2018 lambda 中国高校计算机大赛

大数据挑战赛决赛分享

目 录

CONTENTS

团队介绍

赛题初见

解决方案

方案总结

1 团队介绍

厦门大学 XIAMEN UNIVERSITY

清华大学 Tsinghua University

快手

自我介绍

- 姓名：Andrew Ng
- 学校：厦门大学
- 专业：金融硕士

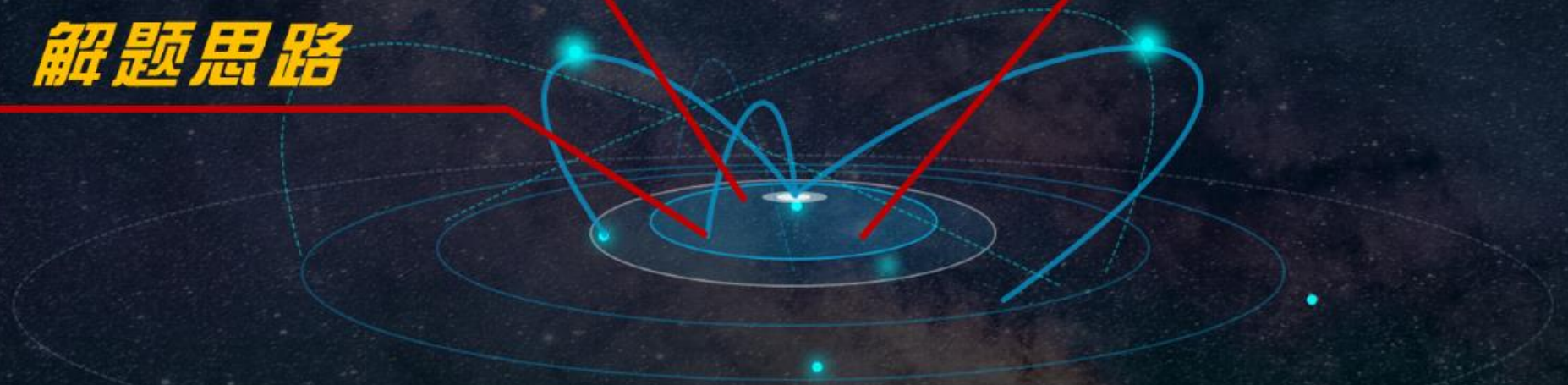
厦门大学 XIAMEN UNIVERSITY 清華大學 Tsinghua University 快手

