**电视广告分户推送与营运规划**

摘要

本文针对电视广告分户推送与广告竞价交易问题，运用因子分解机与深层神经网络（DeepFM）和协同过滤算法，建立了历史记录和用户及广告特征的推荐算法模型，和保证电视台利益最大前提下提升收视率与销售量的竞价交易模型，使用MATLAB/python/C++编程得出了合理结论和解决方案，最终对模型准确度和合理性进行了分析。

针对问题一要求，依据用户-广告数据集建立回归模型研究具有不同特征用户对不同特征的广告的偏好，并给出了具有相似特征用户对指定特征广告的偏好函数，使用DNN+FM的DeepFM模型建立了回归分析。在问题二中，我们根据CCTV1的分时段广告竞价底价，首先建立了收视率和广告底价之间的函数关系，利用收视率的历史纪录，建立了收视率-时段模型。在第三问中，我们首先通过作图分析了第一问基于特征工程的回归模型的不足之处，使用协同过滤算法针对用户历史记录和极大似然估计建立了新的综合推荐模型。在第四问中，我们考虑了电视台收视率，电视台广告收入，竞价方案对电视台利益的关系，提出了长期利益函数，由于利益函数中performance()复杂度过高难以求解，提出了一种简化方案进行拟合，建立了基于广义第二密封竞价的竞价交易模型。

关键词：广告推荐算法，特征提取，协同过滤算法，极大似然估计，广义第二密封竞价，

**目录**

1. 问题背景
   1. 相关数据
   2. 问题概括
2. 问题分析
3. 模型假设
4. 名词解释及符号说明
5. 模型建立与求解
   1. 数据预处理
   2. 问题一的分析与求解
      1. 问题分析
      2. 模型I ：DeepFM模型建立与求解

5.3问题二的分析与求解

5.2.1问题分析

5.2.2模型II：时间序列分析模型建立与求解

* 1. 问题三的分析与求解

5.3.1问题分析

5.3.2模型III：协同过滤与当前最热估计的推荐模型

5.3.3 用户点击数分布拟合

5.3.4 用户历史数与协同吗过滤精度函数

5.3.5 联合推荐模型

5.4 问题四的分析与求解

5.4.1 问题分析

5.4.2 周期利润模型的建立与求解

6 模型检验

6.1模型I算例及检验

6.2模型II算例及检验

6.3模型III算例及检验

6.4模型IIII算例及检验

参考文献

附录：

# 1.问题背景

## 1.1 相关数据

### 1.1.1 用户-广告数据

本文数据来源于2019腾讯广告算法大赛的数据集。数据集为历史n天的曝光广告的数据（特定流量上采样），包括对应每次曝光的流量特征（用户属性和广告位等时空信息）以及曝光广告的设置和竞争力分数；测试集是新的一批广告设置（有完全新的广告id，也有老的广告id修改了设置），要求预估这批广告的日曝

光 。）

数据分为用户特征属性和广告数据文件，结构如下：

#### 1.1.1.1用户特征属性文件

注：用户特征字段未知的均使用 0 表示。每列特征取值都用加密后的 id 表示，均

为随机映射。不同列的id取值区间会重复。各字段使用制表符(\t)分隔，每列的具

体业务含义如下：

用户id：此处和上面曝光日志文件中的用户id关联

年龄（Age）：每个取值随机映射为[1-N]的唯一id

性别(Gender)：男/女

地域(area)：每个省/市用唯一id标识，可能多标签，使用逗号分隔不同id

婚恋状态（Status）：单身/已婚等状态，可能去多值，使用逗号分隔

学历(Education)：博士/硕士/本科/高中/初中/小学

消费能力（ConsuptionAbility）：高/低

设备（device） ：IOS/Android, 不区分版本号

工作状态（work） ：在校大学生/商旅人士/政府公职人员/科研教育者/ IT互联

网工作者/医护工作者, 可能取多值，逗号分隔

连接类型(ConnectionType)：无线/2G/3G/4G

行为兴趣(behavior)：每个兴趣点一个id，可多值，逗号分隔

#### 1.1.1.2广告数据文件

广告是指广告主创建的广告创意（或称广告素材）及广告展示相关设置，包含广

告的基本信息（广告名称、投放时间等）、广告的推广目标、投放平台、投放的广

告规格、所投放的广告创意、广告的受众（即广告的定向设置）以及广告出价等

信息。

a. 广告静态数据

该类广告属性一般从广告创建后无法修改。所有id类数据均为加密后随机映射。

各列用制表符分隔，含义如下：

广告id：和曝光日志中的广告id相关联

创建时间：广告创建时的时间戳

广告账户id：广告所在账户的唯一标识，账户结构分为四级：账户——推广计

划——广告——素材

商品id：广告推广目标的唯一标识，若推广目标是落地页，则该字段为空

商品类型：广告推广目标的类型，枚举型

广告行业id：广告所属的行业类别标识

素材尺寸：不同广告位对素材的尺寸要求不同，同一个广告可能有多个不同尺

寸的素材，用逗号分隔

### 1.1.2 收视率数据

本研究通过文献调研，寻找到了对于电视广告定价影响最大的因素，分别是：收视率、满意度和覆盖率。在不同时段内广告竞价底价的主要影响因素为收视率，而满意度和覆盖度随时段的变化极小。故在建立分时段的广告竞价底价的模型时，只选取收视率作为变量，将满意度和覆盖度视为常数。

本研究通过查找到CCTV1的广告竞价底价随时间的变化的数据和在3月1日到4月25日CCTV1在八个时间段的收视率。根据CCTV1在不同时间段对广告的定价，我们将一天的时间分为八个时间段，分别是：凌晨时段、早晨时段、上午时段、中午时段、下午时段、傍晚时段、黄金时段、后黄金时段。利用3月1日到4月25日的不同时段的收视率的均值，对不同时间段的广告底价定价进行拟合。

具体时间段的划分如表一所示，收集到的CCTV1不同时间段的广告定价如表二所示。不同时间段收视率的均值如表三所示。研究中所用到的数据集的来源为中央电视台官网。

|  |  |
| --- | --- |
| 凌晨时段 | 0：00-6：00 |
| 早晨时段 | 6：00-9：00 |
| 上午时段 | 9：00-12：00 |
| 中午时段 | 12：00-14：00 |
| 下午时段 | 14：00-17：00 |
| 傍晚时段 | 17：00-19：00 |
| 黄金时段 | 19：00-22：00 |
| 后黄金时段 | 22：00-24：00 |

表一 时间段的划分

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 5秒 | 10秒 | 15秒 | 20秒 | 25秒 | 30秒 |
| 凌晨时段 | 20100 | 30400 | 38000 | 51700 | 60800 | 68400 |
| 早晨时段 | 20700 | 31200 | 39000 | 53000 | 62400 | 70200 |
| 上午时段 | 25240 | 38080 | 47600 | 64740 | 76160 | 85680 |
| 中午时段 | 48500 | 73300 | 91600 | 124600 | 146600 | 164900 |
| 下午时段 | 30750 | 46425 | 58025 | 78900 | 92850 | 104450 |
| 傍晚时段 | 36300 | 54700 | 68400 | 93000 | 109400 | 123100 |
| 黄金时段 | 93400 | 140900 | 176100 | 239500 | 281800 | 317025 |
| 后黄金时段 | 46100 | 69600 | 87000 | 118400 | 139250 | 156700 |

