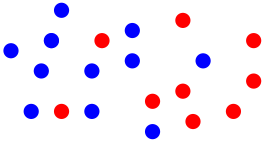
Scalable Personalised Item Ranking through Parametric Density Estimation

Riku Togashi, Masahiro Kato, Mayu Otani, Tetsuya Sakai and Shin'ichi Satoh

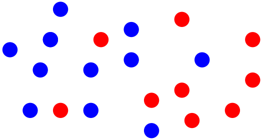
SIGIR '21, July 11–15, 2021

Problem Definition

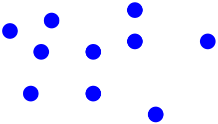


Two-class problem

Problem Definition

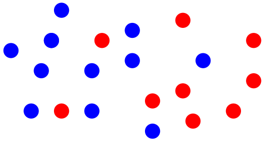


Two-class problem

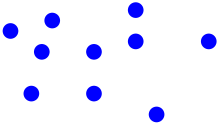


One-class problem

Problem Definition



Two-class problem



One-class problem

Learning from implicit feedback

Learning from one-class problem, i.e. we are given only positive examples.

Approaches

- Pointwise approach
  Solve a binary classification problem to predict if a user prefers an item
  Model: trying to estimate $p(y = +1|u, i)$

Approaches

- Pointwise approach
  Solve a binary classification problem to predict if a user prefers an item
  Model: trying to estimate $p(y = +1|u, i)$
  - Advantage: efficiency
  - Disadvantage: ranking ineffectiveness

Approaches

- Pointwise approach
  Solve a binary classification problem to predict if a user prefers an item
  Model: trying to estimate $p(y = +1|u, i)$
  - Advantage: efficiency
  - Disadvantage: ranking ineffectiveness

- Pairwise ranking:
  A binary classification problem to predict which of any given two items a user will prefer

Approaches

- Pointwise approach
  Solve a binary classification problem to predict if a user prefers an item
  Model: trying to estimate $p(y = +1|u, i)$
  - Advantage: efficiency
  - Disadvantage: ranking ineffectiveness
- Pairwise ranking:
  A binary classification problem to predict which of any given two items a user will prefer
  - Advantage: ranking effectiveness
  - Disadvantage: severely inefficient due to the quadratic computational cost (i.e. problematic in large-scale setting)

Approaches

- Pointwise approach
  Solve a binary classification problem to predict if a user prefers an item
  Model: trying to estimate $p(y = +1|u, i)$
  - Advantage: efficiency
  - Disadvantage: ranking ineffectiveness
- Pairwise ranking:
  A binary classification problem to predict which of any given two items a user will prefer
  - Advantage: ranking effectiveness
  - Disadvantage: severely inefficient due to the quadratic computational cost (i.e. problematic in large-scale setting)
- Special case: IRGAN
  - Advantage: provides an optimized negative sampler using SGD
  - Disadvantages:
    (1) difficult to tune hyper-parameters (2) inefficient training

This paper

Bayes rule

$$p(y = +1|u, i) = \frac{p(i|u, y = +1)p(y = +1|u)}{p(i|u)}$$

This paper

Bayes rule

$$p(y=+1|u,i)=\frac{p(i|u,y=+1)p(y=+1|u)}{p(i|u)}$$

Estimating pdf of positive items for each user using MLE

- Advantages: is efficient and effective in ranking

Notation

$\mathcal{U}$: users; $\mathcal{U} = \{u_j\}_{j=1}^{N}$

$\mathcal{I}$: items; $\mathcal{I} = \{i_j\}_{j=1}^{N}$

data : N i.i.d user feedback logs $(u_1, i_1, y_1), \ldots, (u_N, i_N, y_N)$

y : binary labels, $y \in \{+1, -1\}$. But here $y_j = 1$ for all j.
In implicit feedback we can observe only positive samples.

task : Given a user, a ranker sorts items in the order in which they are most likely to interact with the user.

model : modelling $p(i|u, y = +1)$ or simply $p(i|u)$.

Modelling the problem (personalized ranking)

Using exponential family of distributions

$$p_f(i|u) := p_0(i)e^{f_u(i)-A(f_u)}$$

parameterized by f_u.

Modelling the problem (personalized ranking)

Using exponential family of distributions

$$p_f(i|u) := p_0(i)e^{f_u(i)-A(f_u)}$$

parameterized by f_u.

- p_0: prior density function on $\mathcal{I}$,

Modelling the problem (personalized ranking)

Using exponential family of distributions

$$p_f(i|u) := p_0(i)e^{f_u(i)-A(f_u)}$$

parameterized by f_u.

