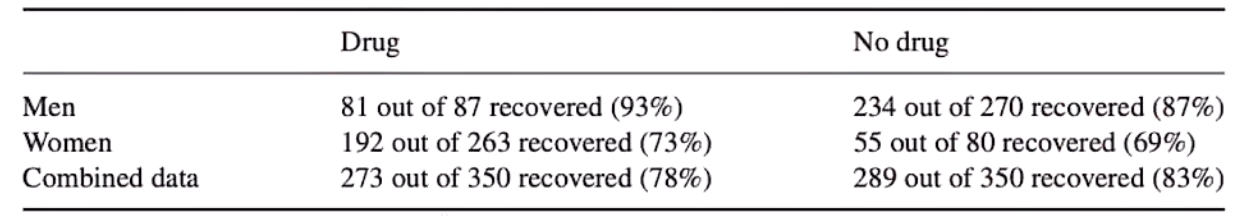
**什么是因果推断？**

**Simpson's Paradox**

Simpson's paradox主要是指**分组和全局统计**可能导出**不一致**结论的数据。

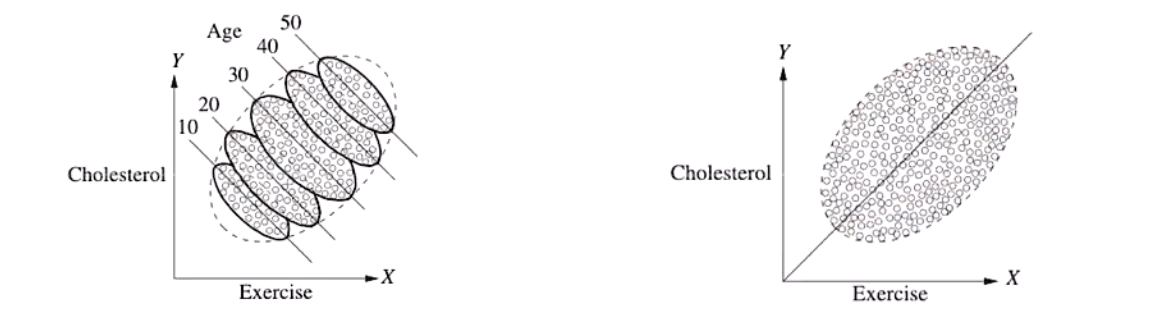
1. 考虑用药在不同性别上的恢复率差异。



**数据背后的含义**：

* 雌激素对疾病的治疗有负向效果。不管用不用药，女性比起男性恢复率更低。
* 在这个数据中，女性比起男性用药更多，分组不均。

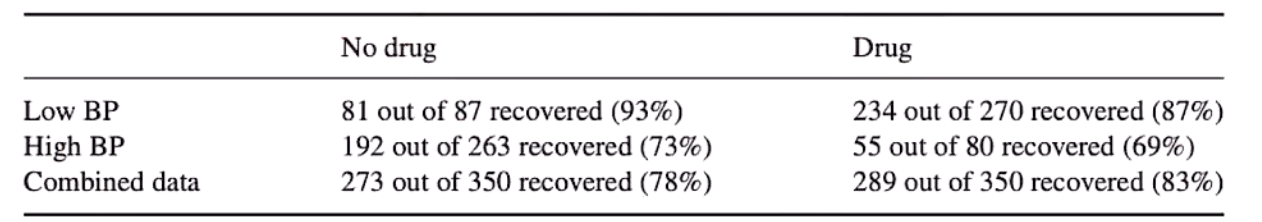
1. 胆固醇含量和运动量的关系。



**数据背后的含义**：

* 年龄越大的人运动量会更多。
* 年龄大的人倾向于更高的胆固醇含量，不管运不运动。

1. 考虑用药在 **治疗后的血压水平上**的恢复率差异。



**数据背后的含义**：

* 药物是通过降低血压来恢复的。
* 药物也有副作用。

**因果和相关性**

举个例子，研究发现，国家的巧克力的销量和诺贝尔奖的获奖人数是正相关的，背后的原因是可能和国家的经济教育水平是相关的。即 **经济教育水平高 -> 巧克力销量高**，同时**经济教育水平高 -> 诺贝尔奖获奖人数多**。即国家的经济教育水平叫做混杂因子(confounder)，那么就造成了巧克力的销量和诺贝尔奖的获奖人数这种伪相关(bias)。

如果想消除这种混杂因子带来的bias，就需要用到因果推断，概率上使用do算子表示成P(Y|do(X=x))，即X对于Y的因果效应。

**什么是因果推断**？

* 因果推断主要研究如何更加科学地识别变量之间的因果关系。
* 强调原因的唯一性。

**因果性/相关性的定义**

* 因果性：在**操作/改变**X之后，Y也会随着这种**操作/改变**也变化，则说明X就是Y的因(cause)。
* 相关性：在**观测**到的数据分布中，X和Y相关。换句话说，如果观测到X的分布，也可以推测出Y的分布。

**因果性和相关性的差异**

* 相关性通常是对称的，而因果性通常是非对称的。相关性不一定说明因果性。
* 因果性一般会在统计层面产生相关性。
* 相关性是被动的观察，因果关系是主动干预。

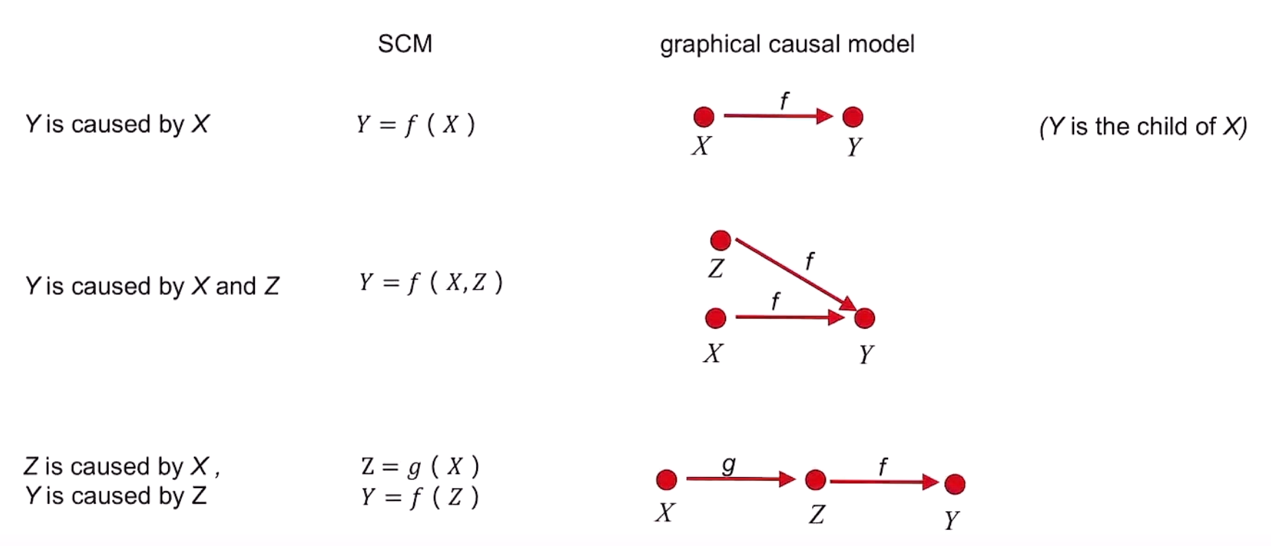
|  |  |  |
| --- | --- | --- |
|  | 相关性 | 因果性 |
| 表达式形式 | P(Y|X=x) | **P(Y|do(X=x))** |
| 实现条件 | 观察 | 干扰 |
| 差异 | 在X=x的数据中，观察Y的分布，没有修改数据的分布。关注感兴趣的部分，改变了看世界的角度。 | 强制某个变量X等于x，对于X的其余依赖变量不影响变化，只影响最终的结果，属于改变了数据的分布。 |