表二 不同时间段广告竞价底价

|  |  |
| --- | --- |
| 凌晨时段 | 0.0456 |
| 早晨时段 | 0.2935 |
| 上午时段 | 0.4313 |
| 中午时段 | 0.9217 |
| 下午时段 | 0.4255 |
| 傍晚时段 | 0.3757 |
| 黄金时段 | 1.5366 |
| 后黄金时段 | 0.4170 |

表三 不同时间段收视率均值

## 1.2 问题概括

问题一概括; 根据用户特征，推荐具有某特征的广告

问题二概括; 更具电视台当前时段收视率数据，建立电视台下一时段收视率规律。

问题三概括：基于用户历史行为进行推荐，缺少历史行为时使用极大似然估计。

问题四概括： 提出保证电视台利益最大的竞价策略。

# 2.问题分析

建立基于特征工程的用户-广告推荐模型，对于具有某一特征的广告，建立回归模型，判断某特征用户是否喜欢

求解电视台收视率变化规律，根据以往收视率历史纪录预测将来某一时段收视率，进而预测该时段广告竞价底价。

求解基于用户历史行为和商品销售量的个性化推送，通过用户历史行为搜索具有相似行为的用户，进行推荐；当缺少用户数据时使用极大似然估计，从最热商品中选择推荐，用户池实时更新动态反馈;最终建立用户分布函数，求解推荐器权重。

求解广告竞价模型，站在理性电视台角度权衡收视率与当前利润，建立评估模型，使用广义二次密封竞价策略。

# 3.模型假设

基本假设1：同一用户对于指定广告在手机上看到的感受和在电视上看到的感受是相同的。

基本假设2：用户对手机上看到的广告的点击正相关地反应了用户对广告内容的兴趣。

基本假设3：用户的历史记录定义为用户在其他各种渠道和广告的交互情况。

基本假设4：一般性地，假设广告点击行为相似的用户具有相似的兴趣。

基本假设5：用户对广告的兴趣的主要贡献来源于用户对广告内容的兴趣，次要来源于广告本身形式。

基本假设6：假设用户在手机上对广告的点击正相关与用户对广告对应商品的购买量。

基本假设7：用户点击的某一广告的次数，正相关与用户对广告的兴趣。

基本假设8：具有相似显性特征的用户在某种程度上具有相似的兴趣。

基本假设9：一般性地，具有最大置信率的推荐方案是基于用户点击过的广告的选择。

基本假设10：假设某一时期内市场份额是有限的。

基本假设11：假设销售量对推荐方案的影响仅作用在按类型推荐时销量优先（基于极大似然估计）

基本假设12：竞价交易模型中所有广告主和电视台都是理性的。

基本假设13：如果在一定时间内给用户推送他并不感兴趣的广告或者广告质量低的广告，则用户会对该电视台的收视率会降低。

# 4.名词解释及符号说明

|  |  |
| --- | --- |
| **DNN** | **深层神经网络** |
| **FM** | **因子分解机** |
| ***"embedding\_size": 8,***  ***"deep\_layers": [32, 32],***  ***"deep\_layers\_activation": tf.nn.relu,***  ***"epoch": 30,***  ***"batch\_size": 1024,***  ***"learning\_rate": 0.001,***  ***"optimizer\_type": "adam",*** | **DeepFM参数，分别表示:embedding后向量的长度，神经网络全连接层节点，激活函数，周期数，每次输入的样本数，学习率，所使用的优化器，** |
| **Gini Normalization** | **模型风险区分能力进行评估。** |
| **Arma** | **自回归滑动平均模型** |
| **P** | **自回归滑动平均模型阶数** |
| **Q** | **自回归滑动平均模型阶数** |
| **多元线性回归（MLR）** |  |
| **协同过滤算法** |  |
| ***Exploitation*** | **开发推荐思路** |
| ***exploration*** | **探索推荐思路** |
| **S（x,y）** | **用户相似度** |
| **Softmax** | **函数形如** |
| **p（x）** | **用户点击量分布函数** |
| **f（x）** | **协同过滤推荐精度函数** |
| **竞价广告利润系数：** | **表示某条广告被单词点击电视台的收益** |
| **候选广告池：** | **i个独立竞价广告ad** |
| **推荐广告池：** | **j个可选的推荐给用户的广告ad** |
| **广告池：{}** | **i个数据中所有广告** |
| **用户池：God = {},** | **为长为m向量表示一个用户对m个广告的历史纪录** |
| **广告选择向量： *X*** | ***one-hot编码的，*形如[0, 1, 1, 0, 0, ……, 0, 1], 表示从中选择进入 的广告选择方案** |
| **广告表现张量**：***Performance(x, A)*** | ***为广告组合为x时，广告组合在用户池中的表现张量。现实意义为广告池中每个广告符合用户需求的次数。*** |
| **利润系数向量：*Profit*** | ***广告组合为x时，对应的单次点击利润系数向量；形如：*[,,…]** |
|  | **单位时间内电视台收视率** |
| ***Pr*** | **电视台利益** |
|  | **编号为i的广告被推荐的置信度** |
|  | **编号为i的用户被推荐的广告的列表** |
|  | ***编号为i的用户*** |
| ***=- Recommendation()*** | **联合推荐器，返回一个包含推荐广告按照置信度高低排序的列表** |
| ***L*** | ***用户喜爱度*** |
| ***A*** | ***用户接受度*** |
| ***BRE*** | ***基本竞争力*** |
| ***Q*** | ***广告质量度*** |
| ***P*** | ***出价*** |

# 5.模型建立与求解

## 5.1 数据预处理

### **5.1.1 用户-广告数据集**

选取50w条广告数据和1w条用户数据，选取用户历史behavior作为历史记录，使用桶排序O(n)，建立特征和用户对应关系。

对用户数据统计特征分析如下：

count mean std min 25% 50% 75% max

user\_id 18034 248377 144133 47 123930 248129 372281 499959

user\_age 18034 528 279 0 333 601 787 988

user\_gender 18034 2 0 1 2 3 3 3

user\_education 18034 4 2 1 2 5 6 8

user\_consuption 18034 2 0 1 1 2 3 3

user\_device 18034 2 0 0 2 2 2 4

user\_connection 18034 2 1 1 2 2 4 5

建立特征组合如下：

gender: 3 [3, 2, 1]

education 8 [7, 6, 2, 5, 8, 3, 1, 4]

consuption 3 [2, 3, 1]

device 4 [2, 3, 0, 4]

work 20 ['0', '4', '5', '2', '3,2', '3', '1', '6,1', '6', '2,4', '2,1', '6,5', '6,4', '3,6', '2,5', '2,6,4', '2,6', '3,2,6', '2,6,1', '2,6,5']

桶排序方法统计得到的广告特征偏好如下（部分）：

{'gender': '3', 'education': '7', 'consuption': '2', 'device': '2', 'work': '0', '3': 3218.0, '18': 7520.0, '1': 4426.0, '13': 1697.0, '2': 309.0, '8': 3207.0, '5': 1868.0, '15': 24.0, '12': 4.0, '14': 0.0}