赛题初见

赛题重述

数据探索

解题思路

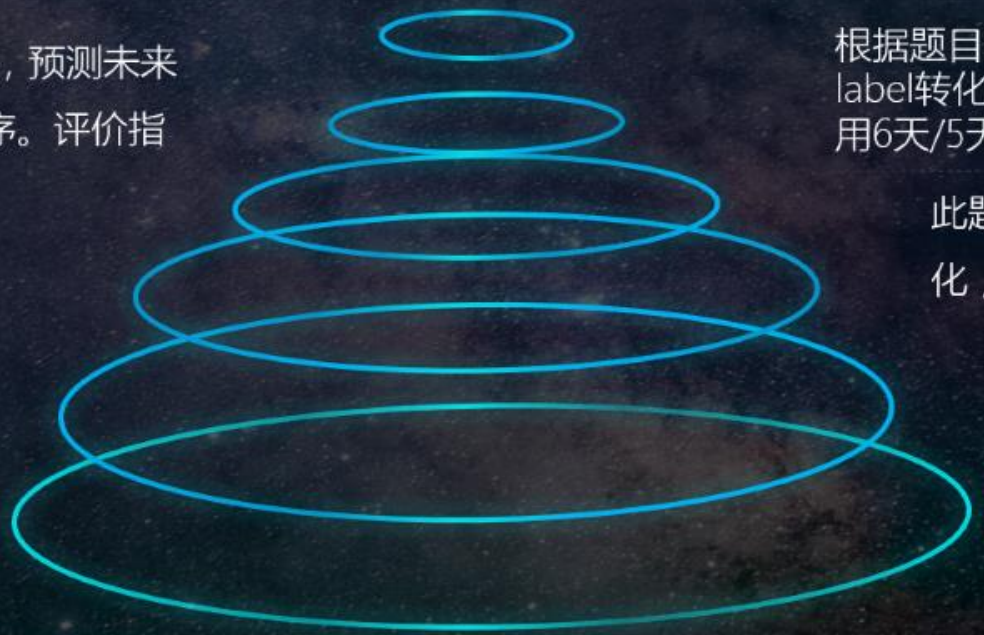


2 赛题初见

厦门大学 XIAMEN UNIVERSITY　清华大学 Tsinghua University　快手

「赛题重述」

基于30天内用户使用快手app的相关行为，预测未来7天内用户是否活跃的概率，对此进行排序。评价指标为AUC。

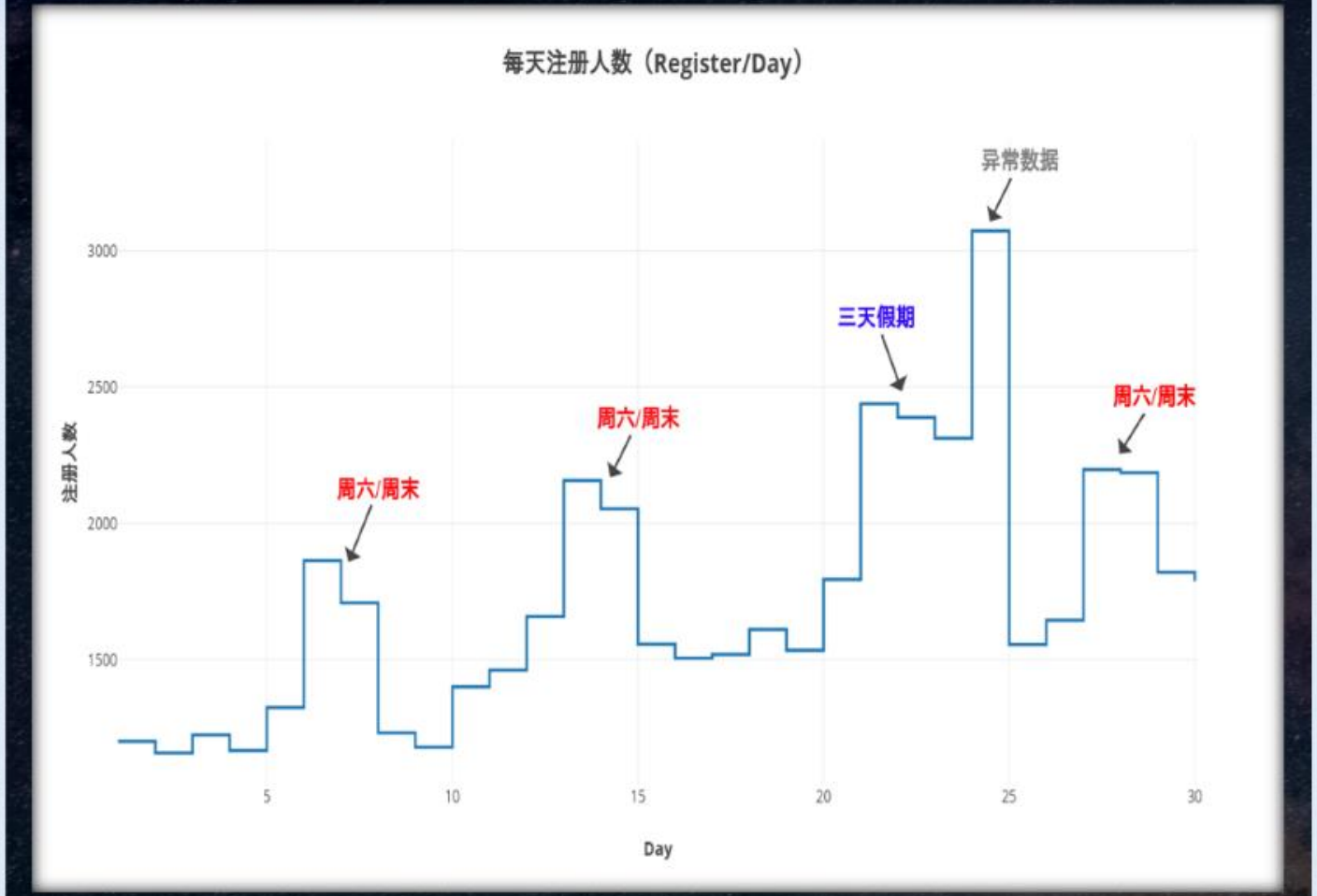


「解题思路」

根据题目要求预测未来7天是否活跃，可以按7天进行打label转化为二分类问题（7天时间长，噪音大，是否可以用6天/5天或4天打label，这样噪音较少）

此题为时间序列题，可将时间序列面板化，使用传统机器学习模型

使用RNN(Many-to-Many)


每天注册人数（Register/Day）
异常数据
三天假期
周六/周末
周六/周末
周六/周末
注册人数
3000
2500
2000
1500
5
10
15
20
25
30
Day

