Debiasing item-to-item recommendations with small annotated datasets

RecSys'20

$I \rightsquigarrow$ item inventory

$(S_1, \cdots, S_N) \rightsquigarrow N$ session vectors.

S_k : $|I|$ entries. $S_k[i] = \text{engage}(u_k, i)$ for i in I.

like / share / buy ...

user k

full info setting

can also be viewed : been drawn i.i.d. directly from an overall distribution of session vector $P(S)$

$S_k[i] \in \{0, 1\}$, binary random variable.

Goal of item2item Recommendation

$$\underset{i \in I \setminus j}{\operatorname{argmax}} \; P(S[i] = 1 \mid S[j] = 1)$$

!! Remarks :

① "session" not neccessarily means session , can also be all historical interactions

② can also be extended to topK problem easily

In partial info setting :

$$S_k^{obs}[i] = S_k[i] \times O_k[i] .$$

binary & latent observation variable , $\in \{0,1\}$

$$O_k \in \{0, 1\}^{|I|} \sim P(O \mid X_k)$$

Assumption here : (conditionally independent)

$$P(O[i] = 1 , O[j] = 1 \mid X_k) = P(O[i] = 1 \mid X_k) \cdot P(O[j] = 1 \mid X_k)$$

MLE :

$$\hat{P}^{MLE}(S[i] = 1 \mid S[j] = 1) = \frac{\hat{P}^{MLE}(S[i] = 1 , S[j] = 1)}{\hat{P}^{MLE}(S[j] = 1)}$$

$$= \frac{\cancel{\frac{1}{N}} \sum_k \mathbb{1}(S_k[i] = 1 \wedge S_k[j] = 1)}{\cancel{\frac{1}{N}} \sum_k \mathbb{1}(S_k[j] = 1)}$$

← assumed to be non-zero

‼ remark on this:

① maximizing this eq. tends to find the popular item i.
thus it tends to have greater chance of co-occurence with j.

Inverse Propensity Score

$$p_{k,i} := P(O_{k[i]} = 1 \mid X_k)$$

② still uses the full info setting S.

$$\hat{P}^{IPS}(s_{[i]} = 1 \mid s_{[j]} = 1) = \frac{\sum_k p_{ki}^{-1} p_{kj}^{-1} \mathbb{1}\{ s^{obs}_{k[i]} = 1 \wedge s^{obs}_{k[j]} = 1 \}}{\sum_k p_{kj}^{-1} \mathbb{1}\{ s^{obs}_{k[j]} = 1 \}}$$

→ unbiased estimator of $P(s_{[i]} = 1 \wedge s_{[j]} = 1)$

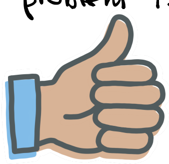
→ unbiased estimator of $P(s_{[j]} = 1)$

$$= Z_j \sum_k p_{ki}^{-1} p_{kj}^{-1} \mathbb{1}\{ s^{obs}_{k[i]} = 1 \wedge s^{obs}_{k[j]} = 1 \}$$

W/ Slutsky's Theorem, we know that the ratio above is a consistent estimator of the true conditional prob.

Now the problem is how to estimate $p_{k,i}$ accurately.

strategy of this paper:



$p_k = f_\theta(x_k)$ → chose a specific function : ! use a model to learn this

learning θ from small annotated dataset

annotated ↓ relevance of items.

Section 4:

Annotated dataset:

$$C_j = \{ (i_1, i_2) \mid rel(i_1 \mid j) > rel(i_2 \mid j) \}$$

$rel(i_1 \mid j)$ ↓ how relevant i is to a given item j.

‼ hard to get the absolute relevance. so only get pairs relative relevance relationship pairs

$rel(i_1 \mid j) > rel(i_2 \mid j)$

$$\Leftrightarrow P(s_{[i_1]} = 1 \mid s_{[j]} = 1) > P(s_{[i_2]} = 1 \mid s_{[j]} = 1)$$

$$\Leftrightarrow \log \hat{P}^{IPS}(s_{[i_1]} = 1 \mid s_{[j]} = 1) - \log \hat{P}^{IPS}(s_{[i_2]} = 1 \mid s_{[j]} = 1) > 0$$

⇒ Hinge Loss :

$$l(C_j;\theta) = \sum_{(i_1,i_2)\in C_j} \max\left\{0, -\log \hat{P}^{IPS}_{\theta}(s[i_1]=1 \mid s[j]=1) + \log \hat{P}^{IPS}_{\theta}(s[i_2]=1 \mid s[j]=1)\right\}$$

$$\Rightarrow \quad \operatorname*{argmin}_{\theta} \frac{1}{L}\sum_{j=1}^{L} l(C_j;\theta)$$

Exponential model:

For $p_k = f_\theta(x_k)$, here we set

$$p_{ki}^{-1} = \exp\left\{\theta^T \cdot \phi(x_k)\right\}$$

x_k : current context

$\phi(\cdot)$: feature map : $x_k \to$ vector representation

$$p_{ij}^{-1} = \exp\left\{\theta^T \cdot \phi(i,j)\right\}$$

Plug this into above equations.

$$\log \hat{P}^{IPS}_{\theta}(s[i]=1 \mid s[j]=1) = \log\Big[Z_j \sum_k \exp\{\theta^T\cdot\phi(k,i)\}\exp\{\theta^T\cdot\phi(k,j)\} \cdot \mathbb{1}\{S^{obs}_{k}[i]=1 \wedge S^{obs}_{k}[j]=1\}\Big]$$

$$= Z_j + \theta^T\cdot\phi(i,j) + \log \underbrace{\sum_k \mathbb{1}\{S^{obs}_{k}[i]=1 \wedge S^{obs}_{k}[j]=1\}}_{:=\mathrm{COUNT}(i,j)}$$

Final Loss Function:

$$l(D_j;\theta) = \sum_{(i_1,i_2)\in C_j} \max\left\{0, -\theta^T\phi(i_1,j) + \log \mathrm{COUNT}(i_1,j) + \theta^T\phi(i_2,j) + \log \mathrm{COUNT}(i_2,j)\right\}$$