{'gender': '3', 'education': '7', 'consuption': '2', 'device': '2', 'work': '4', '3': 586.0, '18': 1364.0, '1': 826.0, '13': 317.0, '2': 60.0, '8': 581.0, '5': 335.0, '15': 4.0, '12': 1.0, '14': 0.0}

{'gender': '3', 'education': '7', 'consuption': '2', 'device': '2', 'work': '5', '3': 348.0, '18': 772.0, '1': 433.0, '13': 176.0, '2': 28.0, '8': 318.0, '5': 188.0, '15': 4.0, '12': 1.0, '14': 0.0}

{'gender': '3', 'education': '7', 'consuption': '2', 'device': '2', 'work': '2', '3': 297.0, '18': 744.0, '1': 433.0, '13': 176.0, '2': 24.0, '8': 279.0, '5': 181.0, '15': 0.0, '12': 0.0, '14': 0.0}

{'gender': '3', 'education': '7', 'consuption': '2', 'device': '2', 'work': '3,2', '3': 0, '18': 0, '1': 0, '13': 0, '2': 0, '8': 0, '5': 0, '15': 0, '12': 0, '14': 0}

{'gender': '3', 'education': '7', 'consuption': '2', 'device': '2', 'work': '3', '3': 1119.0, '18': 2701.0, '1': 1576.0, '13': 610.0, '2': 110.0, '8': 1169.0, '5': 692.0, '15': 8.0, '12': 3.0, '14': 0.0}

{'gender': '3', 'education': '7', 'consuption': '2', 'device': '2', 'work': '1', '3': 1290.0, '18': 2959.0, '1': 1816.0, '13': 690.0, '2': 121.0, '8': 1237.0, '5': 791.0, '15': 12.0, '12': 5.0, '14': 0.0}

{'gender': '3', 'education': '7', 'consuption': '2', 'device': '2', 'work': '6,1', '3': 0, '18': 0, '1': 0, '13': 0, '2': 0, '8': 0, '5': 0, '15': 0, '12': 0, '14': 0}

{'gender': '3', 'education': '7', 'consuption': '2', 'device': '2', 'work': '6', '3': 109.0, '18': 220.0, '1': 158.0, '13': 57.0, '2': 8.0, '8': 96.0, '5': 53.0, '15': 0.0, '12': 0.0, '14': 0.0}

{'gender': '3', 'education': '7', 'consuption': '2', 'device': '2', 'work': '2,4', '3': 4.0, '18': 9.0, '1': 3.0, '13': 1.0, '2': 0.0, '8': 4.0, '5': 3.0, '15': 0.0, '12': 0.0, '14': 0.0}

{'gender': '3', 'education': '7', 'consuption': '2', 'device': '2', 'work': '2,1', '3': 12.0, '18': 39.0, '1': 3

### 5.1.2 收视率数据

求均值

## 5.2 问题一的分析与求解

### 5.2.1 问题分析

提取了用户的年龄，地域，工作，受教育程度，连接方式，设备类型等特征，提取了广告的尺寸，类型，广告行业类型等特征，首先根据特征对用户分类，根据同类用户对不同类

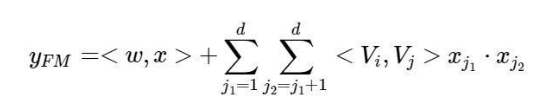
广告的偏好加权得到了该类用户-广告偏好，建立回归模型，使用DeepFM求解。

### 5.2.2 DeepFM模型建立与求解

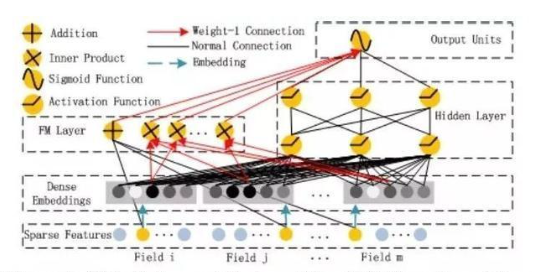
FM(Factorization Machine)主要是为了解决数据稀疏的情况下，特征怎样组合的问题。在广告分类的问题中，根据用户与广告位的一些特征，来预测用户是否会点击广告.

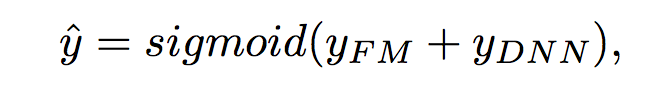
=

其中为特征，使用one-hot编码，取值为0/1。考虑到特征组合维度过高，使用deepFM算法进行求解。从而预测对应类型用户对对应类型广告的偏好，使用**Gini Normalization作为性能指标**。DeepFM使用FM算法作为因子分解机：



DNN部分使用Dense Embeddings压缩特征。

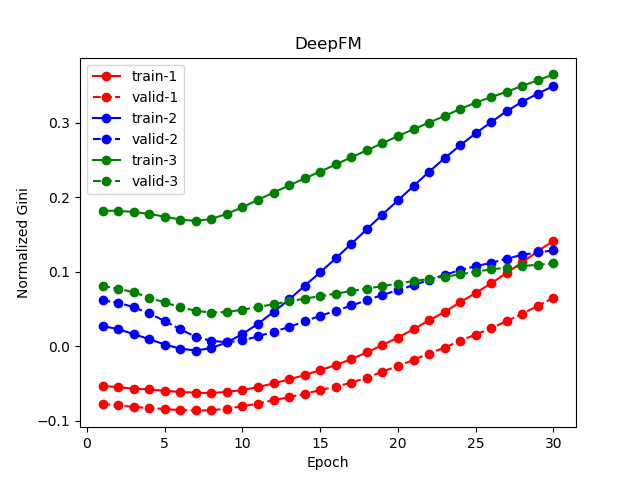




DeepFM参数见4：

训练集为1w用户偏好特征，测试集规模为2000

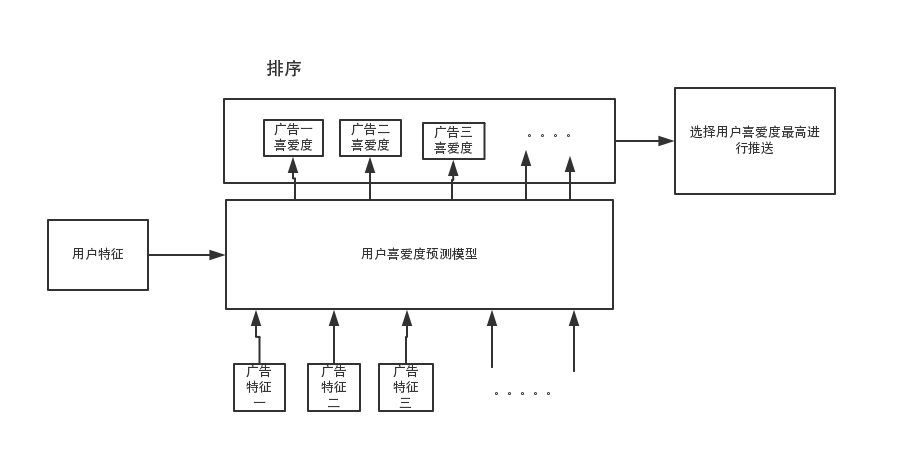
效果如下；



可见性能一般，这个问题将在第三问中详细分析。

**5.2.3 建立分类匹配推送静态模型**

根据以上的训练好的用户喜爱度预测模型，我们可以对某一特征的用户预测其对于某一特征的广告的喜爱程度。所以在向具有某一特征的用户在候选广告中推荐广告时，我们可以利用用户喜爱度预测模型对用户对于各种广告的喜爱程度预测后进行排序，将排名第一位的广告，即用户喜爱度最高的广告推荐给用户。分类匹配推送静态模型结构如下：