数据探索

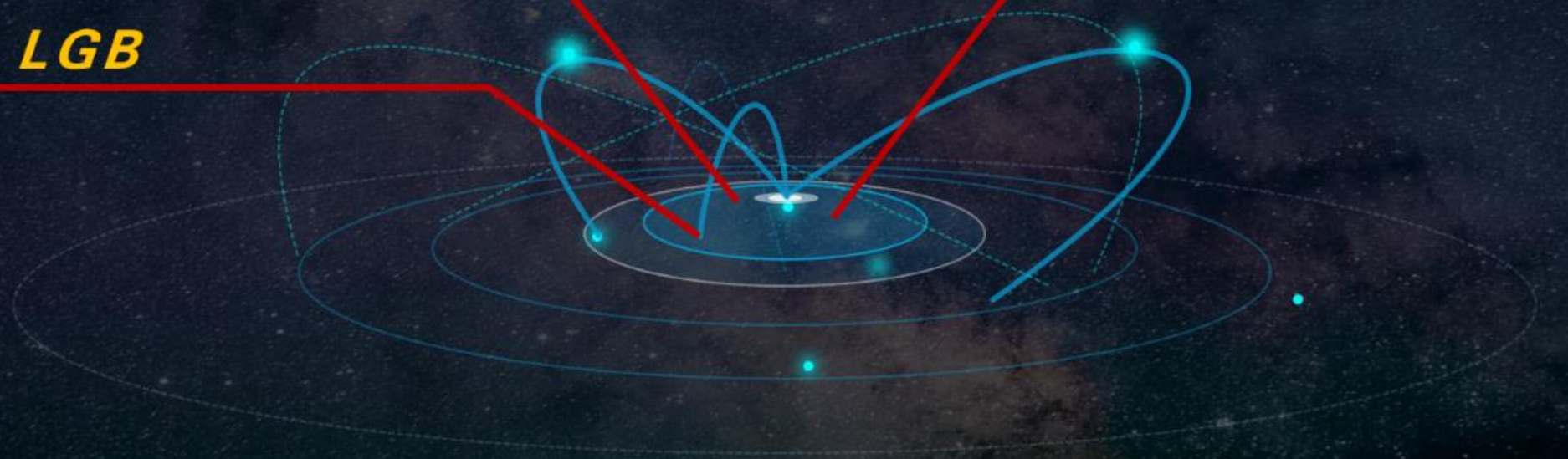
厦门大学 XIAMEN UNIVERSITY 清華大學 Tsinghua University 快手

「解决方案」

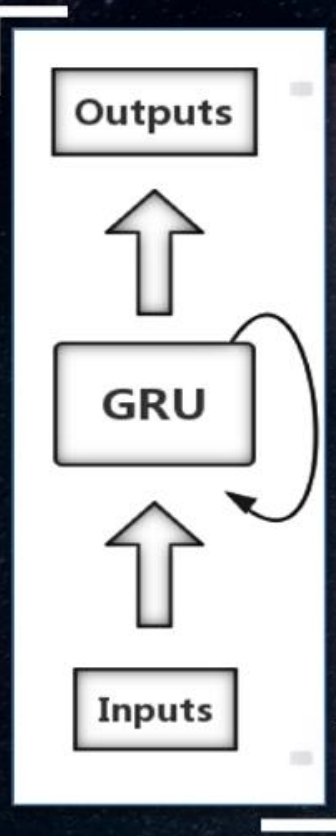
RNN(many-to-many)

LGB

融合方案



Batch选择与采样策略



不同user,id
长度不同

Register Day	Seq_length	User_id Set	Sampling Strategy (Occurrence Times)
1	30	{1, 5, 24, 90, ...}	[30] * (30 - 7)
2	29	{46, 56, 193, ...}	[29] * (30 - 7)
3	28	{...}	[28] * (30 - 7)
4	27	{...}	[27] * (30 - 7)
...	...	.....	.....
23	8	{...}	[8] * (8 - 7)
24	7	{...}	None
25	6	{...}	None
26	5	{...}	None
27	4	{...}	None
28	3	{...}	None
29	2	{...}	None
30	1	{...}	None