- p_0: prior density function on $\mathcal{I}$,
- $f_u(i)$: the scoring function to learn (u's preference score for i),

Modelling the problem (personalized ranking)

Using exponential family of distributions

$$p_f(i|u) := p_0(i)e^{f_u(i)-A(f_u)}$$

parameterized by f_u.

- p_0: prior density function on $\mathcal{I}$,
- $f_u(i)$: the scoring function to learn (u's preference score for i),
- $A(f_u) = \log \int_{\mathcal{I}} p_0(i)e^{f_u(i)}di$ (normalization: $\int_{\mathcal{I}} p_f(i|u)di = 1$),

Modelling the problem (personalized ranking)

Using exponential family of distributions

$$p_f(i|u) := p_0(i)e^{f_u(i)-A(f_u)}$$

parameterized by f_u.

- p_0: prior density function on $\mathcal{I}$,
- $f_u(i)$: the scoring function to learn (u's preference score for i),
- $A(f_u) = \log \int_{\mathcal{I}} p_0(i)e^{f_u(i)}di$ (normalization: $\int_{\mathcal{I}} p_f(i|u)di = 1$),
- Assumption: p_0 uniform distribution,

Modelling the problem (personalized ranking)

Using exponential family of distributions

$$p_f(i|u) := p_0(i)e^{f_u(i)-A(f_u)}$$

parameterized by f_u.

- p_0: prior density function on $\mathcal{I}$,
- $f_u(i)$: the scoring function to learn (u's preference score for i),
- $A(f_u) = \log \int_{\mathcal{I}} p_0(i)e^{f_u(i)}di$ (normalization: $\int_{\mathcal{I}} p_f(i|u)di = 1$),
- Assumption: p_0 uniform distribution,
- $p_f(i|u) \propto e^{f_u(i)} \Longrightarrow p_f(i|u) =^{rank} f_u(i)$

Risk

$$R(f) = \mathbb{E}_u\Big[\underbrace{\underbrace{\hat{\mathbb{E}}_+}_{\text{empirical } \mathbb{E} \text{ over } + \text{ items}}}_{\hat{\mathbb{E}}_{p(i|u,y=+1)}} [-\log p_f(i|u)] + \underbrace{\underbrace{KLD\big(p_f(\cdot|u)||p_0(\cdot)\big)}_{\text{Personalization term}}}_{\text{incorporates both + and - items}} \Big]$$

Risk

$$R(f) = \mathbb{E}_u\Big[\underbrace{\underbrace{\hat{\mathbb{E}}_+}_{\text{empirical } \mathbb{E} \text{ over } + \text{ items}}}_{\hat{\mathbb{E}}_{p(i|u,y=+1)}} [-\log p_f(i|u)] + \underbrace{\underbrace{KLD\big(p_f(\cdot|u)||p_0(\cdot)\big)}_{\text{Personalization term}}}_{\text{incorporates both + and - items}} \Big]$$

$$R(f) = \mathbb{E}_u\Big[\hat{\mathbb{E}}_+[-f_u(i)] + \underbrace{\mathbb{E}_{p_f(i|u)}[f_u(i)]}_{\to \text{negative instances in pairwise and pointwise}} \Big]$$

Risk

$$R(f) = \mathbb{E}_u\Big[\underbrace{\underbrace{\hat{\mathbb{E}}_+}_{\text{empirical } \mathbb{E} \text{ over } + \text{ items}}}_{\hat{\mathbb{E}}_{p(i|u,y=+1)}}[-\log p_f(i|u)] + \underbrace{\underbrace{KLD\big(p_f(\cdot|u)||p_0(\cdot)\big)}_{\text{Personalization term}}}_{\text{incorporates both + and - items}}\Big]$$

$$R(f) = \mathbb{E}_u\Big[\hat{\mathbb{E}}_+[-f_u(i)] + \underbrace{\mathbb{E}_{p_f(i|u)}[f_u(i)]}_{\rightarrow\text{negative instances in pairwise and pointwise}}\Big]$$

$$R(f) = \mathbb{E}_u\left[\hat{\mathbb{E}}_+[-f_u(i)] + \frac{\mathbb{E}_{p_0}[f_u(i)e^{f_u(i)}]}{\mathbb{E}_{p_0}[e^{f_u(i)}]}\right]$$