## 5.3 问题二的分析与求解

### 5.3.1 问题分析

本题为设计卖方分时段竞卖合理底价估算模型，所以该问题的解决方式是利用已有的历史数据分时段的对底价进行估算，影响广告分时段定价的最重要的因素为收视率，所以我们采集了CCTV1收视率的历史数据和CCTV1的历史分时段定价，首先拟合收视率和定价的关系，然后基于时间序列分析利用历史数据预测了未来的收视率。利用预测的得到的未来的收视率和收视率与定价的关系，建立出了卖方分时段竞卖合理底价估算模型。

### 5.3.2 时间序列模型建立与求解

由于收视率与广告定价有较好的线性关系，故使用线性回归对该问题进行拟合。设收视率为X，广告定价为Y，常数为b，模型公式如下：

CodeCogsEqn.gif

由于数据中存在同一时间段不同时长的广告的数据，且广告定价与广告时长并不是简单的线性关系，所以对已有的六种不同秒数的广告，分别基于收视率进行模型的拟合。

拟合得到的模型如表四所示：

|  |  |
| --- | --- |
| 五秒 | Y= 48716.7287X+13149.1555 |
| 十秒 | Y= 73502.1071X+19858.4348 |
| 十五秒 | Y= 91852.7787X+24832.9060 |
| 二十秒 | Y= 124924.2895X+33777.0034 |
| 二十五秒 | Y=146999.3028X +39725.8402 |
| 三十秒 | Y=165361.3255X+44703.3984 |

表四 分时段的价格随收视率变化

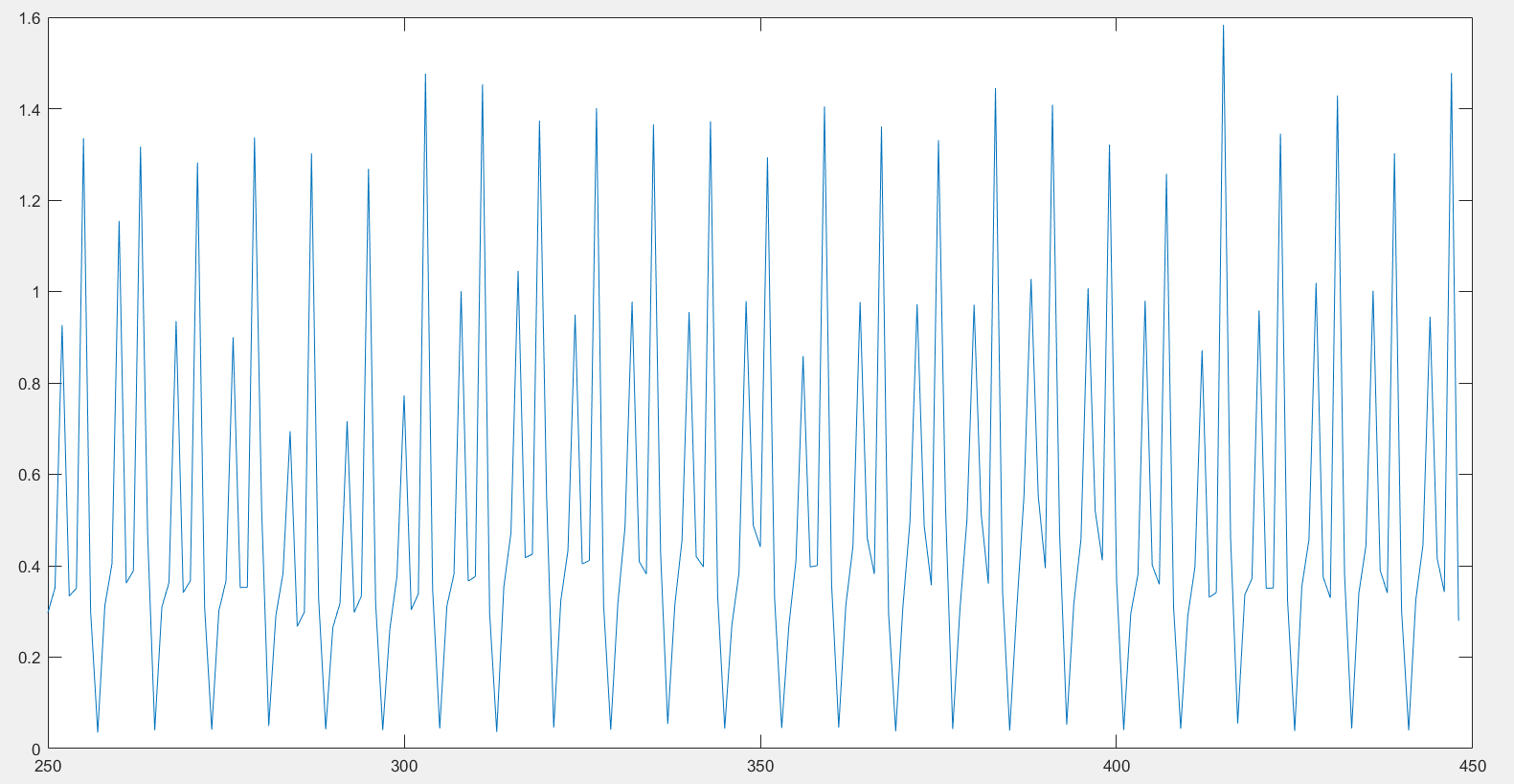
拟合得到的模型的拟合优度如表五所示：

|  |  |
| --- | --- |
| 时间 | 拟合优度R^2 |
| 五秒 | 0.9262 |
| 十秒 | 0.9263 |
| 十五秒 | 0.9262 |
| 二十秒 | 0.9260 |
| 二十五秒 | 0.9262 |
| 三十秒 | 0.9261 |

表五 拟合优度

由此可见用收视率与广告定价有很明显的线性关系，可以使用已经建立好的线性模型对已知收视率的时间段进行广告竞价底价的预测。

由于利用上述模型对广告底价预测需要利用预测的收视率的值。所以有必要利用现有的收视率的数值对将来的收视率数值进行预测。由于收视率的数值随时间的变化有明显的周期性（如图一所示），故可以利用时间序列分析，对时间序列进行建模。



