1. 用户特征

（1）历史消费行为：  
 i) 消费券次数/总的消费次数

ii) 折扣价格/总的价格

折扣率的平均值，券满减比例的平均值，作为两个特征

iii) 不用消费券的最高的折扣率 /统计所有的人

iv) 购买距离的各档比例

10个特征

/\*v) 该用户对某商品的关注度

对同一个

连续几天都点击同一个物品的消费券 \*/

vi）该用户对某商店的关注度

/\*点击+购买+领取

有券在某店的购买记录+没券在某店的购买记录

没券在某店的购买记录\*/

分母都是总的

分成两个：有券的和没券的分别统计

6) 用户等级：

初步做法，把各个折扣率和券的额度满减分别作为特征

具体划分看数据。如果折扣率和券的个数有限，直接每一种作为一个特征，否则相近的几种合并。另外是优惠券众数的统计。

折扣率到一定程度才买。根据折扣率划分等级

券的额度/满减

7) 购买行为是否集中（后期，当前暂时不考虑）

是不是会出现问题：某个时间段买很多

8）领取券和使用券的时间差值

先平均

9）对各档的商户访问的频率

10）常访问店，

每个用户常不用券访问的店的并集。

每个店作为一个特征，然后如果对man 1, ，是他常访问的店，该特征值为1否则为0

11）用户领券不用的比例（后期看有没有必要拆成针对各个档位商店）

12）对用户消费的时间计算比例，按照上中下旬

二、消费券

（1）只点击比例/（15天内）购买比例/（15天外）购买比例/领取不购买的比例

（3）Discount\_rate

1. 对消费券分级，各级分开统计（后来再做）

根据满减的比例等

1. 券被使用的比例以及方差（每个月的情况的方差）（后期改成各个消费者等级的券的使用比例）

测试数据里面出现了，没有历史记录的，那么就按照平均值计算

三、商户特征

（1）距离user的距离 – 有可能没有信息

距离为i档的用户的个数/该店总的用户的个数

距离为i档的消费记录的个数/该店总的消费记录的个数

（2）使用的消费券的数量（额度）/所有商店使用的优惠券的数量(额度)

（3）商户的热度，最近1个月卖出的数量

各个月平均的消费的比例，销售曲线

Pi=第I个月消费件数/第I月所有店消费记录个数

求pi的平均值，方差，最近一个月的pi

（4）客户群

各档客户的比例

（5）//消费券使用率，领取率

（6）商户卖的东西是不是常见

根据同一个人一定时间内买的次数。。

求每个用户消费次数最多（大家都去过）的N个店，求交集

去过的用户/（距离区间内的总用户）（参考df,停用词和面包牛奶）

（7）店的货物质量

该店的销售曲线 – 梯度求平均值

（8）没有购物券的时候的被访问频率，根据这个对商户分档

被访问次数/总的记录

（9）发的各档券的比例

档和用户对券划分一样，按照折扣比例和满减比例分档

========================================================================================================

四、商品特征

1. /\*商品的热度，最近多少人买。复杂化就是，考虑最近的人有没有用消费券去买。
2. aparior 商品和用户的关联度\*/
3. 该商品的消费券的使用比率