Risk

$$R(f) = \mathbb{E}_u\Big[\underbrace{\underbrace{\hat{\mathbb{E}}_+}_{\text{empirical } \mathbb{E} \text{ over } + \text{ items}}}_{\hat{\mathbb{E}}_{p(i|u,y=+1)}} [-\log p_f(i|u)] + \underbrace{\underbrace{KLD\big(p_f(\cdot|u)||p_0(\cdot)\big)}_{\text{Personalization term}}}_{\text{incorporates both + and - items}} \Big]$$

$$R(f) = \mathbb{E}_u\Big[\hat{\mathbb{E}}_+[-f_u(i)] + \underbrace{\mathbb{E}_{p_f(i|u)}[f_u(i)]}_{\rightarrow\text{negative instances in pairwise and pointwise}} \Big]$$

$$R(f) = \mathbb{E}_u\Big[\hat{\mathbb{E}}_+[-f_u(i)] + \frac{\mathbb{E}_{p_0}[f_u(i)e^{f_u(i)}]}{\mathbb{E}_{p_0}[e^{f_u(i)}]}\Big]$$

Risk Approximation

$$\hat{R}(f, \mathcal{U}_B, \mathcal{I}_B) = \underbrace{\frac{1}{|\mathcal{U}_B|}\sum_{u\in\mathcal{U}_B}}_{\mathbb{E}_u} \Bigg(\underbrace{-\frac{1}{|\mathcal{I}_u^+|}\sum_{i\in\mathcal{I}_u^+} f_u(i)}_{\hat{\mathbb{E}}_+[-f_u(i)]} + \frac{\sum_{i\in\mathcal{I}_B} f_u(i)e^{f_u(i)}}{\sum_{i\in\mathcal{I}_B} e^{f_u(i)}} \Bigg)$$

Relation to IRWGAN

Min-Max objective of IRWGAN for user u:

$$\min_{q_u \in \mathcal{P}} KLD(q_u || p_0) + WD(P_u || Q_u)$$

Relation to IRWGAN

Min-Max objective of IRWGAN for user u:

$$\min_{q_u \in \mathcal{P}} KLD(q_u || p_0) + WD(P_u || Q_u)$$

$$= \min_{q_u \in \mathcal{P}} \left\{ KLD(q_u || p_0) + \sup_{f_u \in \mathcal{F}} \left[\hat{\mathbb{E}}_{+}[f_u(i)] - \mathbb{E}_{q_u(i)}[f_u(i)] \right] \right\}$$

Relation to IRWGAN

Min-Max objective of IRWGAN for user u:

$$\min_{q_u \in \mathcal{P}} KLD(q_u || p_0) + WD(P_u || Q_u)$$

$$= \min_{q_u \in \mathcal{P}} \left\{ KLD(q_u || p_0) + \sup_{f_u \in \mathcal{F}} \left[\hat{\mathbb{E}}_+[f_u(i)] - \mathbb{E}_{q_u(i)}[f_u(i)] \right] \right\}$$

$$\overbrace{=}^{\textit{duality}} \sup_{f_u \in \mathcal{F}} \left\{ \hat{\mathbb{E}}_+[f_u(i)] + \min_{q_u \in \mathcal{P}} \left[KLD(q_u || p_0) - \mathbb{E}_{q_u(i)}[f_u(i)] \right] \right\}$$

Relation to IRWGAN

Min-Max objective of IRWGAN for user u:

$$\min_{q_u \in \mathcal{P}} KLD(q_u || p_0) + WD(P_u || Q_u)$$

$$= \min_{q_u \in \mathcal{P}} \left\{ KLD(q_u || p_0) + \sup_{f_u \in \mathcal{F}} \left[\hat{\mathbb{E}}_+[f_u(i)] - \mathbb{E}_{q_u(i)}[f_u(i)] \right] \right\}$$

$$\overbrace{=}^{\textit{duality}} \sup_{f_u \in \mathcal{F}} \left\{ \hat{\mathbb{E}}_+[f_u(i)] + \min_{q_u \in \mathcal{P}} \left[KLD(q_u || p_0) - \mathbb{E}_{q_u(i)}[f_u(i)] \right] \right\}$$

