基于用户时序行为的需求预测算法

1 算法简介

本算法拟通过对用户多时序的浏览（pv）、收藏（fav）、加购（cart）、购买（order）等行为数据，并结合用户背景信息及商品特征信息，进行行为特征建模与客群分类，进而针对不同客群进行需求预测分析。

1.1 数据介绍

实验数据集包括：

- 京享值大于 9000 的用户，共 269805 个 id（年龄标签："1"，15 岁以下；"2"，16-25 岁；"3"，26-35 岁；"4"，36-45 岁；"5"，46-55 岁；"6"，56 岁以上。性别标签："0"，女性；"1"，男性。）
- 美妆护肤、食品饮料、电子数码三大类商品，共 161800 个 sku（均为京东自营商品）
- 时间跨度为：2021/02/01 - 2021/05/31

表 1: 数据介绍

数据类型	解释	数据字段
用户信息	用户 ID	user_id
	用户性别	sex
	用户年龄	age
	用户位置	province
		city
		country
商品信息	商品 ID	sku_id
	商品名称	sku_name
	商品类别信息	first_cate_id、first_cate_name
		second_cate_id、second_cate_name
		third_cate_id、third_cate_name
用户浏览行为（pv）	浏览操作	user_id、sku_id、action_time
用户收藏行为（fav）	收藏操作	user_id、sku_id、action_time
用户加购行为（cart）	加购操作	user_id、sku_id、action_time
用户下单行为（order）	下单操作	user_id、sku_id、action_time
	下单数量及价格	sale_qtty、after_prefr_amount

对数据进行处理：

- 删除“action_pv.csv”表格中“sku_id”和“action_time”为空的行；
- 将“action_pv.csv”、“action_fav.csv”、"action_cart.csv"、“action_order.csv”四个表格合并，并提取用户对同一种商品的行为序列（一个小时内的相同行为被视为一个行为动作，超过一天的行为间隔将被标记为“-”），形成“user_action_sequence”的表格。

1.2 算法优势

基于用户行为特征进行需求预测，可解释性强，应用范围广。

1.3 应用场景

本模型针对不同行为特征的分类客群进行需求预测，因此可在促销、营销等商品销售环节中，有针对性地对某类客群进行营销推广，增大其购买率。

2 模型结构

2.1 隐马尔科夫链模型

用户在购物网站上针对某一种商品的行为包括浏览、收藏、加购物车、下单等，且可能是在连续时间完成的，也可能是经过间隔时间完成的。如图1所示，用户一系列的时序行为可以由“p、f、c、o、i、s”表示，其中“p”为 pageview 是用户对商品的浏览，“f”为 fav 是用户对商品的收藏，“c”为 cart 是用户将商品加购物车，“o”是 order 是用户下单购买该商品，“i”为 interrupt 表示用户间隔一段时间后继续对该商品进行下一个操作，“s”为 stop 表示用户停止对该商品的操作行为。

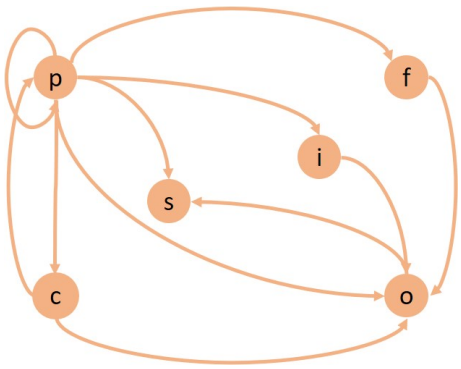

图 1: 用户行为示例图

通过对数据的简单分析，我们发现用户的时序序列包括上万种不同的形式和长度，将用户行为表征为“interested、willing to buy、forget”三种不同的心理状态，我们利用隐马尔科夫链模型进行建模，如图2所示。

隐马尔科夫链模型包含 N 个状态，由于马尔科夫的“无记忆性”，在离散时间 t 的状态仅仅取决于 $t-1$ 时刻的状态，与之前的状态无关，其可以用一个状态转移概率矩阵表示。每个时刻 t 实际观察到的行为 Q_t 取决于其状态变量 O_t，涉及到的数学符号如下：

