decomposition biased user response in shallow interaction modeling-论文框架

论文框架

topic：decomposition biased user response in shallow interaction modeling

场景：搜索/推荐召回粗排–浅交互网络

样本：user(id, side info(age, sex..))/query(id, side info(intention, brand..))，item(id, title, side info), is_click

模型结构：召回 – 双塔，粗排–非复杂dnn

样本策略：全局负采样，batch内负采样

目标：给出样本马太分布/特征不充分情况下的无偏估计

背景&挑战：

现有推荐模型大量使用用户反馈信号，特征用了大量原子特征（**atom feature–唯一决定一个item，如item id**），数据分布马太效应严重，模型学习有偏，具体的，少量 query/item占据绝大多样本，side info不充分条件下（side info不能完全确定一个id），id embedding会潜在编码未纳入模型的side info，从而使side info学习不充分，模型泛化能力不足

验证方式1–模拟（残差独立性分析，参数有偏分析）：

考虑一个最简单情况，假设reponse和feautre的关系是线性的，权重已知

$$y = w_0 \cdot x_0 + w_1 \cdot x_1 + .. + w_k \cdot x_k$$

设计模拟过程验证问题存在：

1 掩盖一些非原子特征

2 针对其中的原子特征（one-hot之后）以马太方式采样

尝试说明 $E(\widehat{w_i}) \neq w_i$ ，其中 $\widehat{w_i}$ 为模型预估权重

验证方式2-ood样本（测试集上不包含训练集覆盖的item id， 长尾id）：

问题和方案：

1）item id 本身是多义的 -> multi head similairity， kernel representation similarity

2）数据分布马太效应，side info学习不充分 -> instrumental variable内生解偶(附-1)， margin distribution ipw

参考资料

[1] Yao T, Yi X, Cheng D Z, et al. Self-supervised learning for large-scale item recommendations[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021: 4321-4330.

[2] https://en.wikipedia.org/wiki/Instrumental_variables_estimation

附

附-1 内生解偶

基础模型-对比学习

采用谷歌在推荐系统使用对比学习思想[1]修正质量模型结构，其结构如下图，数据增强有两点：1）网络结构drop out 2）输入特征使用特征相关性随机mask，即随机选择一个特征，然后把与其相关性高的特征都mask掉，从而学习label在多个视角上的投影

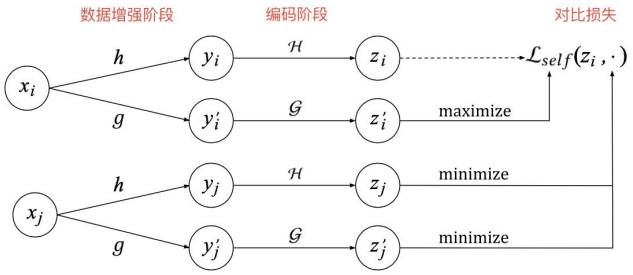


Figure 2: Self-supervised learning framework illustration. Two data augmentations h and g are applied to the input; Encoders $\mathcal{H}$ and $\mathcal{G}$ are applied to the augmented examples $\mathbf{y}_i$ and $\mathbf{y}'_i$ to generate embeddings $\mathbf{z}_i$ and $\mathbf{z}'_i$. The SSL loss $\mathcal{L}_{self}$ w.r.t. $\mathbf{z}_i$ is optimized towards maximizing the similarity with $\mathbf{z}'_i$ while minimizing the similarity between $\mathbf{z}_j$ and $\mathbf{z}'_j$.