$\mathcal{P}$: probability measures supported in $\mathcal{I}$

$\mathcal{F}$: a convex subset of 1-Lipschitz functions

Relation to IRWGAN

Min-Max objective of IRWGAN for user u:

$$\min_{q_u \in \mathcal{P}} KLD(q_u || p_0) + WD(P_u || Q_u)$$

$$= \min_{q_u \in \mathcal{P}} \left\{ KLD(q_u || p_0) + \sup_{f_u \in \mathcal{F}} \left[\hat{\mathbb{E}}_{+}[f_u(i)] - \mathbb{E}_{q_u(i)}[f_u(i)] \right] \right\}$$

$$\overbrace{=}^{\textit{duality}} \sup_{f_u \in \mathcal{F}} \left\{ \hat{\mathbb{E}}_{+}[f_u(i)] + \min_{q_u \in \mathcal{P}} \left[KLD(q_u || p_0) - \mathbb{E}_{q_u(i)}[f_u(i)] \right] \right\}$$

$\mathcal{P}$: probability measures supported in $\mathcal{I}$
$\mathcal{F}$: a convex subset of 1-Lipschitz functions

$$R_{WGAN}(f) = \mathbb{E}_u \left[\hat{\mathbb{E}}_{+}[-f_u(i)] + \underbrace{\mathbb{E}_{p_f(i|u)}[f_u(i)]}_{\textit{optimal } q_u^*(i) = p_f(i|u)} \right]$$

Relation to IRWGAN

Min-Max objective of IRWGAN for user u:

$$\min_{q_u \in \mathcal{P}} KLD(q_u || p_0) + WD(P_u || Q_u)$$

$$= \min_{q_u \in \mathcal{P}} \left\{ KLD(q_u || p_0) + \sup_{f_u \in \mathcal{F}} \left[\hat{\mathbb{E}}_+[f_u(i)] - \mathbb{E}_{q_u(i)}[f_u(i)] \right] \right\}$$

$$\overbrace{=}^{\text{duality}} \sup_{f_u \in \mathcal{F}} \left\{ \hat{\mathbb{E}}_+[f_u(i)] + \min_{q_u \in \mathcal{P}} \left[KLD(q_u || p_0) - \mathbb{E}_{q_u(i)}[f_u(i)] \right] \right\}$$

$\mathcal{P}$: probability measures supported in $\mathcal{I}$
$\mathcal{F}$: a convex subset of 1-Lipschitz functions

$$R_{WGAN}(f) = \mathbb{E}_u \left[\hat{\mathbb{E}}_+[-f_u(i)] + \underbrace{\mathbb{E}_{p_f(i|u)}[f_u(i)]}_{\text{optimal } q_u^*(i) = p_f(i|u)} \right]$$

$$R^*_{WGAN}(f) = R(f)$$

Datasets statistics and effectiveness (nDCG)

Dataset	User #	Item #	Interaction #	Density
Gowalla	29,858	40,981	1,027,370	0.00084
Yelp2018	31,668	38,048	1,561,406	0.00130
Amazon-Book	52,643	91,599	2,984,108	0.00062

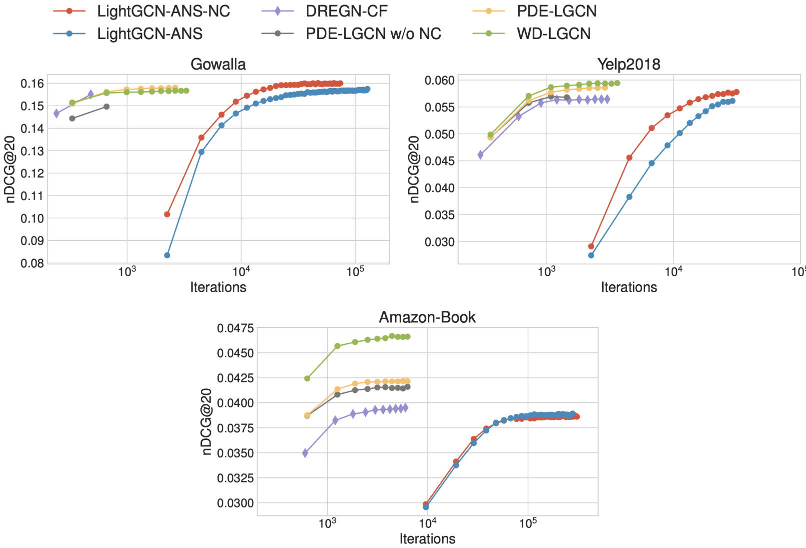
Datasets statistics and effectiveness (nDCG)

Dataset	User #	Item #	Interaction #	Density
Gowalla	29,858	40,981	1,027,370	0.00084
Yelp2018	31,668	38,048	1,561,406	0.00130
Amazon-Book	52,643	91,599	2,984,108	0.00062