（4）稀缺商品还是普遍商品

同一个人买的次数/时间

特征离散

数据预处理：

1、重复数据

2、垃圾店

和无消费记录的人，直接赋值，不参与训练

垃圾券，几乎没人用

3、处理15天的数据，15天外的使用/总的使用记录 的比例

4、统计各种优惠券，如果<100就逐个考虑，否则按规则分类

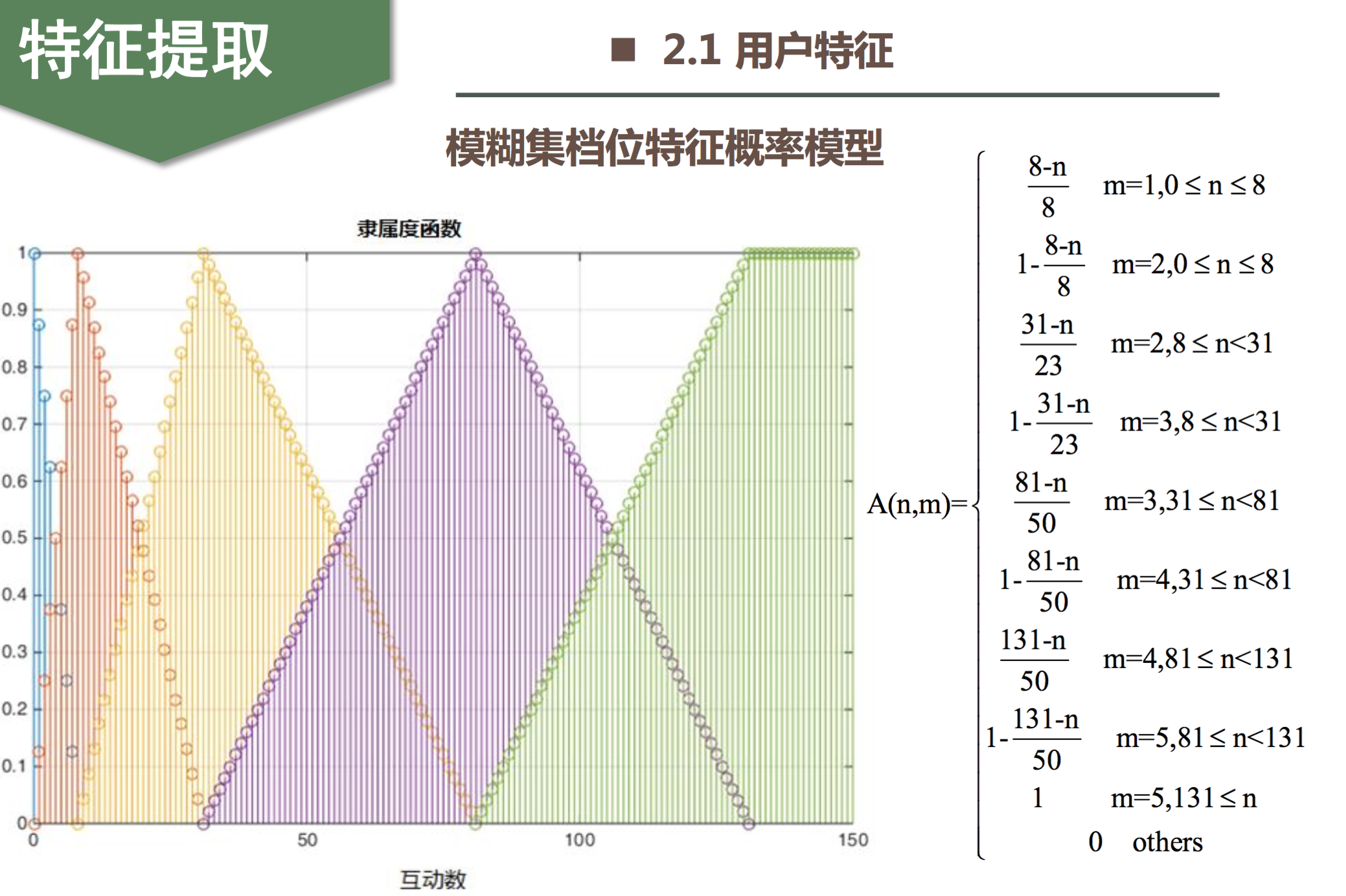
5、券是否通用，考察券和店的关联（有该店券但是没用）

有这家的券但是这次没用，后来消费了这家店，但是没用券

6、一定距离区间内，去这家店的人数占这家店区间内总人数的比例

7、横轴是距离，纵轴是去的人次，商店不做区分

这个曲线应该是急剧下降，取占比例很大的距离



二期：

1. 用户的相关性，地理相关性，商店偏好相关性，购物券相关性
2. 训练样本的时间维度选取，是否只选择最后一个月的数据
3. 用户购买行为之间的关联度，根据关联规则，很多店都有同样的一批用户，则这些用户关联