重点要mask掉的是“唯一标识样本特征”，与google工作相比，我们不太强调对所有样本mask的随机性，而是选择与原子特征（如query本身，item id等）相关性最高的特征子集合，以较高的概率mask掉

升级模型-内生解耦

对原子特征做进一步思考，一次搜索请求，排序需要决策是item的排序位置，item是treament，response是是否点击，模型使用双塔结构捕捉item和点击之间关系，样本使用用户反馈日志。针对这个设置，我们有两个判断：

1. **原子特征item id等内生的**
2. **非原子特征可以作为instrument variable**

根据之前针对支付宝搜索bias分析，马太效应严重，item对应原子特征item id等在这个场景下是内生的，举个极端例子，两个都非常相关的itemA和B， A有足够的曝光，B没有曝光，模型从日志中学习很容易得出A相关B不相关的结论，而非原子特征内生性相对弱很多，可以作为**instrument variable**，关于内生和instrument variable详细见[2]，附-1给出简单介绍

经济领域一种借助instrumental variable解决变量内生方法是two-stage least squares，先将内生变量投影到instrumental variable，再用response对投影后的内生变量做回归

Stage 1: Regress each column of **X** on **Z**, ($X = Z\delta + \text{errors}$):

$$\hat{\delta} = (Z^{\mathrm{T}} Z)^{-1} Z^{\mathrm{T}} X,$$

and save the predicted values:

$$\widehat{X} = Z\hat{\delta} = Z(Z^{\mathrm{T}} Z)^{-1} Z^{\mathrm{T}} X = P_Z X.$$

In the second stage, the regression of interest is estimated as usual, except that in this stage each endogenous covariate is replaced with the predicted values from the first stage:

Stage 2: Regress **Y** on the predicted values from the first stage:

$$Y = \widehat{X}\beta + \text{noise},$$

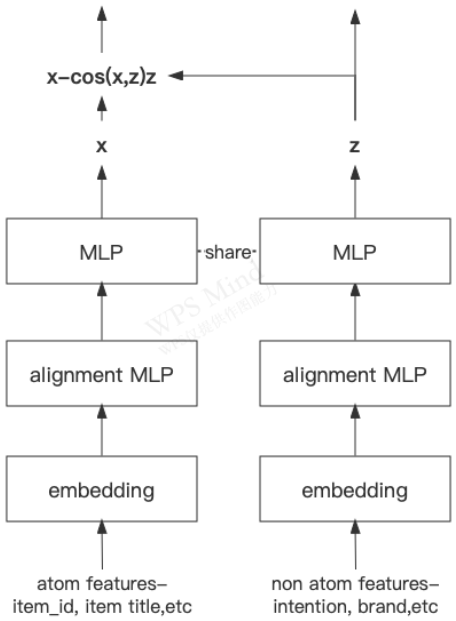
which gives

$$\beta_{2\text{SLS}} = \left(X^{\mathrm{T}} P_Z X\right)^{-1} X^{\mathrm{T}} P_Z Y.$$

stage1投影非常有意思，用机器学习领域方式表达，可以将每个塔表征层拆成两部分，两部分相互对立，第一部分原子特征（内生特征）表征 $\mathbf{x}$ ，第二部分非原子特征（instrumental特征）表征 $\mathbf{z}$ ， $\mathbf{x}$ 可以解偶成两部分 $\mathbf{x} = \mathbf{x}^{\perp} + \widehat{\mathbf{x}}$ ， $\widehat{\mathbf{x}} = \frac{\mathbf{x}^T \mathbf{z}}{||\mathbf{x}||||\mathbf{z}||}\mathbf{z}$ 是 $\mathbf{x}$ 在 $\mathbf{z}$ 上的投影， $\mathbf{x}^{\perp} = \mathbf{x} - \widehat{\mathbf{x}}$ 与 $\mathbf{z}$ 正交，通过对表征层的在计算相似度时候只保留 $\mathbf{x}^{\perp}$ ，这种方式有几个好处

1. $\mathbf{x}^{\perp}$ 更多关注记忆效应， $\mathbf{z}$ 用来做泛化， $\mathbf{x}$ 丢掉 $\widehat{\mathbf{x}}$ ，减少高频交互pair影响，从而使用 $\mathbf{z}$ 学习到更重要权重，增强模型泛化性
2. 可以单独对 $\mathbf{x}^{\perp}$ 加L2正则，约束低频原子特征的噪音影响，在长尾query/item交互上更加稳定
3. instrumental表征 $\mathbf{z}$ 更加稳定，可以结合之前思路，构造ipw，修正z相似损失函数权重修正后loss如下，通过只估计给定非原子特征下该item出现概率，简化ipw估计，减少其方差，同时也有非常好的物理意义，即非原子特征的泛化能力是通过item体现的，而不是通过pv体现，这样就和instrumental variable的定义也联系了起来（instrumental variable 通过变量x起作用）
   a. $l^{'} = l([x_q^{\perp}, z_q], [x_i^{\perp}, z_i]) + \beta \cdot \frac{1}{g(z_i)} \cdot l(z_q, z_i)$ ， 其中 $g(z_i)$ 表示ipwe，
   $E(g(z_i)) = P(i|non\ atom\ features\ of\ i)$