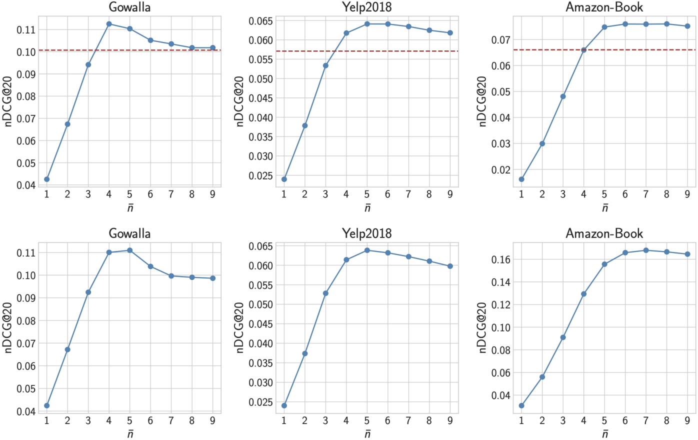
Method	Gowalla		Yelp2018		Amazon-Book	
	R@20	nDCG@20	R@20	nDCG@20	R@20	nDCG@20
NGCF	0.1567	0.1325	0.0575	0.0474	0.0342	0.0262
Mult-VAE	0.1644	0.1340	0.0589	0.0458	0.0413	0.0306
ENMF	0.1512	0.1306	0.0628	0.0515	0.0344	0.0272
LightGCN	0.1828	0.1551	0.0651	0.0532	0.0421	0.0324
DREGN-CF	0.1828	0.1551	0.0685	0.0564	0.0506	0.0395
PDE-LGCN	0.1871	0.1575	0.0706	0.0582	0.0542	0.0422
WD-LGCN	0.1859	0.1567	0.0719	0.0594	0.0580	0.0466

PDE-LGCN: parametric density estimation LightGCN **WD-LGCN:** Wasserstein distance LightGCN

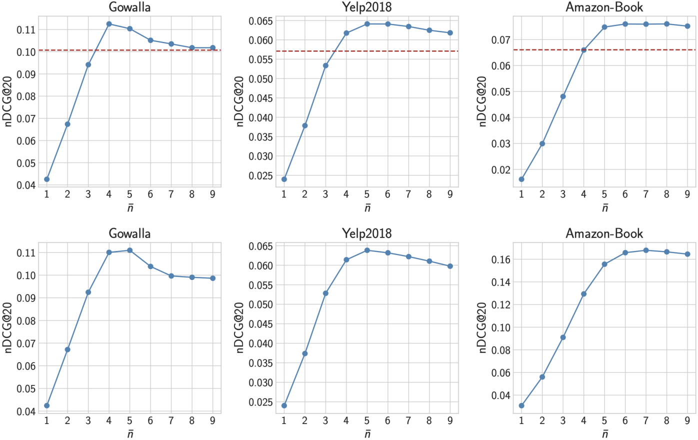
Efficiency


LightGCN-ANS-NC
LightGCN-ANS
DREGN-CF
PDE-LGCN w/o NC
PDE-LGCN
WD-LGCN
Gowalla
Yelp2018
Amazon-Book
nDCG@20
Iterations
10^3
10^4
10^5
0.16
0.15
0.14
0.13
0.12
0.11
0.10
0.09
0.08
0.060
0.055
0.050
0.045
0.040
0.035
0.030
0.0475
0.0450
0.0425
0.0400
0.0375
0.0350
0.0325
0.0300

nDCG in terms of clipping n



nDCG in terms of clipping n



$$\text{clip}(\mathbf{e}, \bar{n}) := \begin{cases} \mathbf{e}, & \text{if } \|\mathbf{e}\| \leq \bar{n} \\ \frac{\bar{n}}{\|\mathbf{e}\|}\mathbf{e}, & \text{if } \|\mathbf{e}\| > \bar{n} \end{cases}$$

What I learned?

What I learned?

- We can learn from only positive labels (i.e. implicit feedback)!

What I learned?

- We can learn from only positive labels (i.e. implicit feedback)!
- Density functions can help to learn efficiently and effectively.

What I learned?

- We can learn from only positive labels (i.e. implicit feedback)!
- Density functions can help to learn efficiently and effectively.
- Exponential family of distributions is a powerful tool for learning to rank.

What I learned?

- We can learn from only positive labels (i.e. implicit feedback)!
- Density functions can help to learn efficiently and effectively.
- Exponential family of distributions is a powerful tool for learning to rank.
- Personalized term in loss functions can be done through KLD with uniform distribution which simultaneously encodes $+$ and $-$ examples implicitly.

QUESTIONS?