办法改进这些方法（如利用位置信息！）
- $\min(30, \max(0, 60 - \text{score}))$ pts附加分

Problem 3

- 了解SFT和GRPO、DAPO的原理，撰写论文阅读报告 阅读报告共15pts
- 分别使用这两种方法在DJIA或标普500或沪深300或其它股票指数上训练大模型（推荐Qwen3 32B），达到与random guess有显著性差异 10+15pts
- 想办法改进这些方法（如使用特征交互，将新闻的隐空间与股价变化的隐空间对齐！），提交文档简单说明你的做法。
- $\max(0, 2 * (\text{ACC} - 52) + \max(0, \text{MAE} - \text{MAE}_{PD}))$ pts附加分（本部分可溢出！）
- 其中 MAE_{PD} 指使用Previous Day的股价来预测下一天股价得到的MAE

Summary

- 满分80分
- 60分基础: 10 (P1) +10 (P2) +15 (阅读报告) +25 (P3)
- P2有30分bonus, 不可溢出
- P3有无上限可溢出的bonus
- 预期码量1500行
- 得分上下限差距较大

你将学会

- 常见时间序列模型的原理与应用
- 大语言模型的原理与应用
- SFT微调、GRPO微调、.....
- 科研初体验（上道了找我发论文！）

Thanks!