**因果推荐的一些基础知识**

**Structural Causal Model**

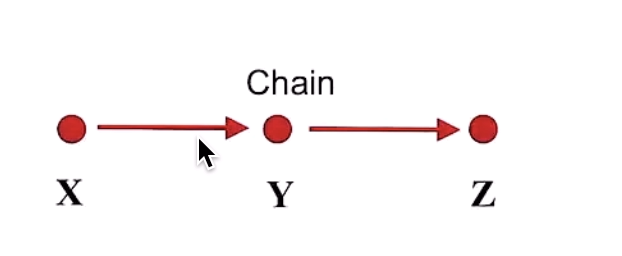
structural causal model，或者叫scm，描述因果机制的概念模型，通常包括三部分组成。分别是外生变量(exogenous variables)，内生变量(endogenous variables)，函数f。可以通过图这种数据结构来进行描述。

* 外生变量 -> 图中的根结点{root nodes}
* 内生变量 -> 图中的根结点的子孙节点{descendant nodes of root nodes}
* 函数 -> 图中的边{edges}



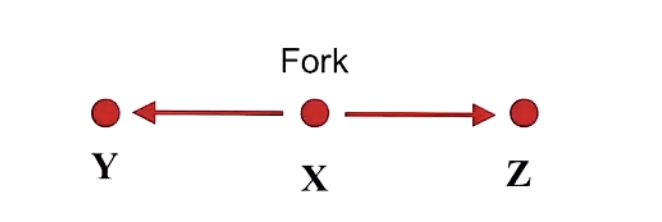
**Graphics Models**

**chain**



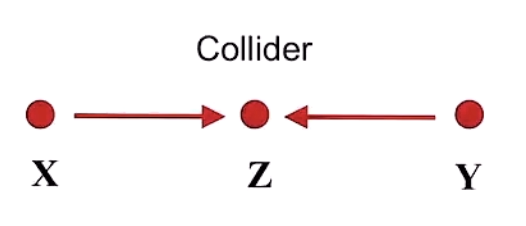
* X和Y可能是依赖/非独立的。
* Y和Z可能是依赖/非独立的。
* X和Z可能是依赖/非独立的。
* 在Y条件概率下，X和Z是独立的。
  + 当Y固定为某个值y的时候，X的变化无法影响到Z。

**fork**



* X和Y可能是依赖/非独立的。
* X和Z可能是依赖/非独立的。
* Y和Z可能是依赖/非独立的。
* 在X条件概率下，Y和Z是独立的。

**collider**



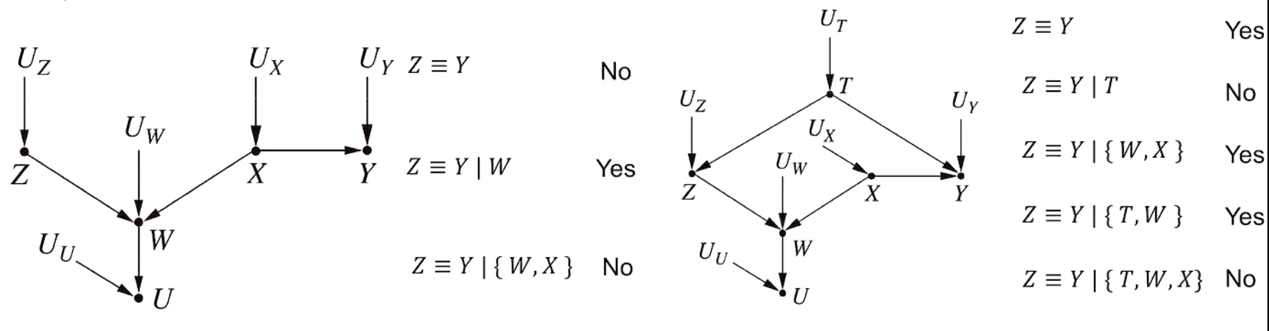
* X和Z可能是依赖/非独立的。
* Y和Z可能是依赖/非独立的。
* X和Y是独立的。
* 在Z条件概率下，X和Y是可能依赖/非独立的。

**d-separation**

现实中的图可能十分复杂，包含很多结构。对于一个DAG而言，d-separation可以快速的判断两个节点是否是条件独立的。这里简单的描述一些规则，原始的比较复杂。

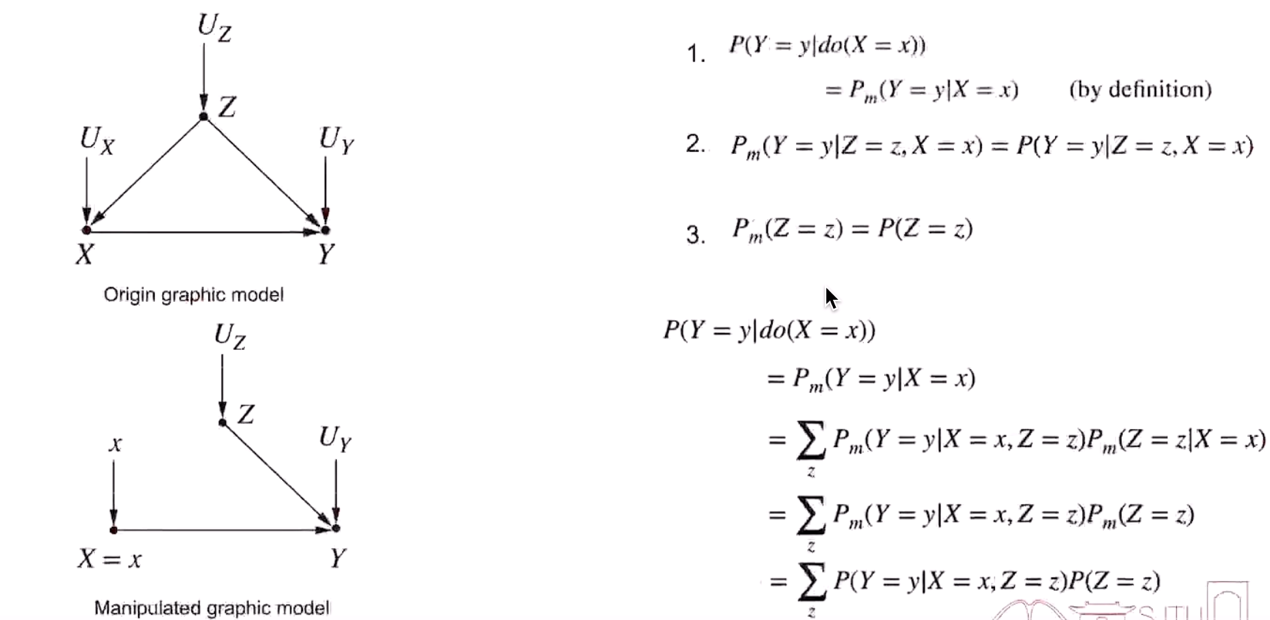
|  |  |  |
| --- | --- | --- |
| 结构 | 非条件 | 条件 |
| chain | unblock | block |
| fork | unblock | block |
| collider | block | unblock |