图一 收视率随时间的变化

具体的建模的方法为：

（1）首先进行时间序列模型的定阶。因为数据具有明显的8个时段的周期性，故对数据做下列差分运算：

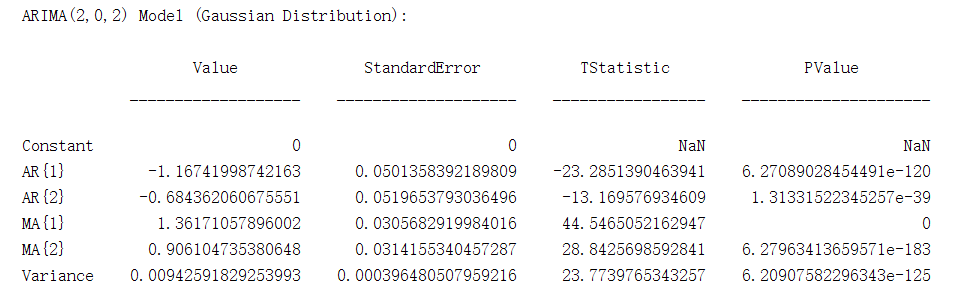
CodeCogsEqn.gif

由于差分后的时间序列是平稳的，所以可以直接对Wt进行ARMA模型拟合。用选取的p,q的各种阶数进行试算，用AIC和BIC准则进行定阶。

根据BIC的值，确定选取的p=2，q=2。

（2）建立模型并进行预测。得到的模型为

模型的拟合优度为：R^2=0.9321，可见模型拟合效果较好。



## 5.4 问题三的分析与求解

### 5.4.1 问题分析

首先分析5.1中DeepFM性能差的原因，首先在特征选取时，用户的特征只包括了用户的一些属性，基于的假设是这些属性的用户对某些属性的商品具有的偏好存在对应关系，但显然地，比如同种地域的用户可能会出现对某种商品的偏好，同济大学的学生喜爱在淘宝上买结构李旭教材，但是同济大学的学生可能是上海的也可能是新疆的。该条假设在群体上虽然成立但是在个体上失去了意义。这一问题将在接下来根据用户历史行为建立模型中得到解决。

首先，从两个角度考虑推荐问题。第一，Exploitation:一个用户近期具有一定的广告浏览记录，即用户点开的广告反应了用户对于某类商品的需求，即基于用户浏览的最多，最频繁的商品广告类，进行推荐，即用户点开什么广告多，就推荐与该广告相似的（此在3.2已述）。

其次，exploration,为了发掘用户可能感兴趣的商品广告，需要探索用户可能感兴趣的类,

我们基于大数据，通过寻找具有相似行为特征的用户进行推荐。

最后，当某一新用户，我们缺少他的历史记录时，我们采用当前最热推荐模型，从极大似然估计的角度进行推荐，即，什么卖的多，我们推什么。

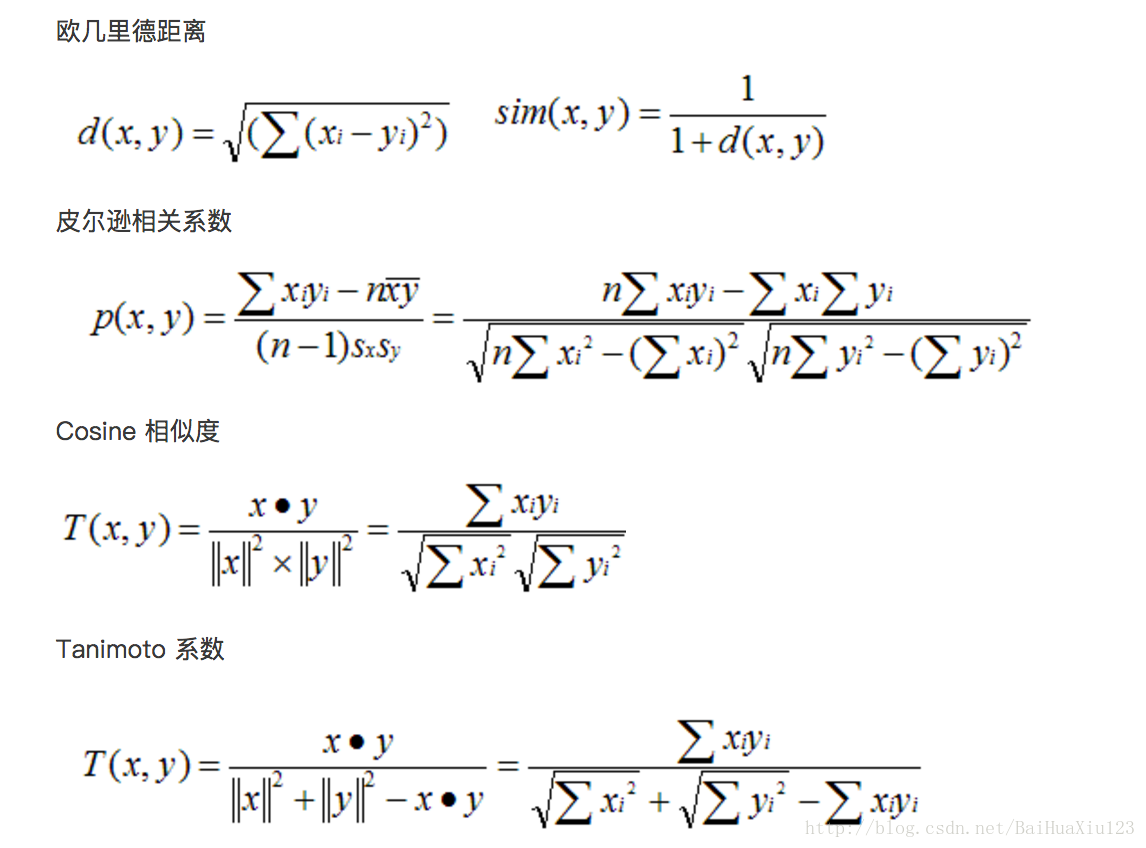
### 5.4.2 协同过滤与当前最热估计的模型建立与分析

#### 5.4.2.1 协同过滤推荐模型

根据某用户的相似行为用户为其推送广告

根据基本假设，同一广告，用户在手机上看到和在电视上看到具有相同的感受。我们使用用户在手机端与广告的互动数据来代替用户在电视上对广告的感受，是十分合理的。

对于一个要进行广告推送的用户，我们已知他之前对广告的历史行为，我们采取寻找与他相似的若干用户，将这些用户普遍喜好的商品广告推荐给该用户。即根据和一个用户有共同喜好的人来给他推荐。 我们将每一条用户信息简化为有序数组***x****,*其中***x[i]***表示x用户对编号为i的广告的喜好次数。对于任意两个用户x,y，我们有当不仅局限于如下方式定义两者相似度：



具体在本问题中，我们采用方差来表示相似度：



显然S(x,y)越小，x,y越相似。由于我们一般认知为相似度越大两者越相似，因此我们修正相似度为方差的相反数，即相似度：



至此我们可以采用协同过滤算法解决此问题。我们首先对n user条用户的历史信息建表，对于每个待推送用户x得到first same个相似用户，并选择这first same个最为偏好的first ad条广告推荐给该用户。当两条广告相似度相同时，我们优先推荐销量大的给用户。

N user条历史信息

建表

用x的数据更新表

推荐给用户

firstad条推送广告

#### 5.4.2当前最热的似然估计

5.4.1模型完成了基于用户及其相似用户的推荐算法。但考虑实际情况，一个用户难以看过所有的广告，而我们寻找的相似用户也是基于用户曾经的喜好记录，因此以上模型可以为用户推荐出很有把握准确的广告，不能保证未被推荐的广告不是足够适合该用户的广告。假设有用户x和广告ad，用户x对ad的商品足够感兴趣但x从未见过广告ad，显然用户x对ad的历史记录为0，3.2基于用户历史记录特征的协同过滤为x匹配到的相似用户很可能也会对ad的爱好程度较低，从而导致合适广告ad无法被推荐给x。

