中国大数据算法大赛-用户购买时间预测

队伍名称：朵拉公波鲁

演讲者：武天老师

2018.07.19

JDATA

目录

JDATA

1 团队介绍

武天老师-吴远皓

2010年进入清华大学精密仪器与机械学系学习，与2014年7月和2017年7月分别取得学士、硕士学位。喜欢编程，本科期间辅修应用计算机科学专业。毕业后加入上汽集团人工智能实验室，目前从事机器学习和深度学习相关算法研发工作。多次获得天池算法比赛top1%

武泰斗-丁文博

20岁毕业于浙江大学自动化专业，后赴德国达姆施塔特工业大学深造，获硕士学位。有丰富的控制论和机器学习实战经验，曾参与德国Ko-HAF项目并完成基于传统机器学习算法和卷积神经网络的驾驶员模型的相关研究。2016年12月回国加入上汽集团人工智能实验室，目前从事机器学习和深度学习相关算法研发工作。

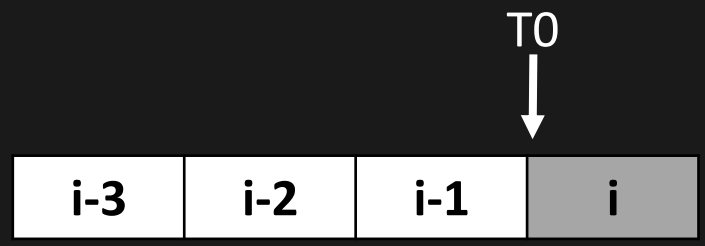
目录

JDATA

2 算法核心思想设计-EDA

- 仿照目标用户集的采样方式可以发现用户量随着时间衰减

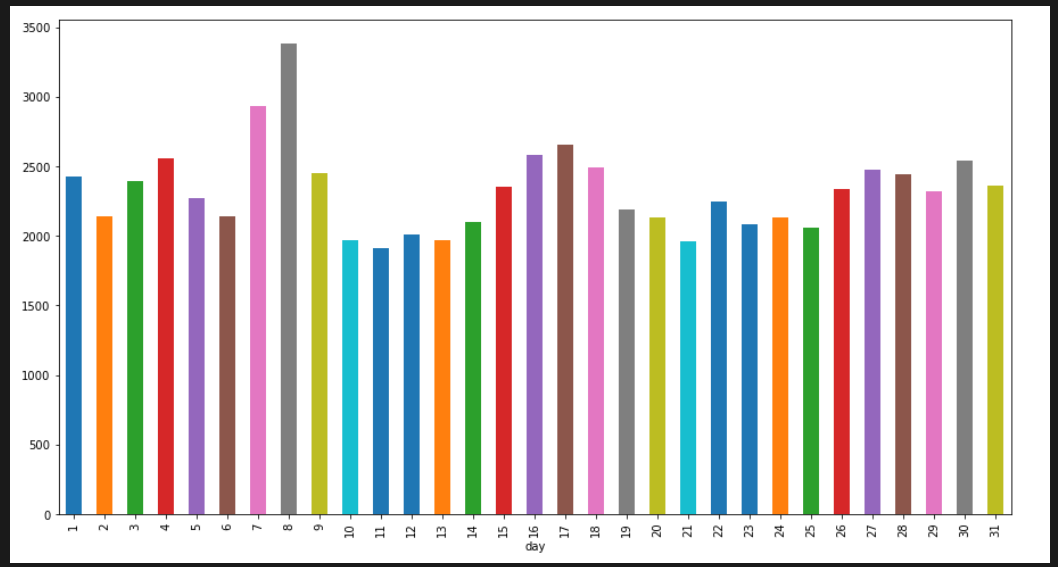
- sku复购：50%以上的sku只会被一个user购买一次，最多的一个sku被同一个用户购买了3375次
- sku买家数：75%的sku只有不超过5个user购买过
- sku购买量：75%的sku被购买不超过3次