例：三条横杠代表相关。



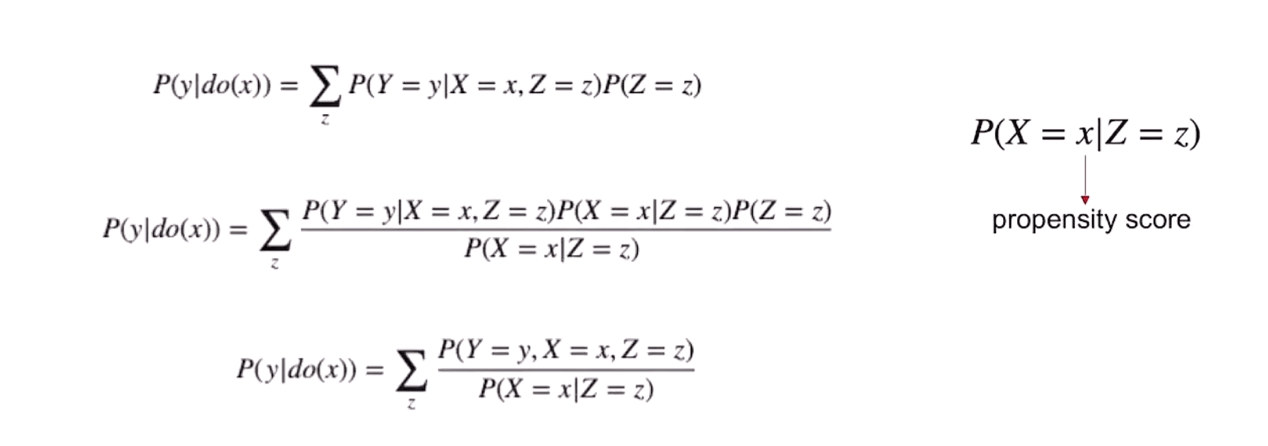
**调整公式**

如何求出对X进行干预后，Y的分布概率变化？通过调整公式，可以把 **干预** 转为 **观测**。将所有指向X的边称之为后门，需要移除所有X的后门。



**Inverse Probability Weighing**

当z的数量特别大的情况下，就会产生计算量很大。另外，z越大的情况下，样本越少，计算出的概率可靠度降低。就需要逆概率加权，来简化计算。propensity score称之为**倾向性得分**。

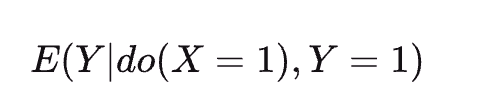


**CounterFactuals**

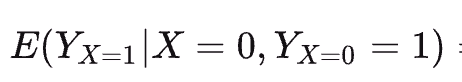
在相同的条件下，两种不同的选择造成了两种不同的结果，反事实基本都可以描述为 **如果当初**...**那么**...的语句。

Eg. 我打算回家，其中在一个交叉路口上，有两条路可以直达我家，一个是高架桥X=1，一个是地下通道X=0，Y表示抵达时间。我选择走了地下通道，结果地下通道在维修施工，导致我回家时间Y=1h。我想，如果我一开始走了高架桥，那么我会到家更快一点。

那么假设用干预来表示，就需要写成如下形式。然而两个Y是不一样的含义，一个代表实际选择X=0所需要的时间Y=1，另外一个代表X=1的时候想要求的时间Y。



所以反事实需要引入不同的下标来进行区分，原始问题就可以写成如下形式。



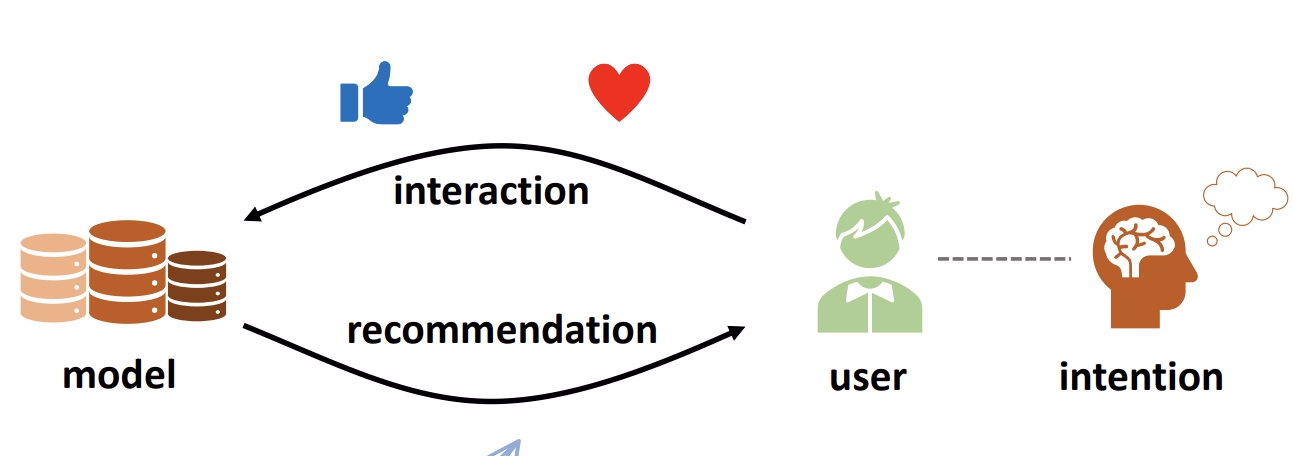
**因果推断在推荐中的应用**

因果推断正常运用的方向有 推荐系统、用户增长、智能营销上，本次只简单介绍一些在推荐系统中比较典型的一些应用，在推荐系统系统比较典型的应用是 流行度消偏。

[**Disentangling User Interest and Conformity for Recommendation with Causal Embedding**](http://staff.ustc.edu.cn/~hexn/papers/www21-dice.pdf)