- T：用户行为序列的长度

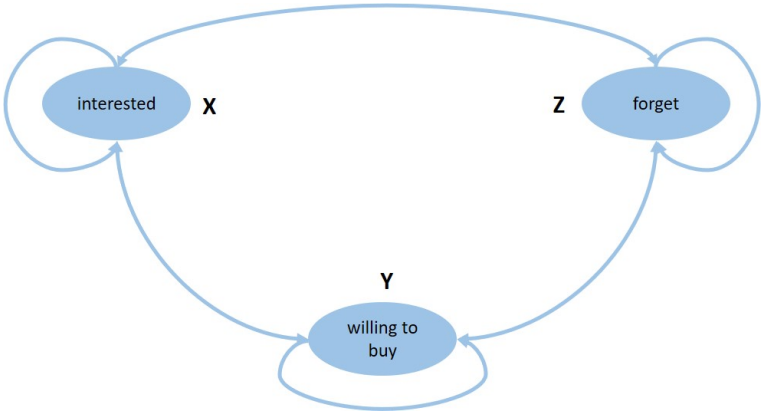

图 2: 隐马尔科夫链状态转移示例图

- N：隐马尔科夫链中的状态数，为 3
- M：用户行为数，为 6
- S：状态集合 $\{X, Y, Z\}$
- V：用户行为集合 $\{p, f, c, o, i, s\}$
- Q_t：t 时刻的状态
- O_t：t 时刻观察到的用户行为
- A：状态转移概率矩阵，维数为 $N \times N$，其中 $a_{ij} = P(Q_{t+1} = j | Q_t = i)$ 为从状态 i 转移到状态 j 的概率
- B：用户行为概率矩阵，维数为 $N \times M$，其中 $b_{jk} = P(O_{t+1} = k | Q_t = j)$ 为当状态为 j 时，观察到的用户行为为 k 的概率
- π：初始状态概率向量，维数为 N，其中 $\pi_i = P(Q_1 = i)$
- $Q = Q_1, \ldots, Q_T$：状态序列
- $O = O_1, \ldots, O_T$：观察到的用户行为序列
- $\lambda = \langle \pi, A, B \rangle$：隐马尔科夫链模型参数集合

在一个参数为 λ 隐马尔科夫链模型中，一个行为序列 O 的概率为：

$$P(O|\lambda) = \sum_{Q} P(O, Q|\lambda) = \sum_{Q_1, \ldots, Q_T} \pi_{Q_1} b_{Q_1 O_1} a_{Q_1 Q_2} b_{Q_2 O_2} \ldots a_{Q_{T-1} Q_T} b_{Q_T O_T} \tag{1}$$

此外，考虑递推的思想计算 $P(O|\lambda)$

1) 正向递推：定义 $\alpha_t(i) = P(O_1, \ldots, O_t, Q_t = i|\lambda)$ 为当时刻 1 到 t 的行为序列为 $O_1, \ldots, O_t$ 且 t 时刻的状态为 i 的概率

$$\begin{aligned} \alpha_1(i) &= \pi_i b_{iO_1} \\ \alpha_{t+1}(j) &= \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_{jO_{t+1}}, 1 \le t \le T-1 \\ P(O|\lambda) &= \sum_{i=1}^{N} \alpha_T(i) \end{aligned} \tag{2}$$

2) 逆向递推：定义 $\beta_t(i) = P(O_{t+1}, \ldots, O_T, Q_t = i|\lambda)$ 为当时刻 $t+1$ 到 T 的行为序列为

$O_{t+1},\ldots,O_T$ 且 t 时刻的状态为 i 的概率

$$\begin{aligned}\beta_T(i) &= 1, 1 \le i \le N \\ \beta_t(i) &= \sum_{j=1}^{N} a_{ij} b_{jO_{t+1}} \beta_{t+1}(j), 1 \le t \le T-1 \\ P(O|\lambda) &= \sum_{i=1}^{N} \pi_i b_{iO_1} \beta_1(i)\end{aligned} \tag{3}$$