我们采取极大似然估计的思想来解决这个问题。我们无法准确判断用户对一个历史记录不佳的商品的喜好程度，于是我们认为一段时间内销量好的商品用户很可能喜欢。对于每个商品我们定义变量热度hot。当用户完成一次购买活动后，对应商品hot值加一。为防止一个商品一段时间过热，导致hot值很高，后期不受欢迎却仍然容易被推荐的情况，我们规定时效t，即hot值为当前t时间段内该物品被购买的次数。

通过此方式，我们解决了用户数据片面性的影响。虽然我们不能保证每次推荐的广告一定对该用户是很合适的，但为一类用户接受新广告提供了可能。

当前最热似然估计模型如下：

P()=

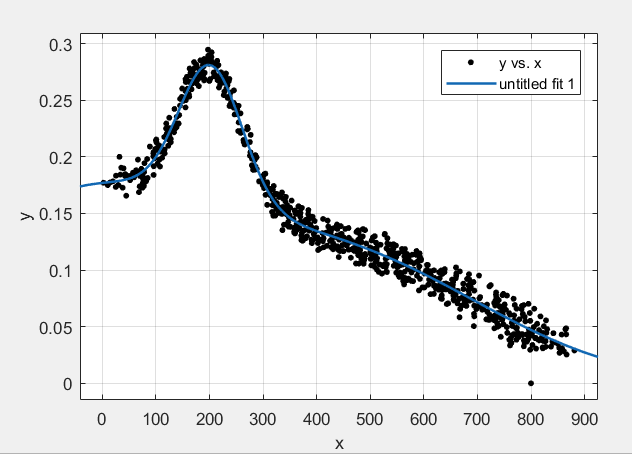
可知，对于某广告，其被推荐的期望为P()，显然地，准确率至少>=P()(此处简化了模型)

（实际上准确率难以建模计算）

#### 5.4.3 联合推荐模型架构

##### 5.4.3.1 用户点击分布函数

通过统计不同广告点击量的用户数，求解用户对广告点击数的分布，拟合结果如下（数据见附录）：



General model Gauss3:

***p(x)*** =

a1\*exp(-((x-b1)/c1)^2) + a2\*exp(-((x-b2)/c2)^2) +

a3\*exp(-((x-b3)/c3)^2)

Coefficients (with 95% confidence bounds):

a1 = 0.1196 (0.1094, 0.1299)

b1 = 201.9 (199.4, 204.5)

c1 = 83.29 (78.06, 88.53)

a2 = -3.338 (-3429, 3422)

b2 = 202.7 (-1509, 1914)

c2 = 404.9 (-4144, 4954)

a3 = 3.499 (-3422, 3429)

b3 = 199.4 (-1434, 1833)

c3 = 413.8 (-4037, 4865)

Goodness of fit:

SSE: 0.06594

R-square: 0.9838

Adjusted R-square: 0.9836

RMSE: 0.009125

##### 5.4.3.2 用户点击数和预测精度关系

为了得到，随着用户历史行为数量上升，与协同过滤算法准确率的函数关系，首先分为80个batch每个batch取20个用户进行预测，取平均准确度，用多项式进行拟合。结果如下（数据见附录）:

Linear model Poly4:

***f(x)*** = p1\*x^4 + p2\*x^3 + p3\*x^2 + p4\*x + p5

Coefficients (with 95% confidence bounds):

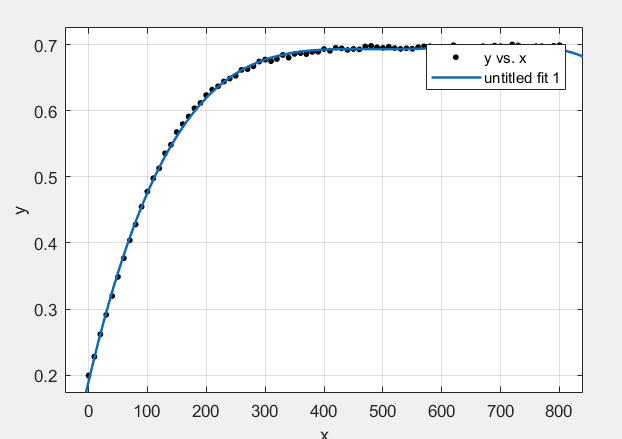
p1 = -5.748e-12 (-6.098e-12, -5.398e-12)

p2 = 1.286e-08 (1.229e-08, 1.342e-08)

p3 = -1.056e-05 (-1.086e-05, -1.026e-05)

p4 = 0.00379 (0.003732, 0.003849)

p5 = 0.1901 (0.1868, 0.1934)

.

##### 5.4.3.3 联合推荐权重

建立联合推荐模型：, x为协同过滤推荐置信度，y为极大似然估计推荐置信度，求解使得z最优的参数，为了求解方便，不妨假设,于是：由5.4.3.1，5.4.3.2，当用户浏览历史纪录数量为x的概率为:

***p(x)*** =

a1\*exp(-((x-b1)/c1)^2) + a2\*exp(-((x-b2)/c2)^2) +

a3\*exp(-((x-b3)/c3)^2)

a1 = 0.1196 (0.1094, 0.1299)

b1 = 201.9 (199.4, 204.5)

c1 = 83.29 (78.06, 88.53)

a2 = -3.338 (-3429, 3422)

b2 = 202.7 (-1509, 1914)

c2 = 404.9 (-4144, 4954)

a3 = 3.499 (-3422, 3429)

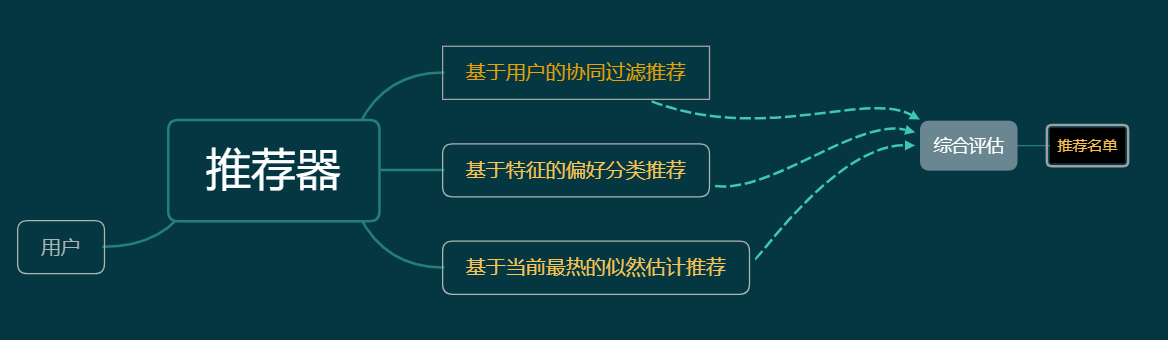
b3 = 199.4 (-1434, 1833)

c3 = 413.8 (-4037, 4865)

f(x) = -5.748e-12 \*x^4 + 1.286e-08\*x^3 + -1.056e-05 \*x^2 + 0.00379 \*x +0.1901，使用当前热度似然估计，使得每个商品对应置信度乘以权重加和，即使最终推荐选择。