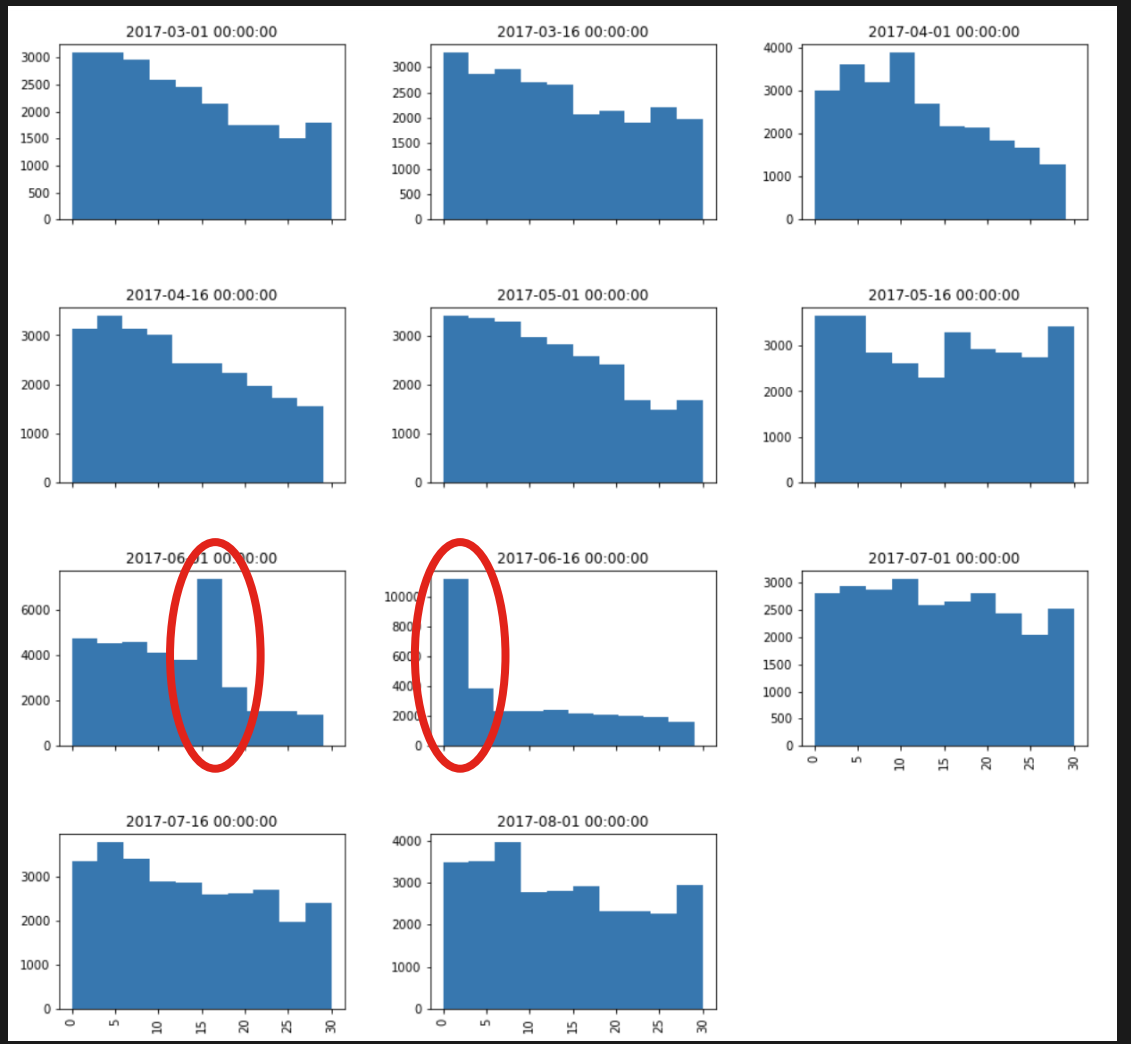


时间	当月下单用户数	前三月购买用户数	当月用户占比
2016-10	8607	15061	0.571476
2016-11	10513	23832	0.44112957
2016-12	13162	28214	0.46650599
2017-1	15222	29800	0.51080537
2017-2	17844	31249	0.57102627
2017-3	22265	36626	0.60790149
2017-4	22285	43845	0.50826776
2017-5	24685	48135	0.5128285
2017-6	31589	51422	**0.61430905**
2017-7	34416	63499	0.54199279
2017-8	34779	80133	0.43401595
合计	235367	451816	-

