第一课时：赛题介绍和线下赛演示

TIANCHI天池

主讲人：李厚意（济贤）

- 毕业学校：西安电子科技大学
- 研究领域：网络编码，数据挖掘
- 个人经历：广州交通大数据竞赛季军，淘宝LBS推荐算法竞赛TOP10，目前就职于蚂蚁金服。网名潜卧终南。
- 个人邮箱：houyi.lhy@alipay.com

史上最接地气的竞赛视频

震撼来袭！！！

这赛题是干啥的？

做的好坏如何评价啊？

啥都不会，如何做这个啊？

想做的更好咋办啊？

一、干啥的

一句话描述：根据用户在手淘上的历史行为记录，预测用户们第二天买什么

- 历史行为记录是什么
- 到底预测的是什么

历史记录：谁在什么时间对什么商品进行什么操作

预测：在第32天中，谁买了什么

数据相貌：

records: 12312542
user_id: 10000
item_id: 2914411
user_item pairs: 4719002 and purchased pairs: 103464 ，CTR is 0.02192497
min time 2014-11-18 00 max time 2014-12-18 23

二、咋评价

- 如果比预测的多：那么提交所有种可能，必得满分
- 如果比预测的准：只预测一条对的，必得满分
- 因此评价指标必须要兼顾到“多”和“准”

Example:
假设第31天（12月19号）发生了20000次交易，我们预测了10000条，预测对了1000条
准确率为：1000/10000=0.1
召回率为：1000/20000=0.05
F1=2*0.1*0.05/(0.1+0.05)

三、咋做

我们先从业务常识去判断，什么样的user-item pair，可能在下一天中发生“购买”:

用户i不停的去看商品j，预测： USERi-ITEMj

用户i将商品j放入购物车，预测：USERi-ITEMj

用户i非常喜欢买东西，预测和USERi相关的所有PAIR

商品j这几天卖得很好，预测和ITEMj相关的所有PAIR

……

条件概率

$$P(H|A)$$

根据三年二班成绩表，预测现在谁在打游戏

姓名	性别	成绩	姓名	性别	成绩
老王	男	70	小强	男	90
婷婷	女	85	小秋	女	85
佳佳	女	90	月月	女	80
翔翔	男	85	南哥	男	70
弃弃	男	90	岚岚	女	85