**Insight**

* 点击原因 可以分解 为 用户的兴趣(interest) 和 用户的从众性(conformity)。
  + 用户的从众性是和物品的流行性是相关的，倾向于点击更流行物品。
* 从用户点击去拆解背后可能的原因，从而将用户的兴趣和从众分离出来。
  + train：兴趣 + 从众
  + infer：兴趣推荐

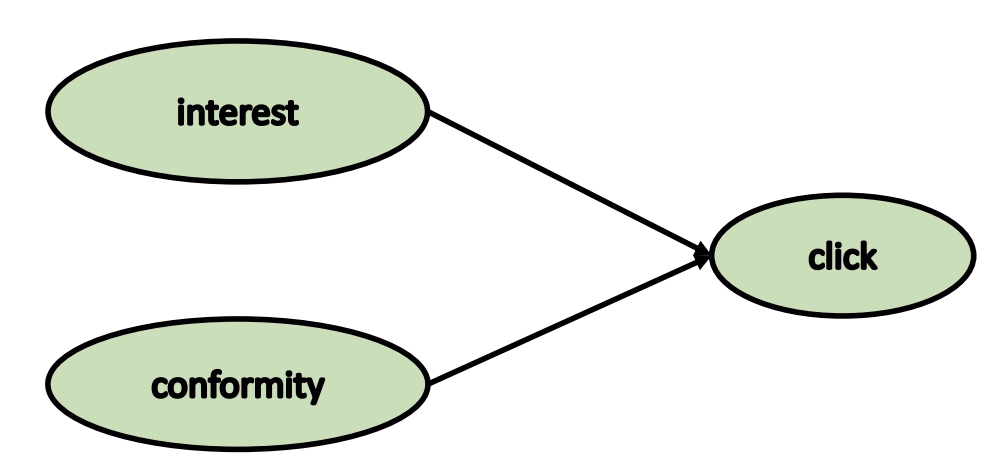


**方案**

**建立因果图**

这是一个对撞因果结构，

* 点击确定的条件概率下，兴趣和一致性是相关。
* 点击不确定的情况下，兴趣和一致性不相关。



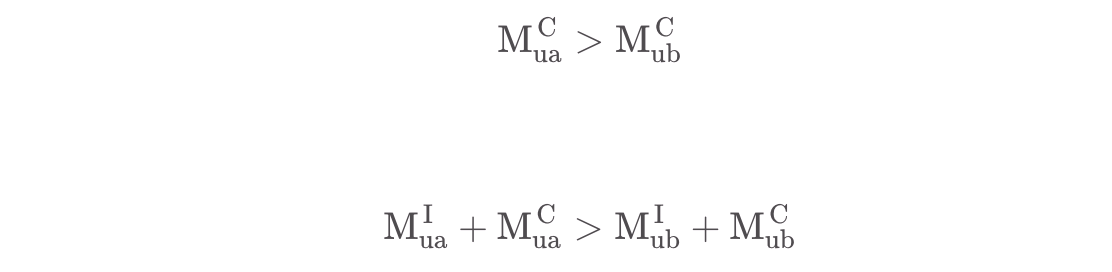
$$$$M^I

$$$$表示user和item的兴趣匹配矩阵，$$$$M^C

$$$$代表user和item的从众性匹配矩阵。将用户的样本组合成pairwise的格式，那么存在如下两种情况

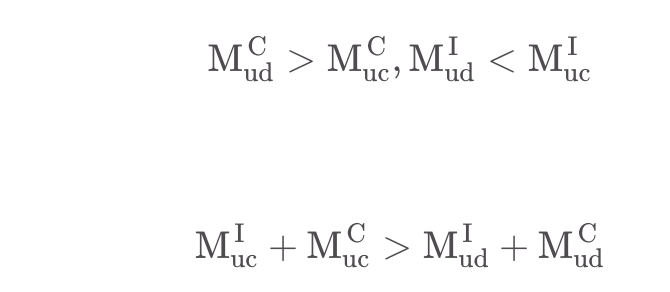
O1: 负样本b的流行度低于正样本a的流行度。

* + 只能说明b热度低于a，不能说明用户对b的兴趣低于a。

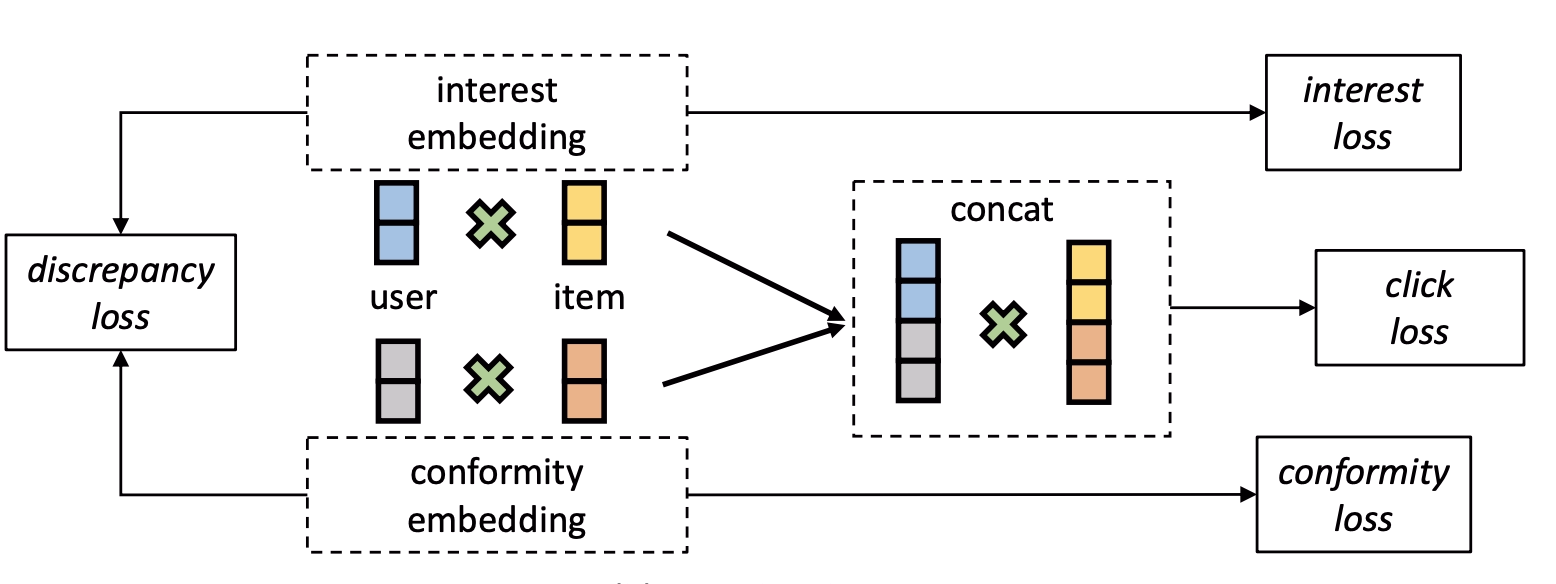


O2: 负样本d的流行度高于正样本c的流行度。