从用户池中随机抽取1000个用户，取 共10个step使用联合预测法，验证精度，取达到最大值时的=0.2

最终联合推荐法如下：



## 5.5 问题四的分析与求解

### 5.5.1 问题分析

该问题可以分解为以下几个子问题，分别是：卖方（电视台）怎样选择竞价广告中的中标者，买房（广告商）怎样出价对自己最优，卖方在确定了候选的推送广告后怎样对广告进行推送。

### 5.5.2更新的分户推荐模型结构如下：

#### 5.5.2.1 竞价交易中的博弈和权衡

总结归纳后，我们发现想要实现在极大化卖方收益的同时极大化收视率和销售量需要考虑以下几点：

1. 由于是在电视台进行分户推送广告，所以同一时间点上会有不同特征的观众收到自己更感兴趣的广告。所以在同一时间点可以同时接受多个广告的报价。
2. 根据基本假设：如果在一定时间内给用户推送他并不感兴趣的广告或者广告质量低的广告，则用户会对该电视台的收视率会降低。
3. 如果电视台只关注自己的短期利益，只接受出价最高的报价，而不去考虑观众对于广告的喜爱程度，会导致将很多用户不感兴趣的广告推送给用户。这样做的结果是会导致电视台的收视率下降。由于电视台的收视率和广告的底价呈正相关，所以会损失电视台的长期利益。
4. 广告质量度Q的度量是根据广告的内容、创意、形式上的独特性和创新性进行打分。广告质量如果很差，观众对于电视台的收视率会降低，不利于电视台的长远收益。所以在计算广告的基本竞争力BRE时要将广告质量度考虑进去。
5. 用户接受度A是根据已有的历史纪录中，用户对于每一种广告的喜爱程度进行度量。如果有多数用户对于某一种类的广告喜爱度很高，则可以说明该广告在推送时的受欢迎程度会更高，会对收视率和产品销售量有正向的积极作用。
6. 电视台面临短期的最大利益和长期的最大利益之间的取舍。若电视台只关注眼前利益，则会推荐给观众并不感兴趣的广告，会使收视率和销售量受到消极影响。而如果完全根据用户的喜好进行广告的推送，在对收视率和销售量起到积极作用的同时使短期的收益减少。所以电视台要同时考虑自己的长期收益和短期收益以实现自己的总体收益的最大化。所以在选择中标的广告时，该问题为最优化问题。（见4.1.2）
7. 电视台选择具体接受哪个广告时，同时考虑每个广告商的报价、每种广告用户的喜爱程度和广告的质量度。
8. 在确定了要推送的几个广告后，电视台根据已经建立的更新的分户推荐模型，根据用户的喜好分别推荐用户最感兴趣的广告，尽量实现推送的广告都是用户感兴趣的，以保证收视率和产品销售情况是在当前条件下最优的。

#### 5.5.2.2 最优化问题

|  |  |
| --- | --- |
| **定义** | **解释** |
| **竞价广告利润系数：** | **表示某条广告被单词点击电视台的收益** |
| **候选广告池：** | **i个独立竞价广告ad** |
| **推荐广告池：** | **j个可选的推荐给用户的广告ad** |
| **广告池：{}** | **i个数据中所有广告** |
| **用户池：God = {},** | **为长为m向量表示一个用户对m个广告的历史纪录** |
| **广告选择向量： *X*** | ***one-hot编码的，*形如[0, 1, 1, 0, 0, ……, 0, 1], 表示从中选择进入 的广告选择方案** |
| **广告表现张量**：***Performance(x, A)*** | ***为广告组合为x时，广告组合在用户池中的表现张量。现实意义为广告池中每个广告符合用户需求的次数。*** |
| **利润系数向量：*Profit*** | ***广告组合为x时，对应的单次点击利润系数向量；形如：*[,,…]** |
|  | ***单位时间内电视台收视率*** |
| ***Pr*** | ***电视台利益*** |

设 =

= if : 1 else :0

=

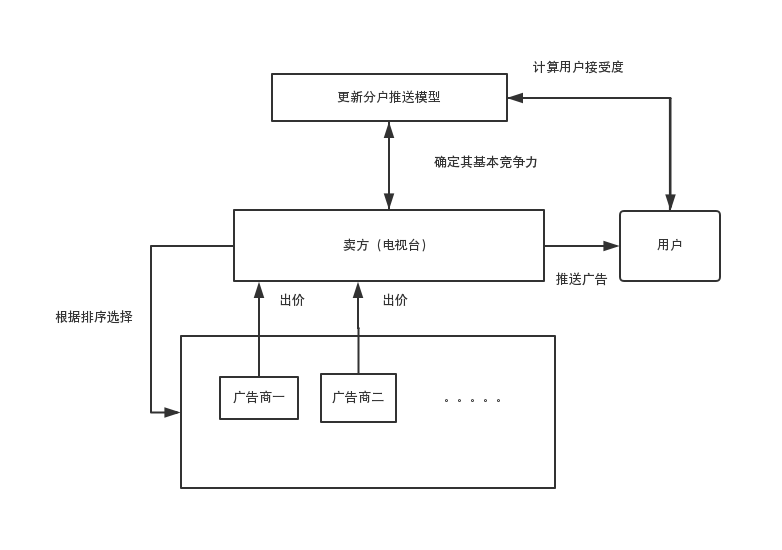
***Performance(X, God) = Performance(, God) = Performance([, ……, ],God) = [, ……, ]***

**根据以上定义，得出卖方的利益函数：**

***: ,***

#### 5.5.2.3 竞价交易总体模型

针对以上问题，我们提出了能够兼顾上述问题并且尽量达到全局最优化的竞价交易模型，能够在极大化卖方收益的同时，尽量保证收视率和销售量。总体框架如下图所示



由基本假设知：电视台在同一时间段能够进行分户推送的广告种类的数量限定为n个。

首先由广告商进行出价，卖方（电视台）收到广告商的出价后，基于在前文中建立的更新的分户推送模型计算出每一种广告的用户接受度A，同时在审核广告时对广告的质量进行打分。更新的分户推送模型对用户接受度进行计算的依据是全体用户的历史行为，。在卖方确定了每个广告的广告质量度Q和用户接受度A 后，计算出每一个广告商的基本竞争力BRE。

对所有广告的基本竞争力进行排序，选出排名靠前的n个广告商接受他们的出家，每个广告商得实际出价根据广义的第二密封竞价模型进行计算（详见4.3）。电视台在确定了n个将要分户推送的广告后，利用更新的分户推送模型对这n个广告进行分户推送，尽量保证将广告推送到对该种广告感兴趣的用户处。

#### 5.2.2.4 用户接受度(A)的计算

用户喜爱度的计算对于实现卖方对于自身最有决策的选择至关重要，广告喜爱度体现了用户对于该广告的普遍喜爱程度。在两个竞价方出价相同时，如果某一方的广告的用户喜爱度较高则可以说明该广告更能起到提升收视率和销售量的效果，对于电视台的长期收益是更为有益的。利用前文中提出的更新的分户匹配推送模型可以对用户喜爱度进行计算，计算的方式如下：