单塔结构变更(以item侧为例)：



增加低频原子特征惩罚loss： $l_{low\_fre} = \frac{1}{m}\sum_i \mathbf{x}_i^2$

增加x和z的regression loss，模拟2 stage least square第一阶段: $l_{regression} = \frac{1}{m}\sum_i (\mathbf{x}_i - \mathbf{z}_i)^2$

附−2

Instrumental variable methods allow for consistent estimation when the explanatory variables (covariates) are correlated with the error terms in a regression model. Such correlation may occur when:

1. changes in the dependent variable change the value of at least one of the covariates ("reverse" causation),
2. there are omitted variables that affect both the dependent and independent variables, or
3. the covariates are subject to non−random measurement error.

First use of an instrument variable occurred in a 1928 book by Philip G. Wright, best known for his excellent description of the production, transport and sale of vegetable and animal oils in the early 1900s in the United States,[5][6] while in 1945, Olav Reiersøl applied the same approach in the context of errors−in−variables models in his dissertation, giving the method its name.[7]

Wright attempted to determine the supply and demand for butter using panel data on prices and quantities sold in the United States. The idea was that a regression analysis could produce a demand or supply curve because they are formed by the path between prices and quantities demanded or supplied. The problem was that the observational data did not form a demand or supply curve as such, but rather a cloud of point observations that took different shapes under varying market conditions. It seemed that making deductions from the data remained elusive.

The problem was that price affected both supply and demand so that a function describing only one of the two could not be constructed directly from the observational data. Wright correctly concluded that he needed a variable that correlated with either demand or supply but not both — that is, an instrumental variable.

After much deliberation, Wright decided to use regional rainfall as his instrumental variable: he concluded that rainfall affected grass production and hence milk production and ultimately butter supply, but not butter demand. In this way he was able to construct a regression equation with only the instrumental variable of price and supply.[8]

Consider for simplicity the single-variable case. Suppose we are considering a regression with one variable and a constant (perhaps no other covariates are necessary, or perhaps we have partialed out any other relevant covariates):

$$y = \alpha + \beta x + u$$

In this case, the coefficient on the regressor of interest is given by $\widehat{\beta} = \frac{\operatorname{cov}(x, y)}{\operatorname{var}(x)}$. Substituting for y gives

$$\begin{aligned}\widehat{\beta} &= \frac{\operatorname{cov}(x, y)}{\operatorname{var}(x)} = \frac{\operatorname{cov}(x, \alpha + \beta x + u)}{\operatorname{var}(x)} \\ &= \frac{\operatorname{cov}(x, \alpha + \beta x)}{\operatorname{var}(x)} + \frac{\operatorname{cov}(x, u)}{\operatorname{var}(x)} = \beta^* + \frac{\operatorname{cov}(x, u)}{\operatorname{var}(x)},\end{aligned}$$

where β^* is what the estimated coefficient vector would be if *x* were not correlated with *u*. In this case, it can be shown that β^* is an unbiased estimator of β. If $\operatorname{cov}(x, u) \neq 0$ in the underlying model that we believe, then OLS gives a coefficient which does *not* reflect the underlying causal effect of interest. IV helps to fix this problem by identifying the parameters β not based on whether x is uncorrelated with u, but based on whether another variable z is uncorrelated with u. If theory suggests that z is related to x (the first stage) but uncorrelated with u (the exclusion restriction), then IV may identify the causal parameter of interest where OLS fails. Because there are multiple specific ways of using and deriving IV estimators even in just the linear case (IV, 2SLS, GMM), we save further discussion for the Estimation section below.