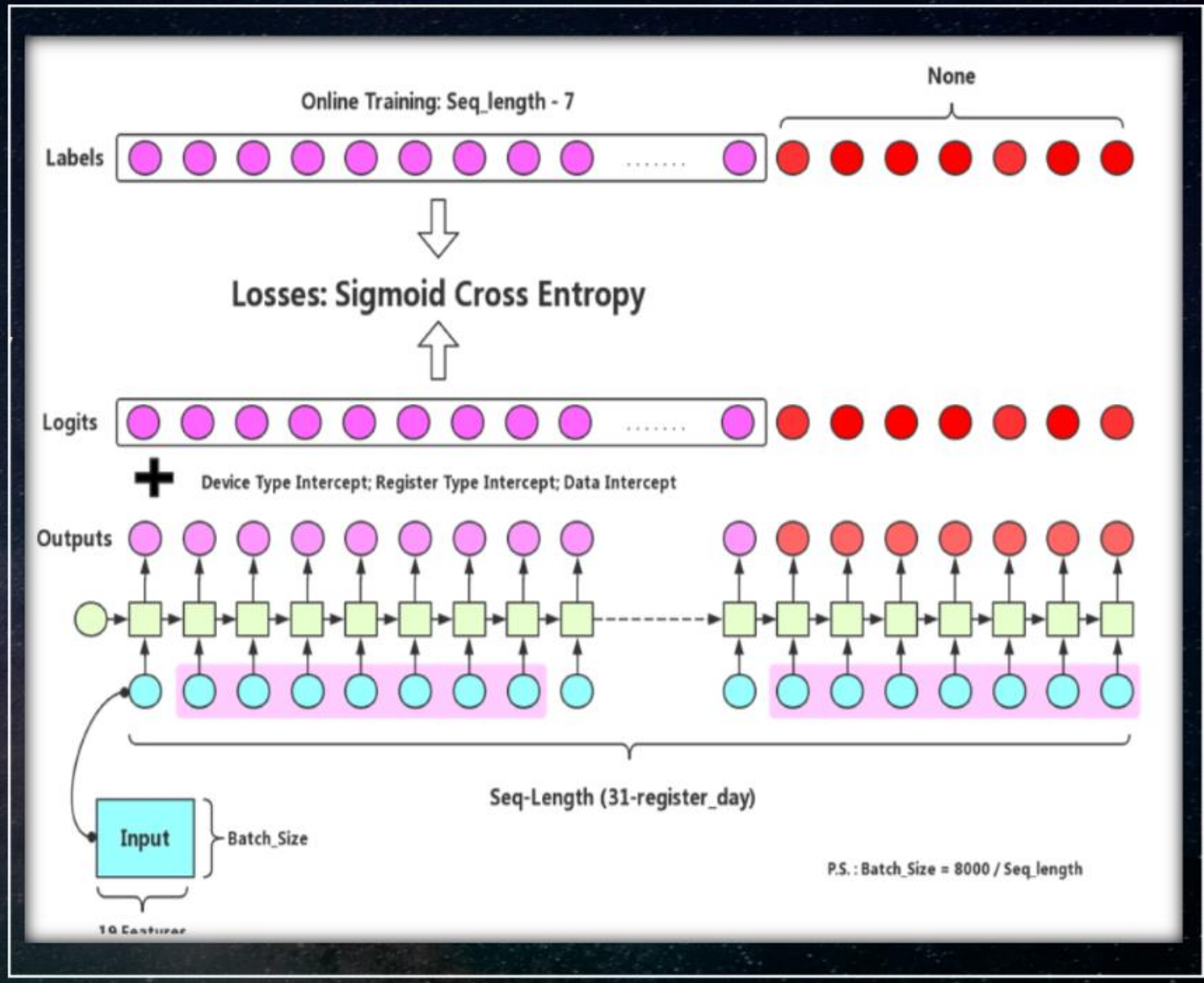
3 解决方案

厦门大学 XIAMEN UNIVERSITY　清華大學 Tsinghua University　快手

RNN (Sampling)

01 线下调参：
1~16线下训练，
23测试

02 线上提交：
1~23训练，
30预测



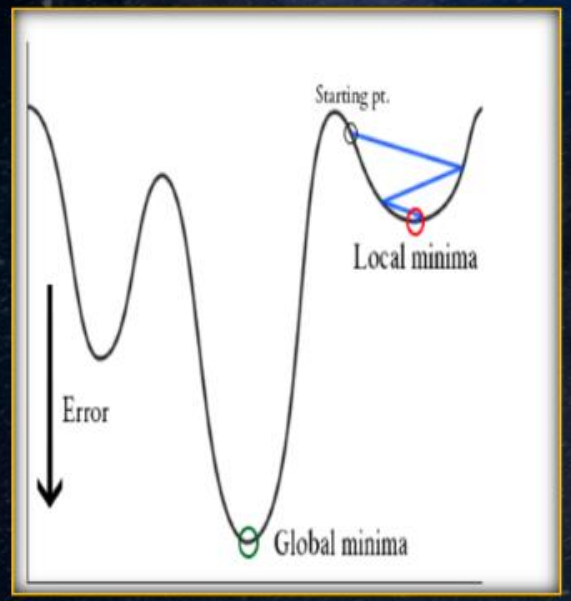
3 解决方案

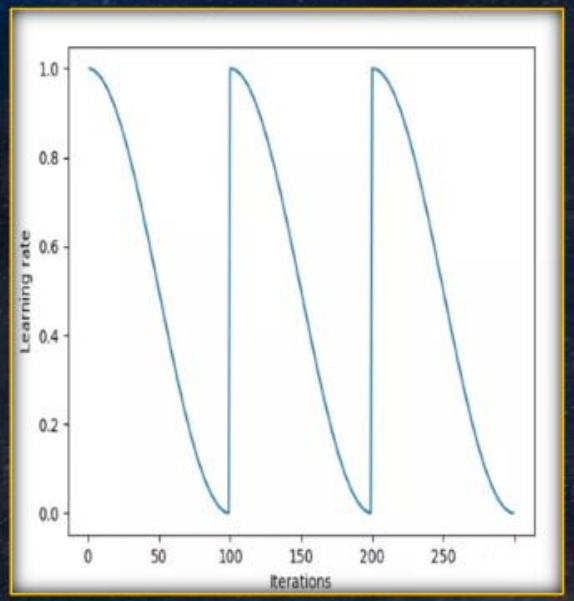
XIAMEN UNIVERSITY Tsinghua University 快手