老湿说：男孩子更爱打游戏

妈妈说：成绩差的很可能就是打游戏打的

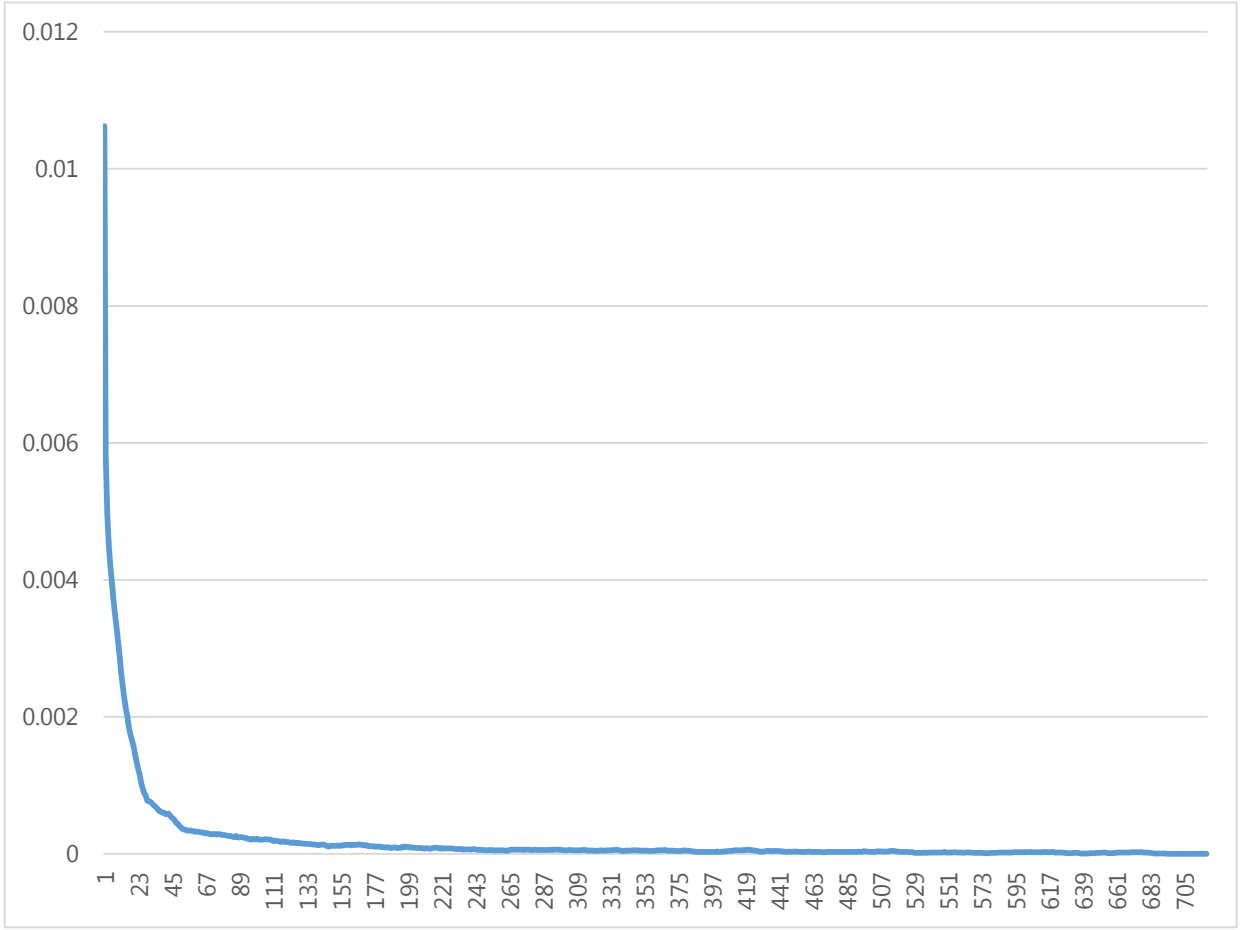
咋做：将业务逻辑转化为规则

规则：找到一个条件概率函数，将其最高的那段取出来

一个有意思的贝叶斯定理介绍：http://mindhacks.cn/2008/09/21/the-magical-bayesian-method/

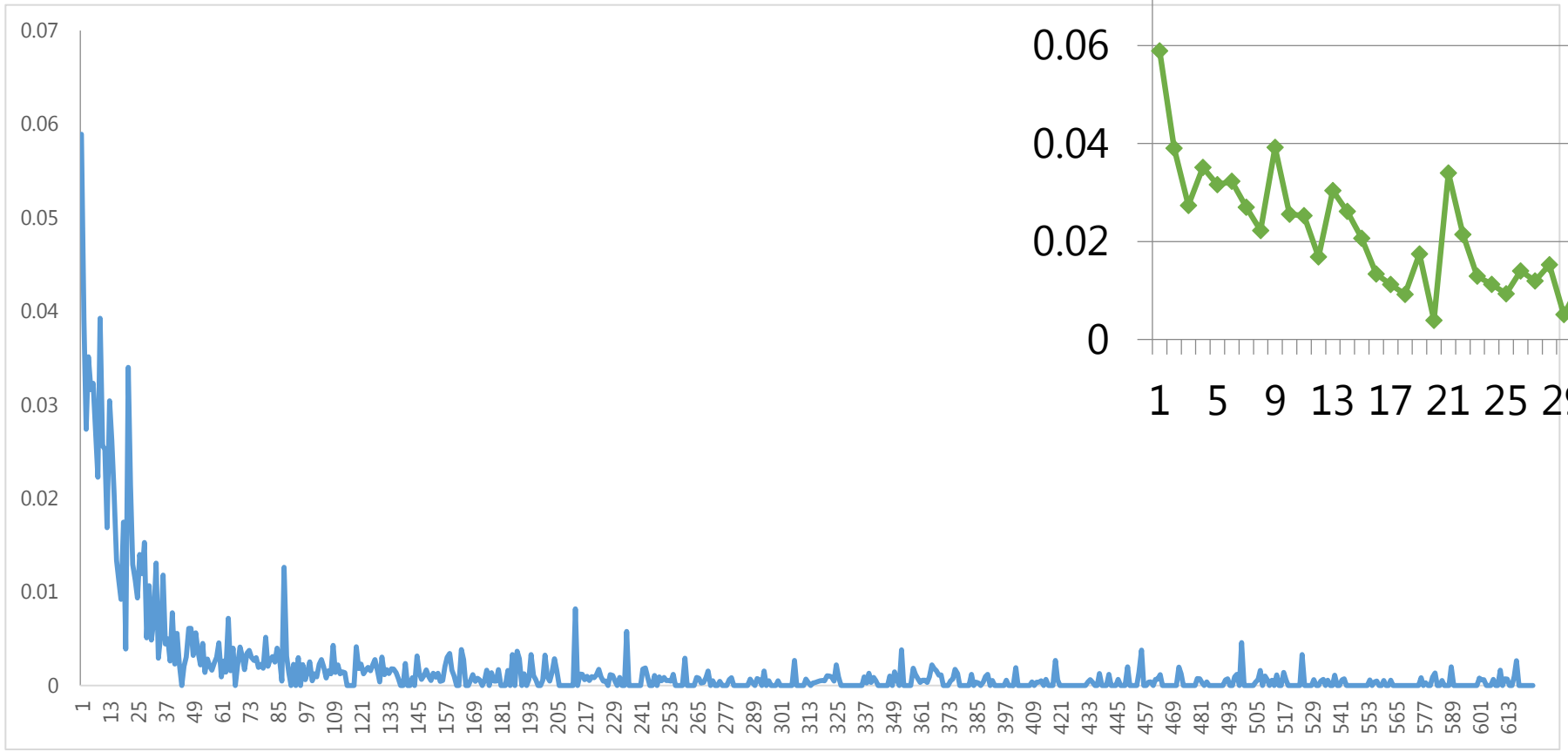
一个简单的分类器Naïve Bayesian Model：
http://www.cnblogs.com/jerrylead/archive/2011/03/05/1971903.html

条件概率函数：基于最大似然的估计



- 通过Nshifts获得更多的样本以增加最大似然估计的准确性
- 120shifts获得遗忘曲线（条件概率函数）

条件概率函数



2015-04-22 21:29:12 9.966025% 9.185804% 10.891089% 62

四、咋做好

如何将多个规则结合在一起:

1. 按照每一个规则是否达标进行打分，打分最高的当作预测

2. 自动的确定分数：Logistics Regression

3. 如果诸多规则和目标之间的关系是非线性的咋办： More Machine Learning Algorithm

请见下一节课....