2 算法核心思想设计-EDA

- 可以看出首单集中在上中旬，618对首单分布有巨大的影响
- 按日期分组统计发现每天都有订单发生，且分布比较均匀



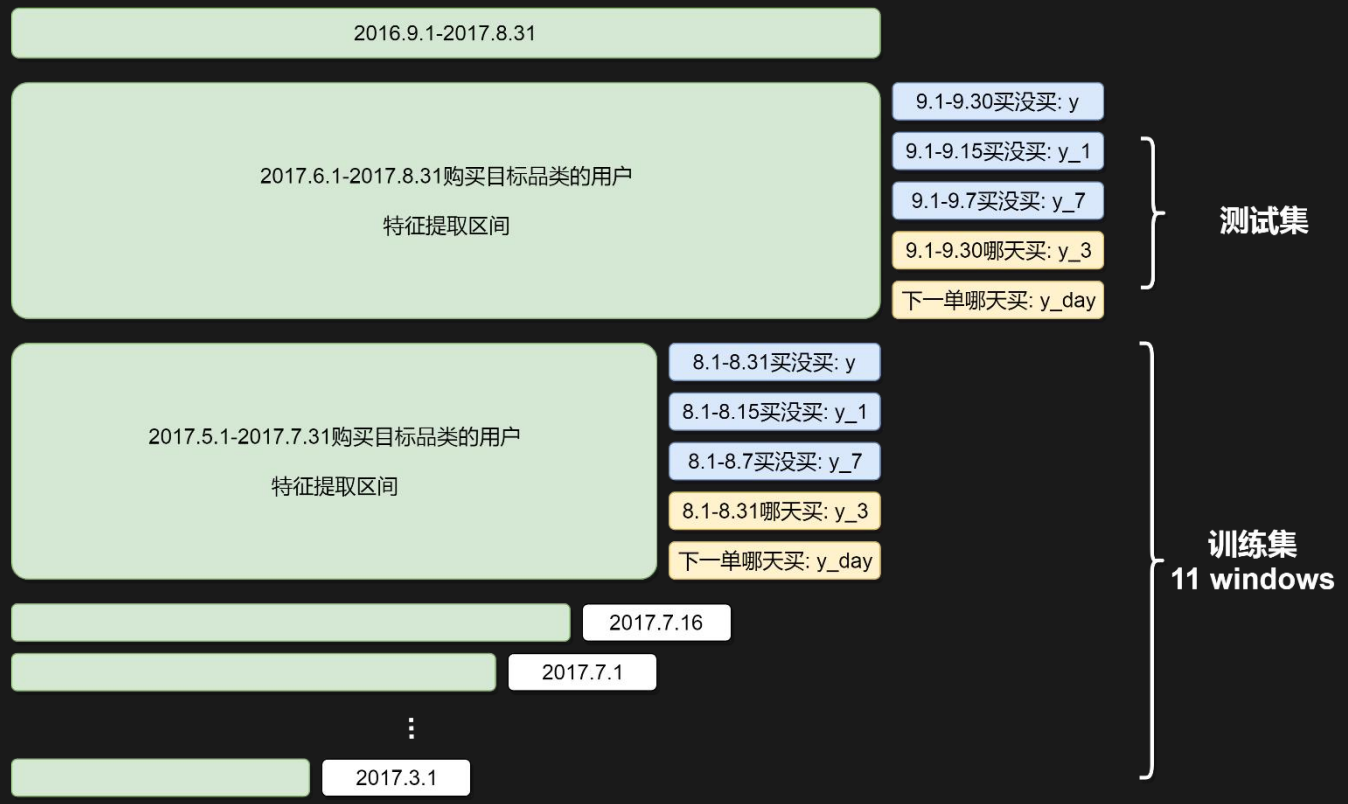


2 算法核心思想设计-核心问题

1. 最靠近测试集的三个月数据的泄漏问题
2. 如何获得更多的训练数据
3. 如何利用数据尽可能多地提供label信息
4. 先做S1还是先做S2
5. 如何描绘SKU之间，User之间的联系
6. 如何更充分地描绘出用户前后行为的联系

2 算法核心思想设计-数据集划分

- 滑窗增加数据量，步长为半个月，共滑了11个窗口作为训练集
- 每个窗口提取5个label，增加信息量
  - 未来1个月内是否购买目标类目商品
  - 未来15天内是否购买目标类目商品
  - 未来7天内是否购买目标类目商品
  - 未来一个月内首单下单时间
  - 下一单的下单时间



2 算法核心思想设计-特征提取

特征工程时候我们有以下几个主要出发点：

1. 将时序信息拍平。即将用户在不同时间的行为进行统计或直接抓取做成特征。
2. 在不同的时间范围提取特征。有的侧重近期行为，有的侧重平均行为。
3. 描绘用户的购物习惯。包括用户喜欢什么东西，喜欢以什么节奏买东西，浏览商品和最终购买商品有什么关联等。
4. 为S2提供足够多的“直接特征”。根据以往的经验，可以通过给模型提供“参考结果”来提升性能。在特征中即为各种不同的上次购买+购买间隔得到的S2预测值。