19 Features	Seq_length（从第一天注册开始）												
第一天标志（Register Day）	1	0	0	0	0	0	....	0	0	0	0	0	0
当天是否有登陆							....						
当天创建视频的次数							....						
当天action_tpye=0的行为次数													
当天action_tpye=1的行为次数													
当天action_tpye=2的行为次数													
当天action_tpye=2的行为次数													
当天action_tpye=3的行为次数													
当天action_tpye=4的行为次数													
当天action_tpye=5的行为次数													
当天在page=0的行为次数							• • • •						
当天在page=1的行为次数													
当天在page=2的行为次数													
当天在page=3的行为次数													
当天在page=4的行为次数													
当天观看video_id的总数（去重）													
当天观看video_id的总数（去除观看自己的）													
当天观看自己video_id的次数													
是否是周末													
是否是三天小长假							....						

「RNN 原始数据输入」

余弦退火+Warm Restarts



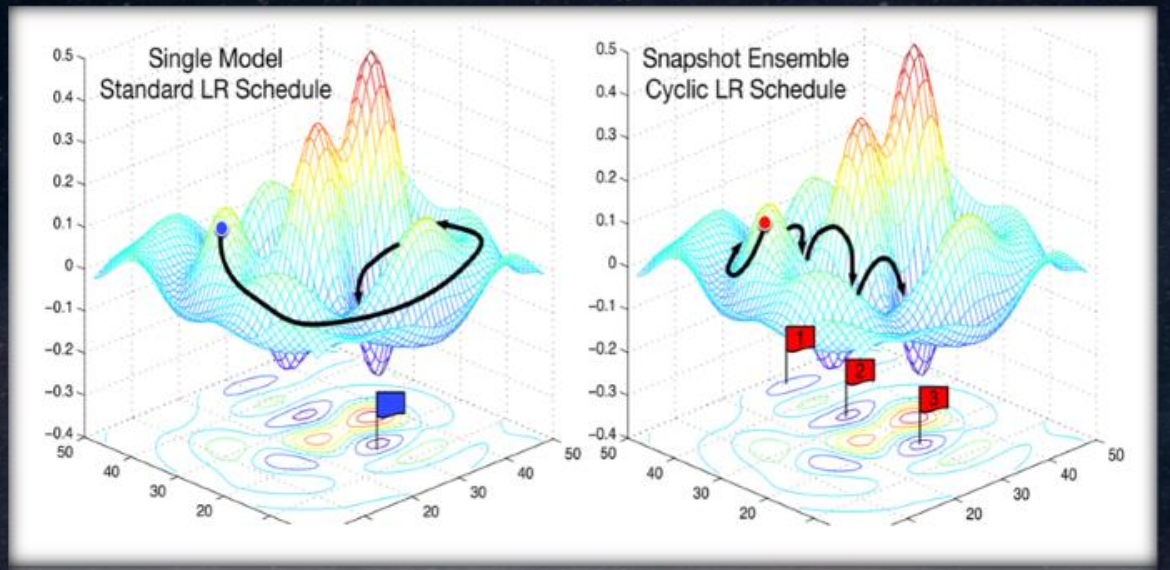


优化

应该越来越接近Loss值的全局最小值时，学习率应该变得更小来使得模型尽可能接近这一点。突然提高学习率，来"跳出"局部最小值并找到通向全局最小值的路径。使用热重启的学习率退火也叫做循环变化学习率（循环长度是20到40个epoch），最初由Smith[1]提出。

[1]. Smith, Leslie N. "Cyclical learning rates for training neural networks." In Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on, pp. 464-472. IEEE, 2017.