1. 利用更新的分户匹配推送模型，对用户池中所有的用户对所有类别的广告进行一次用户喜爱度(L)的预测，每个竞价周期只需要进行一次计算，且该计算的结果可以在后续的分户推荐中起到至关重要的作用。
2. 得到了每一个用户对每一类别的用户喜爱度Lij后，对用户喜爱度进行归一化处理。
3. 将归一化后的用户喜爱度对每一种类型的广告进行求和，以此估计出某一种广告在当前用户中的用户接受度A。

计算公式为：

https://latex.codecogs.com/gif.latex?%5Cdpi%7B120%7D%20A%3D%5Csum_%7Bi%3D1%7D%5En%20%5Cfrac%7BL_i_j%7D%7B%5Csum_%7Bj%3D1%7D%5E%7Bm%7D%7BL_i_j%7D%7D

其中Lij为某一用户对于某一类型广告的用户喜爱度。

利用该种方式计算出的用户接受度可以在宏观上表示该类产品在电视台观众处的接受程度。

#### 5.2.2.5广告质量度(Q)的计算

由于广告的质量如果较差则会对用户的体验有消极影响，会在一定程度上降低收视率，所以计算广告的质量度也是十分重要的。广告质量度的评判标准有以下几个因素：

（1）相关性：包括关键词和创意的相关程度，以及购买词和推广业务的相关性。

（2）创意构思水平：创意围绕关键词撰写得越通顺、越有“创意“，越能吸引潜在客户的关注。

广告质量度的计算公式如下：

https://latex.codecogs.com/gif.latex?%5Cdpi%7B120%7D%20Q%3Dax_1&plus;bx_2&plus;c

其中x1为相关性评分，x2为创意构思水平打分，a和b为权重，且a和b远小于c。

由于广告质量度的判断只是电视台对于广告的一个初次筛选过程，筛选时为电视台针对每个广告的相关性和创意构思水平进行评分。由于此种有较强的主观性，所以该因素对于广告最终基本竞争力的排名影响很小，主要目的在于筛选出质量明显较差，有较大可能会影响到收视率的广告。

#### 5.5.2.6 广义的第二密封竞价制度

本文中的竞价交易体系所使用的是广义的第二密封竞价模型，买方之间不知道彼此的出价金额，只能知道自己在竞争中的排名是不是第一位，一旦交易成功，中标者支付价格的公式为： CodeCogsEqn.gif 同时，由于交易时买家之间并不知道彼此的报价，所以可以达到出价的纳什均衡，每个出价者有足够的激励去报出自己能承受的最高报价。在每个买家报出自己能承受的最大报价的情况下，卖方能够得到相对最大的收益。“一价”会导致系统抖动，造成不稳定。但“二价”中，由于买家的计费价格跟自己的出价没有直接关系，没有动力通过修改自己的出价，在赢得拍卖的同时获得更多的利益，因此避免了出价的抖动。

同时，采用密封的形式可以有效地避免竞价方得知其他人的报价后，选择对自己更有利但对电视台的收益不利的选择。可以实现对电视台收益的最大化

基本竞争力计算公式：

https://latex.codecogs.com/gif.latex?%5Cdpi%7B120%7D%20BRE%3DA%5Ccdot%20Q%5Ccdot%20P

# 模型检验

### 6.1模型I算例及检验

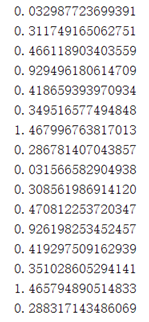
随机收取4组，每组50人，每人推荐k（10，20，30）个，结果如下;

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 推荐数量 | 测试用户组a | 测试用户组b | 测试用户组c | 测试用户组d |
| 精度/10（%） | 52.3 | 49.1 | 55.2 | 49.6 |
| 精度/20（%） | 57.6 | 51.2 | 59.2 | 50.6 |
| 精度/30（%） | 59.2 | 52.6 | 61.0 | 53.2 |

## 6.2 模型II 算例及检验

预测样例：

利用已经建立的ARMA(2,2)预测未来两天的收视率数值为



在2.2和2.3中，我们分别建立了已知收视率对广告竞价底价进行预测以及根据现有的收视率对未来的收视率进行预测的模型。综合以上两个模型，我们可以根据已有的收视率的历史纪录预测未来任意分时段的广告竞价底价模型。

未来分时段收视率

分时段预测广告竞价底价

收视率历史纪录

预测价格

ARMA(2,2)预测

## 6.3 模型III算例及检验

完成推荐过程后，我们根据x的数据对表进行更新。如果表中有x记录，我们将表中x的记录更新为现有记录。若表中无x记录，则将x记录加入表中。从而实现模型更新。

经实际验证，对于较大数据范围该模型有较高准确性。

|  |  |  |
| --- | --- | --- |
| 表中初始记录数（n user） | 每次为用户推荐广告数（first ad） | 成功率 |
| 100 | 5 | 66.7% |
| 100 | 10 | 82.7% |
| 100 | 15 | 83.6% |
| 5000 | 5 | 66.7% |
| 5000 | 10 | 83.3% |
| 5000 | 15 | 88.9% |

注：1.广告共计100000种

2.每次推荐的广告中只要有一条用户喜好即为成功。

## 6.4 模型IIII算例及检验

**具体算例**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 竞价方 | 出价P | 广告质量度Q | 用户接受度A | 基本竞争力BRE |
| 广告A | 1.8339 | 9.9 | 5178 | 94009.75 |
| 广告B | 2.2588 | 9.9 | 9245 | 206737.8 |
| 广告C | 0.8622 | 9.4 | 9670 | 78372.26 |
| 广告D | 0.3188 | 9.8 | 8393 | 26221.75 |
| 广告E | 1.3077 | 9.1 | 8788 | 104577.8 |
| 广告F | 0.4336 | 9.4 | 8715 | 35520.95 |
| 广告G | 3.5784 | 9.9 | 6961 | 246601.5 |
| 广告H | 2.7694 | 9.7 | 8277 | 222346.5 |
| 广告I | 1.3499 | 9.9 | 5855 | 78246.28 |

当电视台确定某一时段应该接受的广告数量n后，根据广告的基本竞争力排行选取前n位广告接受报价。

例：当n=2，α=0.001时，广G和广告H竞标成功。没有中标的第一位是广告B。

根据公式CodeCogsEqn.gif

广告G 的实际出价P1=(2.2588\*9.9\*9245)/(9.9\*6961)+0.001=3.0009

广告H的实际出价P2=(2.2588\*9.9\*9245)/(9.7\*8277)+0.001=2.5760

**模型的优点**

根据4.1中提出的竞价中的博弈和权衡。该模型利用了广义的第二密封竞价制度，在最大程度上激励竞价方报出自己能承受的最大价格，对于卖方来说是有效的极大化自身收益的方式。同时该模型还考虑到了自身利益与收视率和销售情况的权衡，使用广告基本竞争力BRE作为选择参与竞价的广告的标准，在已经建立了更新的分户匹配推送模型的基础之上利用简单的方式对该最优化问题进行了求解，可以尽可能地实现全局的最优。

# 7.模型总结

没什么可说的，面对50w规模的广告和10w规模的用户，效果挺好。第四问没实际数据验证。

**参考文献：没时间写了，有个队员回家了。**

附录：