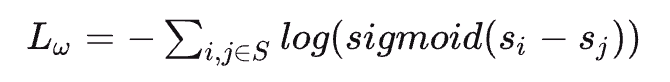
* + 说明d的热度高于c，同时用户对c的兴趣高于d。



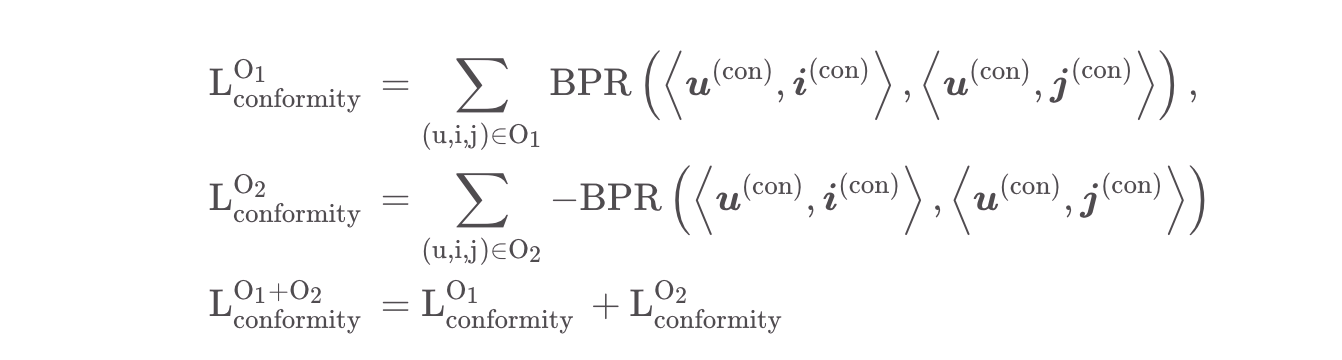
**训练**



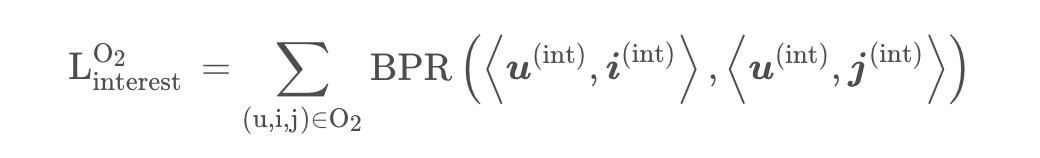
**Pairwise BPR loss**公式：



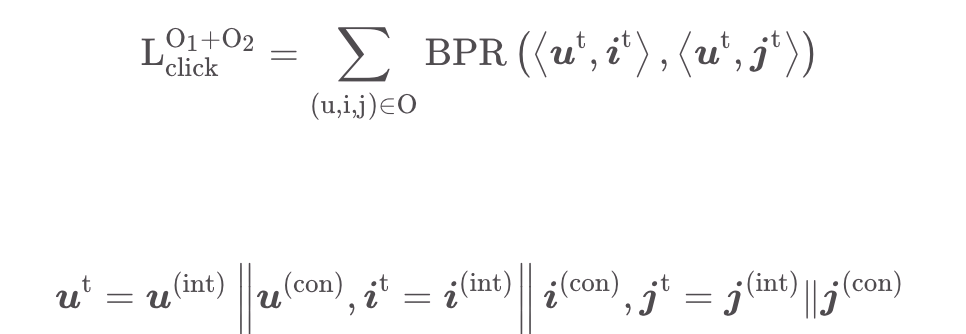
**从众性loss**：对于从众性来说，O1和O2两组情况都有，那么从众性loss就是如下所示



**兴趣loss**：兴趣loss只有O2才有。



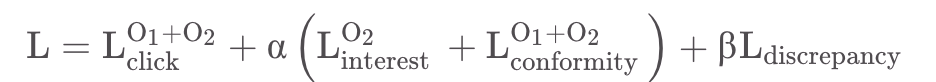
**点击loss**：兴趣和从众性embedding concat在一起。



**分离loss**：两个embedding应该尽可能的不相关。

* L1/L2距离/相关系数计算

**总loss**：多任务学习



[**Model-Agnostic Counterfactual Reasoning for Eliminating Popularity Bias in Recommender System**](https://arxiv.org/pdf/2010.15363.pdf)