2 算法核心思想设计-特征提取

- **User/SKU Embedding**
  - 将User看过买过的所有SKU_ID作为词组成document，利用LDA获得10维Embedding向量。对SKU做类似操作。
  - 可以描绘相似的User/SKU

- **SKU 聚类 → Subcate**
  - Cate的范围太广，101cate包含33232种商品，SKU间联系松散（推测）
  - 将Embedding向量及cate特征作为输入将SKU进行聚类获得子类目（subcate）
  - 更加符合业务逻辑
  - 由于前期考虑了user-subcate 模型，subcate数设置为30

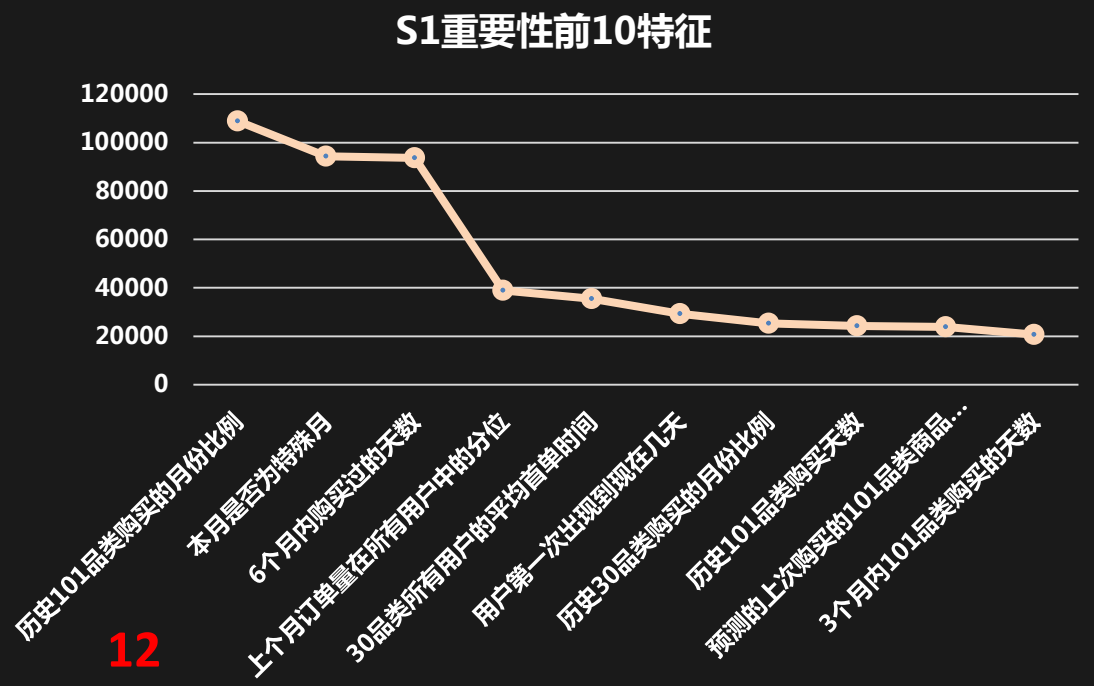
2 算法核心思想设计-特征提取

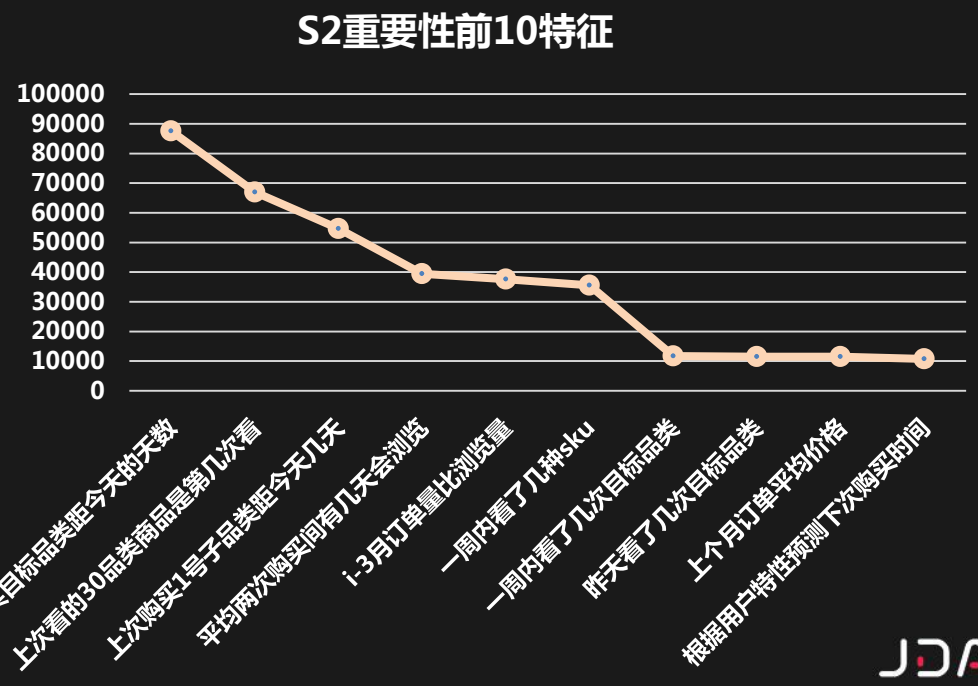
- **构建Session，获取浏览与购买的联系**
  - 将用户的行为按照购买进行切分，将两次购买之间的浏览收藏行为以及同一天内的购买行为进行提取
  - 看过的SKU、在哪天看的这些SKU、看了几次
  - 进一步获得最终购买的商品在他浏览过的商品中价格处于什么分位数，最终购买的商品是他看的第几件商品等特征

- **将最近浏览/购买的行为作为特征**
  - 最近5次浏览商品的subcate_ID、最近5次购买的时间等等
  - 可以结合购买间隔获得预测购买时间

2 算法核心思想设计-特征提取

- 总共构建了多月行为、单月行为、单周行为、单日行为、评论行为、session、user_sku交叉等14类特征，共685维
- S1和S2使用完全相同的特征集合