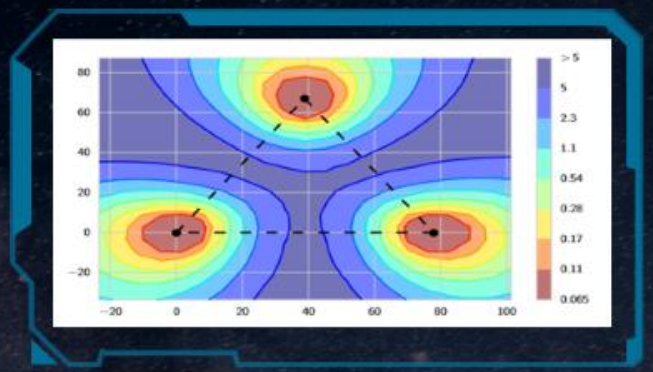
快照集成（Snapshot ensembles）

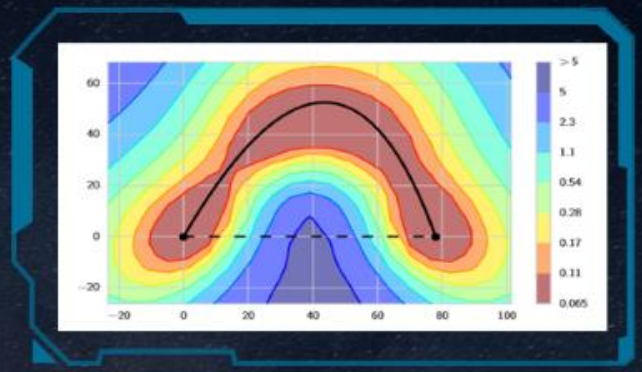


Huang[1]2017年提出快照集成：在训练单个模型时使用热重启来组装一个集合，且基本上无额外的代价。

[1]. Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E., & Weinberger, K. Q. (2017). Snapshot Ensembles: Train 1, get M for free. In Proceedings of ICLR 2017

快速几何集成（FGE：Fast Geometric Ensembling）





FGE使用线性分段循环学习率策略代替余弦；

FGE的循环长度更短——每个循环只有2到4个epoch。（短循环）

由于在足够多的不同模型间，存在低损失的连接通路，沿着那些通路，采用短循环是可行的，而且在这一过程中，会产生差异足够大的模型，集成这些模型会产生很好的结果。因此，与快照集成相比，FGE提高了模型的性能，每次循环经过更少的epoch就能找到差异足够大的模型（这使训练速度更快）。

[1]. Garipov. 2018. “Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs

3 解决方案

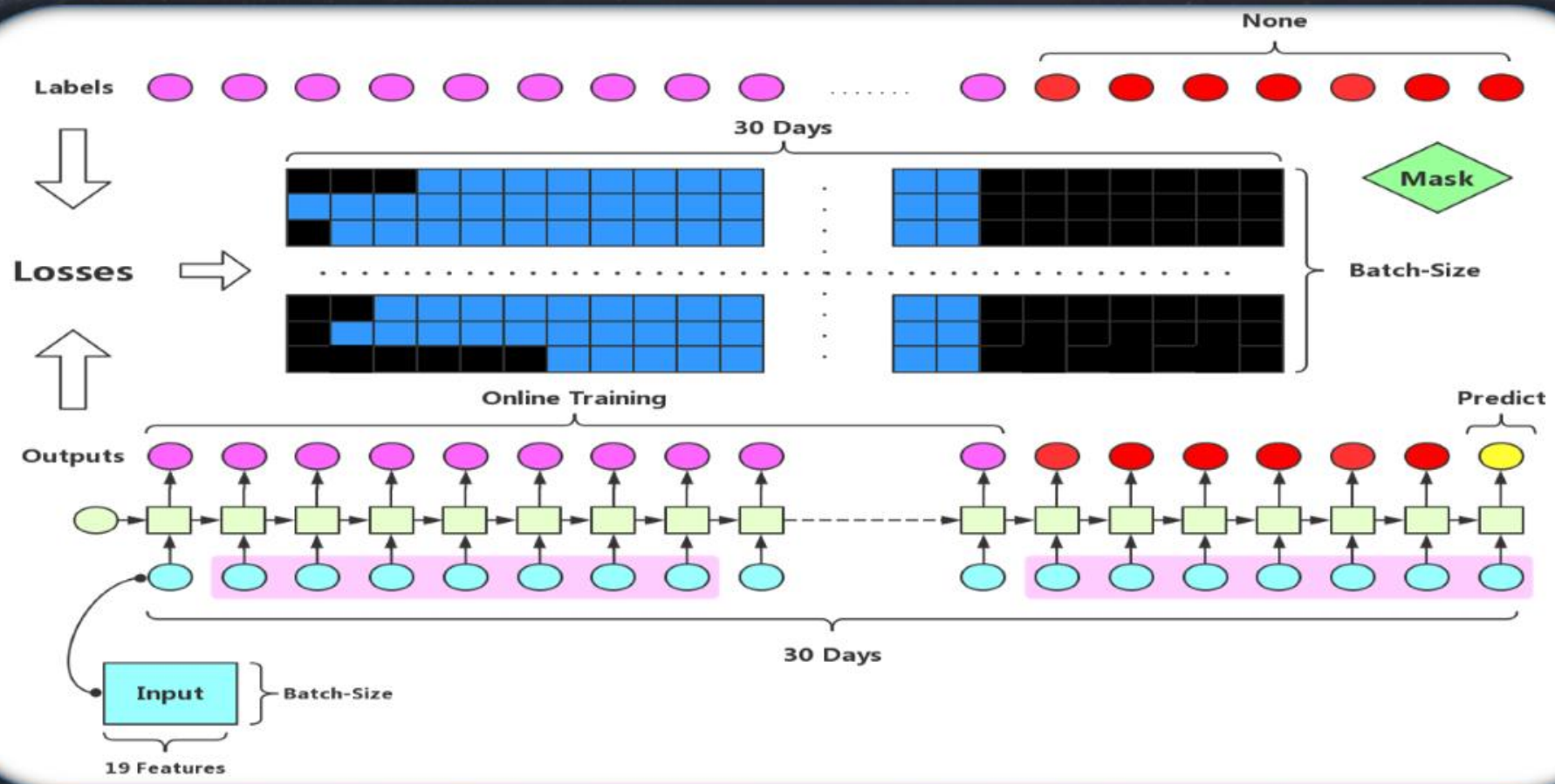
厦门大学 XIAMEN UNIVERSITY　清华大学 Tsinghua University　快手