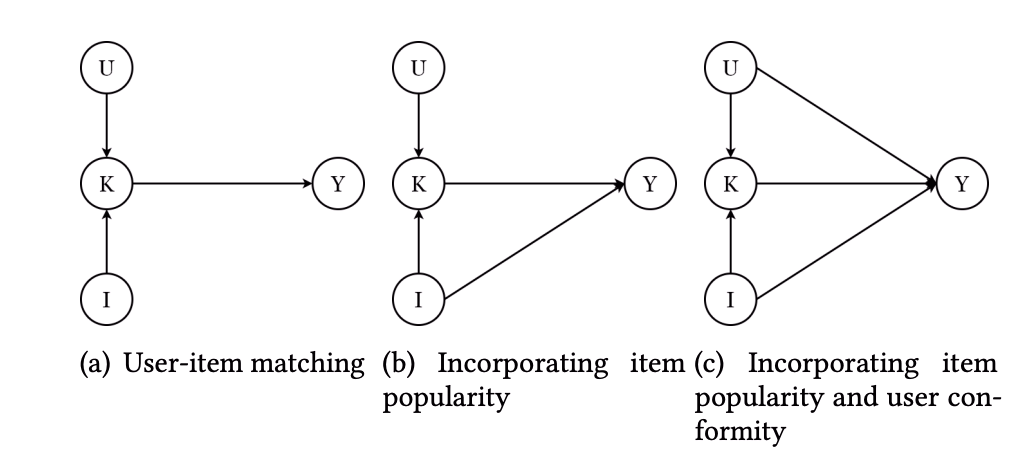
**Insight**

* 用户从众性以及物品的流行性都会直接对点击存在影响
  + 真实推荐的时候需要去掉这两条边的影响

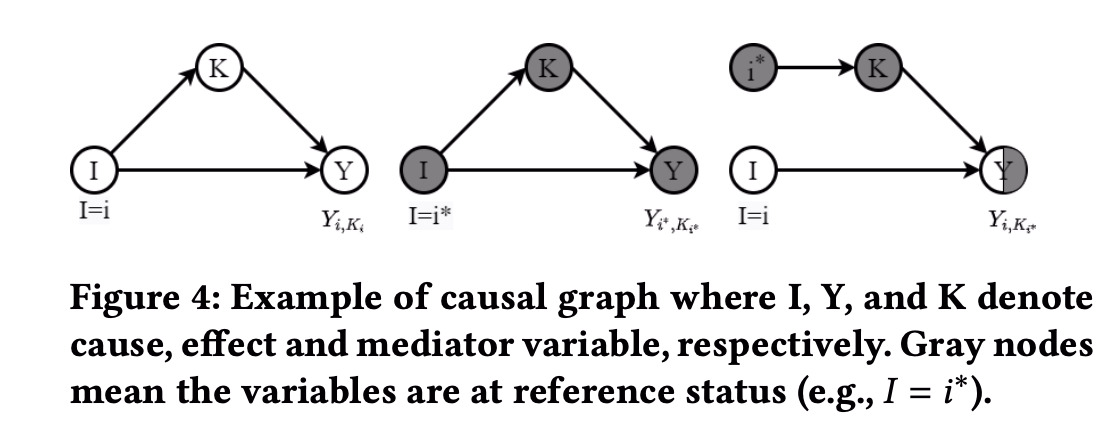
**方案**

**建立因果图**

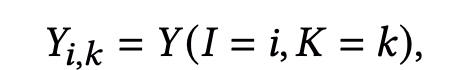
U是用户特征，I是item特征，K是模型打分模块，Y是最终的用户action，0 or 1.



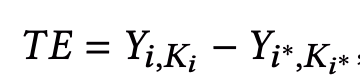
以I为例，图c可以看出I对Y的影响分为 I -> K -> Y，以及 I -> Y，实际中需要去除I->Y这条边。



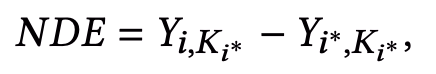
选择一个参考item，比如item特征输入全0 or 不输入。



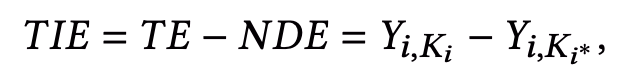
* Total effect(TE): 从参考节点$$$$I^\*
* $$$$到$$$$I
* $$$$，Y的变化，可以视为 所有因素引起的总效应。



* natural direct effect (NDE): 只影响$$$$I^\*
* $$$$-> Y的效应。（部分路径改变，$$$$I
* $$$$-> K -> Y这条路径不变，反事实)。

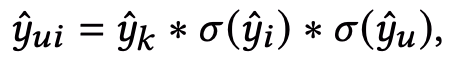


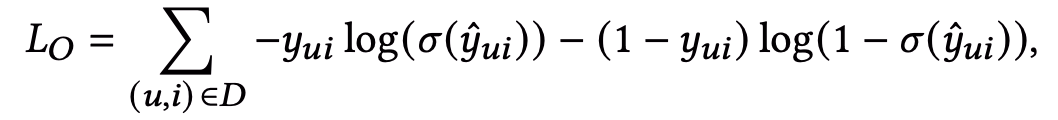
* total indirect effect (TIE) ：代表从I -> K -> Y影响的效应，是去除偏差边的影响。