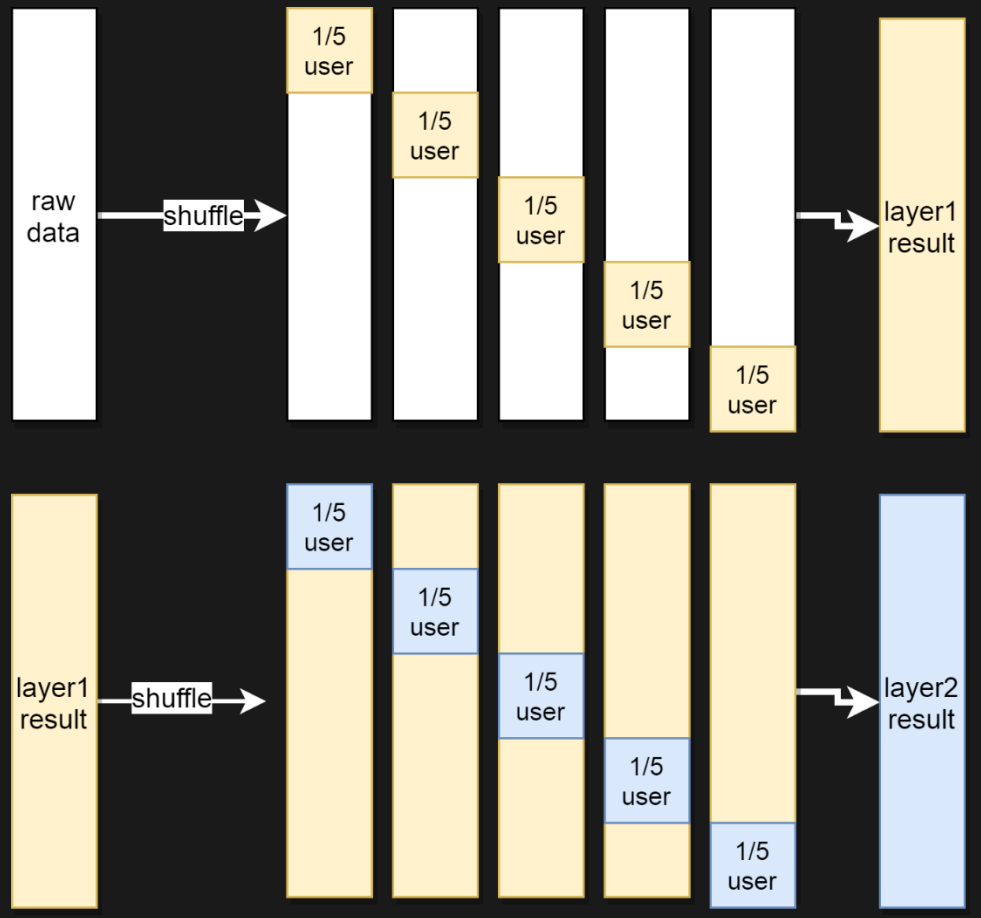




JDATA

2 算法核心思想设计-模型框架

- S1和S2的结果可以互相作为参考
- 用**两层堆叠（stacking）**模型搭起两个问题间的桥梁，完美解决先做哪个的问题
- 在layer1，利用相同特征及3个不同的S1 label训练了**3个S1模型**，利用相同特征及2个不同的S2 label以及3种不同的目标函数（MSE，MAE，S2）训练了**5个S2模型**
- 训练时采用对User_id的**5-fold**交叉验证



JDATA

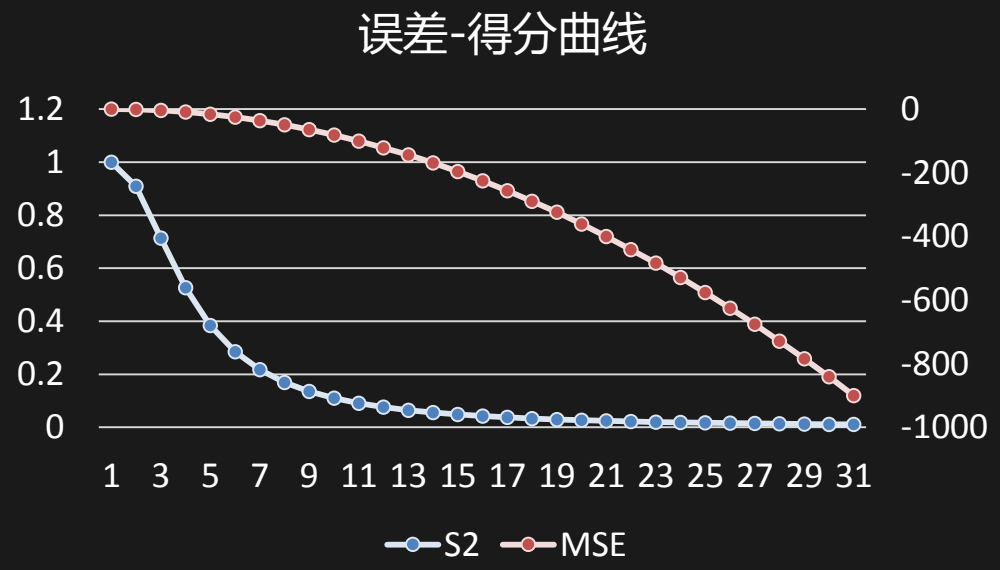
2 算法核心思想设计-训练技巧

$$S_2 = \frac{\sum_{u \in U_r} f(u)}{|U_r|}$$

$$f(u) = \begin{cases} 0, u \notin U_r \\ \dfrac{10}{10 + d_u{}^2}, u \in U_r \end{cases}$$

- 删除由于采样引起泄漏的数据
- 自定义目标函数大幅提升S2得分（0.03）

```
# S2 一阶、二阶导
def my_objective(preds, train_data):
    labels = train_data.get_label()
    d = preds-labels
    x = (10.+np.square(d))
    grad = -20*d/np.square(x)
    hess = 80*np.square(d)*np.power(x,-3)-20*np.power(x,-2)
    return -grad, -hess
```



- 删除2分类问题里可能引起泄漏的overlap数据

JDATA

目录

JDATA

3 比赛经验总结

- 本次比赛的关键点：
  - 对题目的数据抽样方式和评价指标有比较好的理解，并对这两个信息进行了比较充分的利用。首先我们使用了**按用户的5折交叉验证**，**删除了会产生过拟合的数据**，另外我们实现了**直接优化S2分数的目标函数**，大幅提高了成绩；
  - 用**两层模型**结合**多label数据集**充分利用了两个子问题的信息；
  - 提出**subcate**的概念，它比cate更加细致，又比sku更加紧凑。可以通过不同的聚类中心数调节细分程度，把类似的sku放到一个整体框架中考虑，符合业务逻辑；
  - 提出**session**概念，将订单以及订单间浏览进行整合，使信息的联系得到加强，可以获得更好的特征。

THANKS

朵拉公波鲁 为您呈现

JDATA