2.2 由观察到的行为序列得到状态序列

对于一个行为序列，找到其对应的状态序列即为最大化条件概率 $P(Q|O,\lambda)$，由贝叶斯模型即为最大化 $P(Q,O|\lambda)$，我们采用动态规划的思想得到状态序列。

1) 记 $\delta_t(j)$ 为在 t 时刻状态为 j 的所有可能的状态转移路径 $Q_1,\ldots,Q_{t-1},Q_t=j,O_1,\ldots,O_t$ 中的概率最大值：

$$\delta_t(j) = \max_{Q_1,\ldots,Q_{t-1}} P\left(Q_1,\ldots,Q_{t-1},Q_t=j,O_1,\ldots,O_t|\lambda\right) \tag{4}$$

递推得到 $t+1$ 时刻值为：

$$\delta_{t+1}(j) = \left[\max_i \delta_t(i) a_{ij}\right] b_{jO_{t+1}} \tag{5}$$

用 $\Phi_t(j)$ 记录时刻 t 状态为 i 的所有状态转移路径 $(Q_1,\ldots,Q_{t-1},i)$ 中概率最大的转移路径的第 $t-1$ 时刻的状态：

$$\Phi_t(j) = \arg\max_i \delta_{t-1}(i) a_{ij} \tag{6}$$

2) 初始化为：

$$\begin{aligned}\delta_1(j) &= \pi_j b_{jO_1}, j = 1,2,\ldots,N \\ \Phi_1(j) &= 0, j = 1,2,\ldots,N\end{aligned} \tag{7}$$

2.3 估计隐马尔科夫链模型中的参数 λ

给定模型参数 λ 和观测序列 O，在时刻 t 出于状态 i 和在时刻 $t+1$ 出于状态 j 的状态转移概率 $\xi_t(i,j)$：

$$\begin{aligned}\xi_t(i,j) &= P(Q_t=i,Q_{t+1}=j|O,\lambda) \\ &= \frac{\alpha_t(i) a_{ij} b_{jO_{t+1}} \beta_{t+1}(j)}{P(O|\lambda)} = \frac{\alpha_t(i) a_{ij} b_{jO_{t+1}} \beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i) a_{ij} b_{jO_{t+1}} \beta_{t+1}(j)}\end{aligned} \tag{8}$$

给定模型参数 λ 和观测序列 O，t 时刻状态为 i 的概率：

$$\begin{aligned}\gamma_t(i) &= P\left(Q_t=i|O,\lambda\right) \\ &= \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)}\end{aligned} \tag{9}$$

在观测序列 O 下，状态 i 出现的期望值为 $\sum_{t=1}^{T}\gamma_t(i)$；在观测序列 O 下，由状态 i 转移出现的期望值为 $\sum_{t=1}^{T-1}\gamma_t(i)$；在观测序列 O 下，由状态 i 转移到状态 j 的期望值为 $\sum_{t=1}^{T-1}\xi_t(i,j)$。

由此可估计状态参数 λ：

$$\hat{\pi}_i = \gamma_1(i), \forall 1 \le iN \tag{10}$$

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{11}$$

$$\hat{b}_{ik} = \frac{\sum_{t\in 1,\ldots,T:O_t=k} \gamma_t(i)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{12}$$

2.4 基于用户行为的需求预测

在本研究中，商品共为三类（美妆护肤、食品饮料、电子数码）。根据上述隐马尔科夫链模型，我们可求得每个行为序列 O 以最大概率的状态序列 Q，并可得到状态转移概率矩阵 A，即可求得用户对某类商品的行为序列以某一最大概率属于某个转移概率矩阵。

在训练集中，我们通过对用户对某类商品所有的行为序列进行隐马尔科夫链模型的学习，得到其状态转移概率矩阵。在测试集中，当已知用户特征、商品特征，以及用户对某类商品的初始操作时，我们可根据学习到的状态转移矩阵预测后续行为序列，及其是否会购买此商品。利用神经网络学习得到状态转移矩阵和状态-行为转移矩阵。