数据集没有充分的打散（shuffle）：

即使使用了采样策略，但为保证长度一样，使得每个batch都是同一天注册，会导致某些特征集中出现，而导致的有时学习过度、有时学习不足，使得下降方向出现偏差的问题.在使用自适应学习率（Adam）算法的时会更加明显。

改进：进行padding，所有数据长度为30，可将数据集充分打散

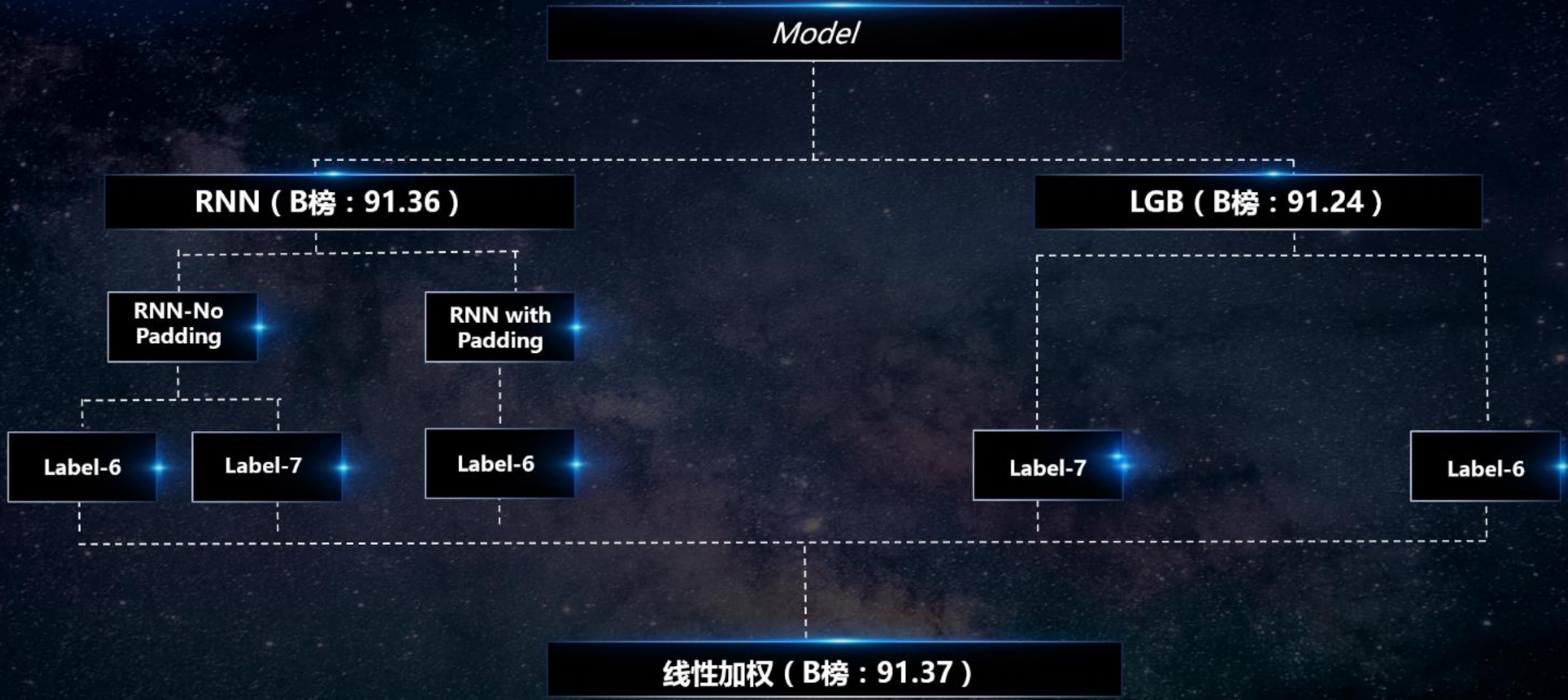
RNN(Padding+Sampling)