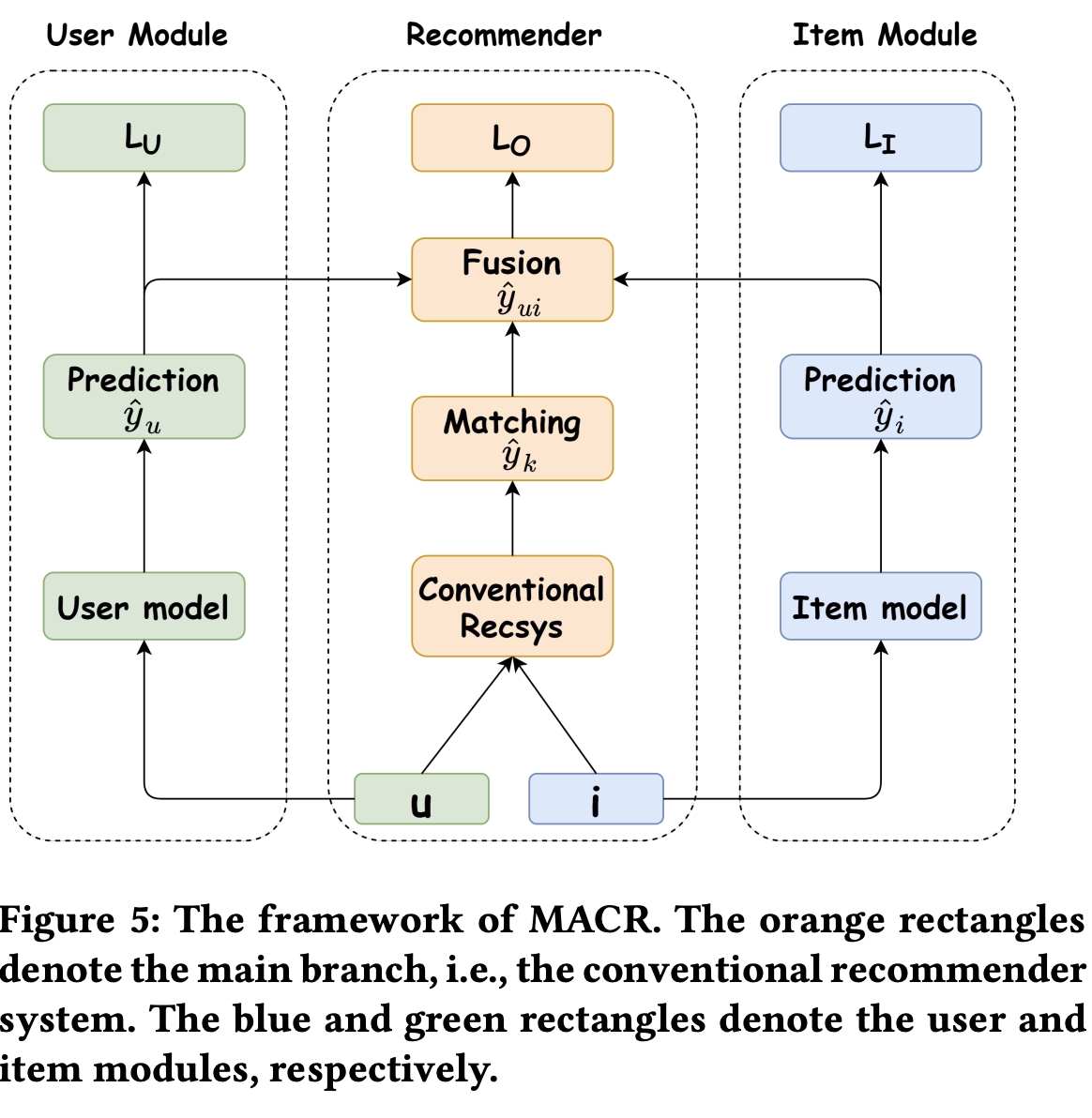


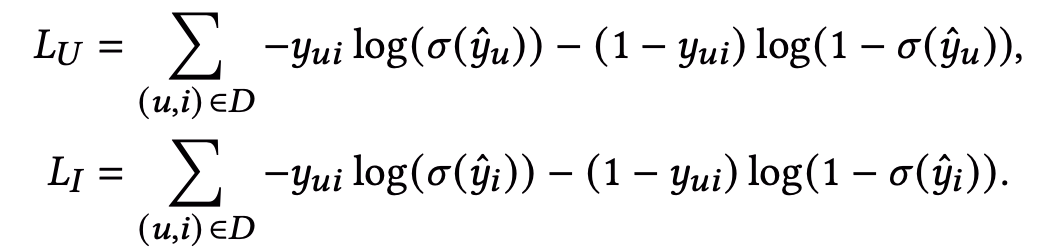
**训练**

假设训练需要预测的函数和loss形式如下，**训练预估函数**是三条边的概率之积。





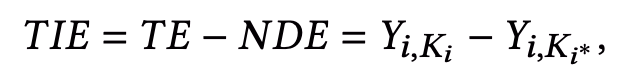




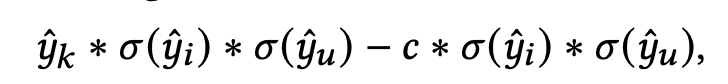


**预估**

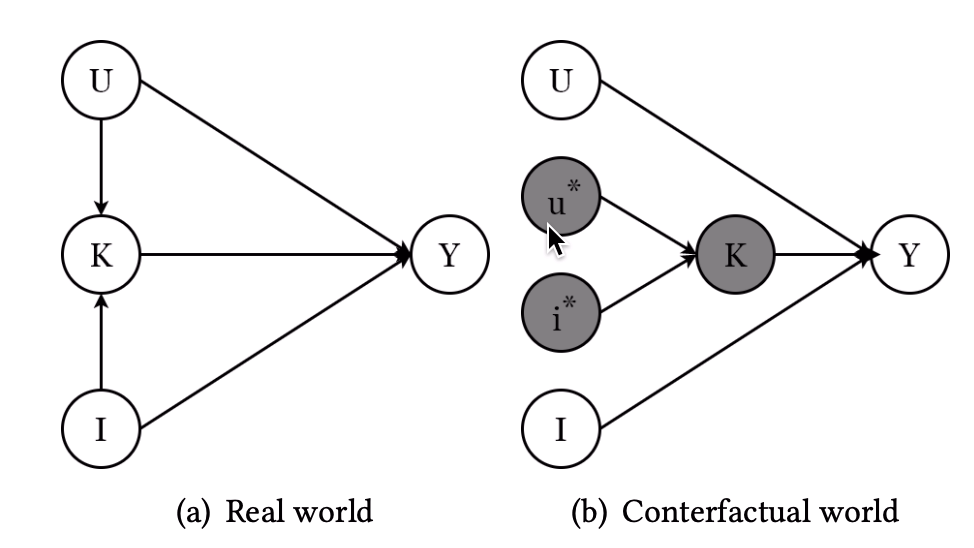
如下所示，根据TIE的公式，也就是说最终的预估结果是 训练预估函数 - 将参考节点只替代k的部分。



也就是



这个c就是将参考节点代入到主网络K的输出，可以视为一个超参进行调节。

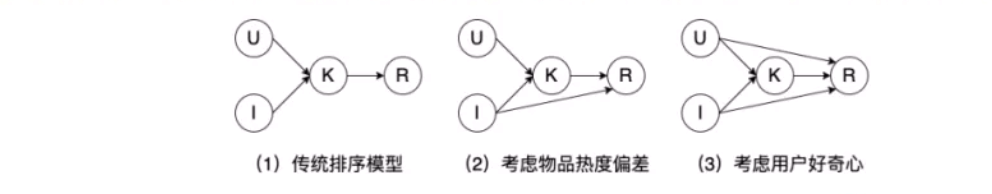


**业界的应用**

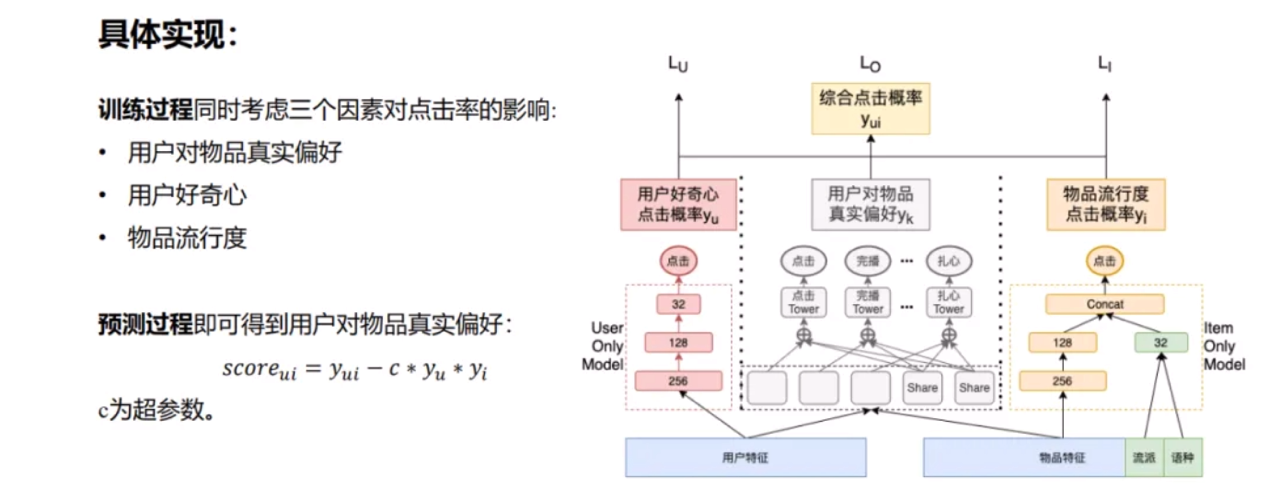
[**QQ音乐排序模型优化**](https://www.6aiq.com/article/1644882812541)

提升用户体验，缓解马太效应。

**因果图**



**方案**

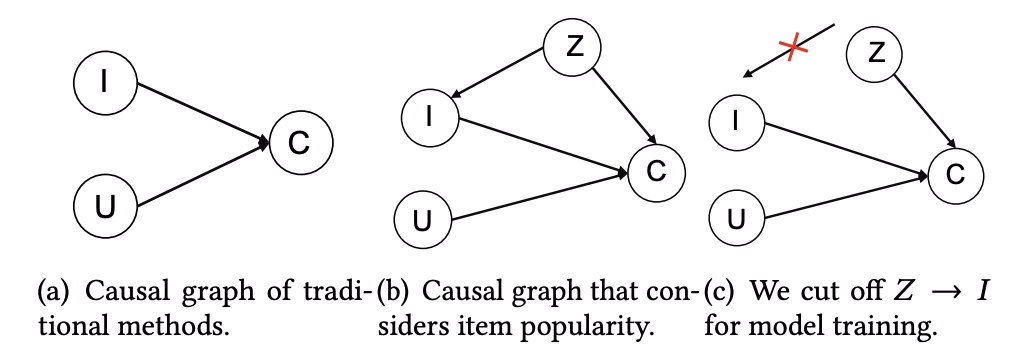


[**Causal Intervention for Leveraging Popularity Bias in Recommendation**](https://arxiv.org/pdf/2105.06067.pdf)

**Insight**

* 流行性会影响item是否曝光，也会影响用户是否交互。
* 流行性的item都是内容比较高的item，所以流行性影响用户是否交互这条边应该保留，而不影响是否曝光。
  + 如下所示，Z是流行性，Z到I 影响I是否曝光的边切断，保留Z影响用户交互的影响。

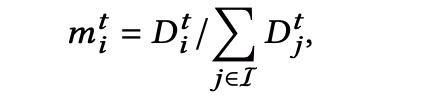
**建立因果图**



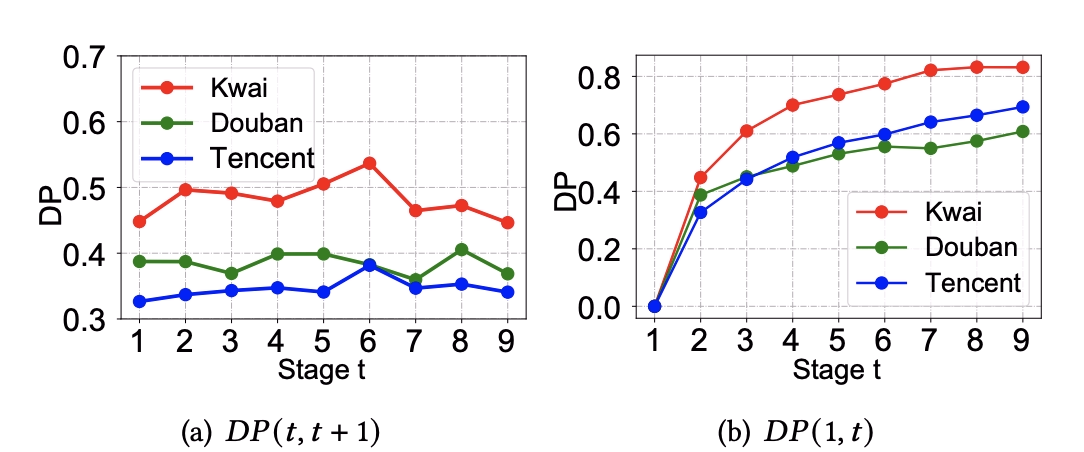
**方案**

**流行度定义**

将时间分为不同的阶段，其中item i在第t个阶段的流行度就是如下所示



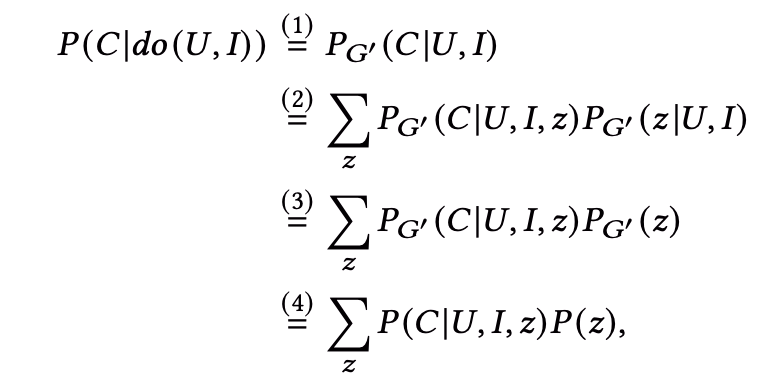
流行是随着时间在漂移的，需要考虑这种影响，比如某个流行的商品突然过季，那么热度可能会逐渐降低。



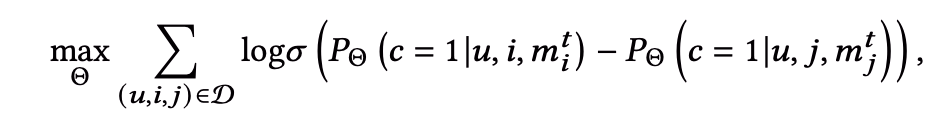
**训练**

通过前面的干预公式知道，如果想切断任何有关I的边，需要对I做干预，$$$$G^{'}

$$$$是切断边之后的图。



训练方式采用pairwise方式，即BPR loss：

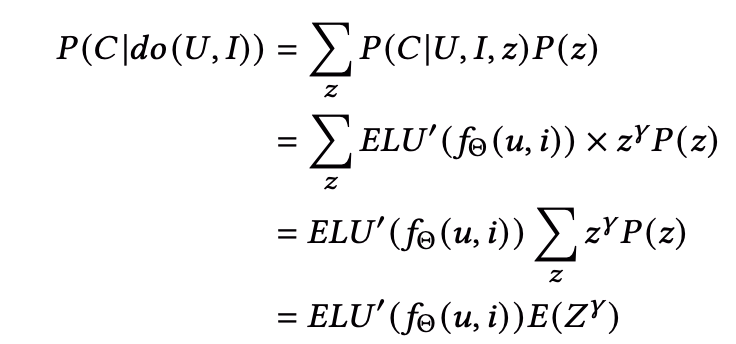


接下来预估$$$$P(C|U,I,Z)

$$$$, 第一步需要将Z解耦，好处是第一方便召回模型使用，第二推理阶段可以不需要inference多次，将其表述为



将其代入上面的do算子中，得到



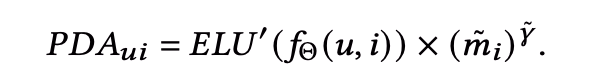
**流行性推理目标调节**

因为已经切断了Z -> I的边，期望在推理的时候可以对一些要流行的item进行boost。则此时引入item的热度预测，可以对流行性item进行好的扶持。这里使用简单的热度预测算子，如下所示，$$$$\alpha

$$$$ 是超参



最终的预估就是



**业界的应用**

[**因果推断在快手推荐场景的应用探索**](https://www.modb.pro/db/331248)