LGB模型

窗口划分	用户/特征	标签	特征
线下训练（调参）	1~9 1~10 1~11 ..... 1~15 1~16	10~16 11~17 12~18 ..... 16~22 16~23	最后一次登录/创作视频/分action_type行为/分page行为等距离现在天数及其相应的行为次数； 最后二次登录/创作视频/分action_type行为/分page行为等距离现在天数及其相应的行为次数； ..... 最后五次登录/创作视频/分action_type行为/分page行为等距离现在天数及其相应的行为次数； 登陆天数一阶/二阶差分的统计量； 创建视频一阶差分统计量； .....
线上训练	1~9 1~10 1~11 ..... 1~22 1~23	10~16 11~17 12~18 ..... 23~29 24~30	
线上提交	1~30	Predict	

3 解决方案


Model
RNN（B榜：91.36）
LGB（B榜：91.24）
RNN-No Padding
RNN with Padding
Label-6
Label-7
Label-6
Label-7
Label-6
线性加权（B榜：91.37）

厦门大学 XIAMEN UNIVERSITY

清华大学 Tsinghua University

快手

方案总结

RNN vs LGB

后期优化

RNN vs LGB

	RNN	LGB
调参难度	参数较多	相对容易
窗口划分	无需滑窗	需滑窗
制造模型差异性	无成本	较困难
数据量	大数据量RNN优势更明显	

4 方案总结

厦门大学 XIAMEN UNIVERSITY　清华大学 Tsinghua University　快手

1.采样策略的优化：
考虑节假日

2. 其他融合方法

3. 前期用Adam，享受Adam快速收敛的优势；后期切换到SGD，慢慢寻找最优解[1]。

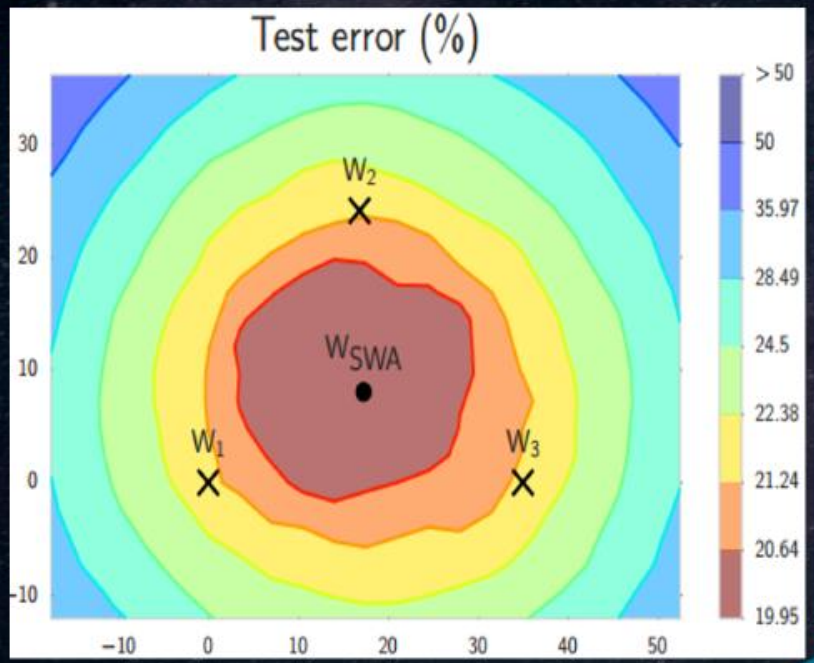
4. 使用随机加权平均(SWA，Stochastic Weight Averaging)[2]

[1]. Nitish Shirish Keskar, Richard Socher (2017). Improving Generalization Performance by Switching from Adam to SGD
[2]. Izmailov (2018). Averaging Weights Leads to Wider Optima and Better Generalization

4 方案总结

XIAMEN UNIVERSITY Tsinghua University 快手

SWA



1. 每次学习率循环结束时产生的局部最小值趋向于在损失面的边缘区域累积，这些边缘区域上的损失值较小（W_1，W_2和W_3）。通过对几个这样的点取平均，很有可能得到一个甚至更低损失的、全局化的通用解（上面左图上的W_{SWA}）。**即在权重空间而不是模型空间对这些点进行平均。**

2. FGE集成对k个模型集成的测试预测需要k倍的计算。但SWA可以被解释为FGE集成的近似值，且只需单个模型的测试时间。

3. 相较于SGD，SWA能够使所取得的解在本质上具有更宽泛的优化。SGD一般收敛于最优点的宽阔平坦区域边界附近的点；此外，SWA能够找到一个位于该地区中心的点。

$$W_{swa} \leftarrow \frac{W_{swa} \cdot n_{models} + W}{n_{models} + 1}$$

W_{swa}存储模型权重的平均值。


谢谢聆听