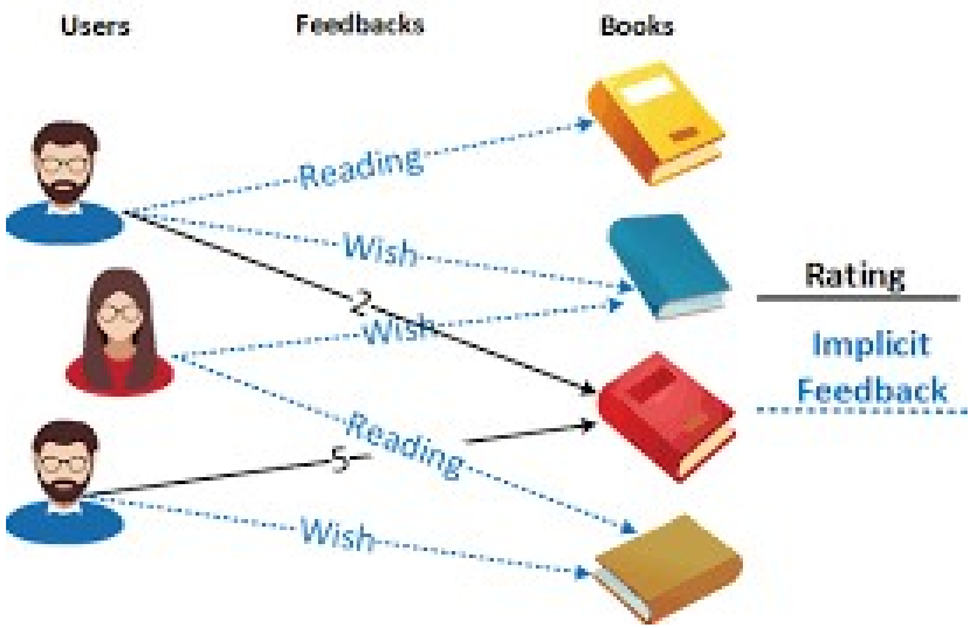
Graduate Certificate in Big Data Analytics

Working with Implicit User Feedback



Dr. Barry Shepherd
Institute of Systems Science
National University of Singapore
Email: barryshepherd@nus.edu.sg

Working with Implicit User Feedback

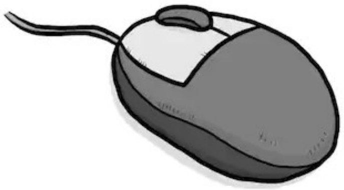
- Issues with Explicit Ratings
  - Very sparse - users tend to be lazy, often don't bother to rate
  - Users may not always tell the truth? E.g. Influenced by peer/friend opinions (all of ,my friends liked that movie so maybe it was good after all!)
  - May not be available at all if no system in place to collect them
  - Watching what the user actually does (e.g. what they view or buy) may be more reliable / accurate than ratings

Examples of Implicit Ratings

- Buying a product
- Viewing a (product) page
- Clicking on a link
- Time spend looking at a page
- Repeat visits
- Referring a page to others

Issues with Implicit Ratings

- Buying something doesn't always mean liking something
- Could be a mistake buy or impulse buy or mistaken page-click etc.
- I may be buying for someone else using my own account



Working with Implicit User Feedback

- Visits to webpages is a commonly used implicit feedback signal:
  - Repeat visits to a page implies liking the page/product (more repeats => more likes)
  - More time on page implies liking the page/product (longer duration => more like)

User (or Session)	page1	page2	page3	page4	page5	page6
1	2	5	4		3	1
2		3		5	3	1
3			5	3		

Implicit ratings matrix

E.g. Assume the ratings here refer to the time spent on page (normalised to 1-5)

- This looks similar to regular (explicit) ratings. Can we proceed as before?
  - We can try – this often works. But, since all users with implicit ratings like the product/page to some degree, if we predict the implicit rating we are predicting the amount of like and not like/dislike

Implicit ratings: Amount of like, 1 → 5

Explicit ratings: dislike 1 → 5 like

Implicit ratings are often treated as Binary

- Its common to assume ANY page view (or similar) is a like, else don't know

User (or Session)	page1	page2	page3	page4	page5	page6
1	2	5	4		3	1
2		3		5	3	1
3			5	3		

User (or Session)	page1	page2	page3	page4	page5	page6
1	1	1	1		1	1
2		1		1	1	1
3			1	1		

- But now we can't use Cosine Similarity (or similar) since result will always be 1

$$\textit{Cosine Similarity} = (1*1 + 1*1 + \dots) / \sqrt{(1^2+1^2 + \dots)*(1^2+1^2 + \dots)} = N/N = 1$$

$$\textit{Euclidean Similarity} = 1 / (1 + \sqrt{(1-1)^2 + (1-1)^2 + \dots}) = 1/1 = 1$$

- To apply Cosine or Euclidean we must assume the NA's => 0 (don't like)
  - Can often work well. e.g. RecommenderLab (an R lib) makes this assumption.

Binary Ratings: Jaccard Similarity

- Measures the similarity between two sets
- Makes no assumptions about the missing values

$$Sim_{jaccard}(A,B) = |A \cap B| / |A \cup B|$$

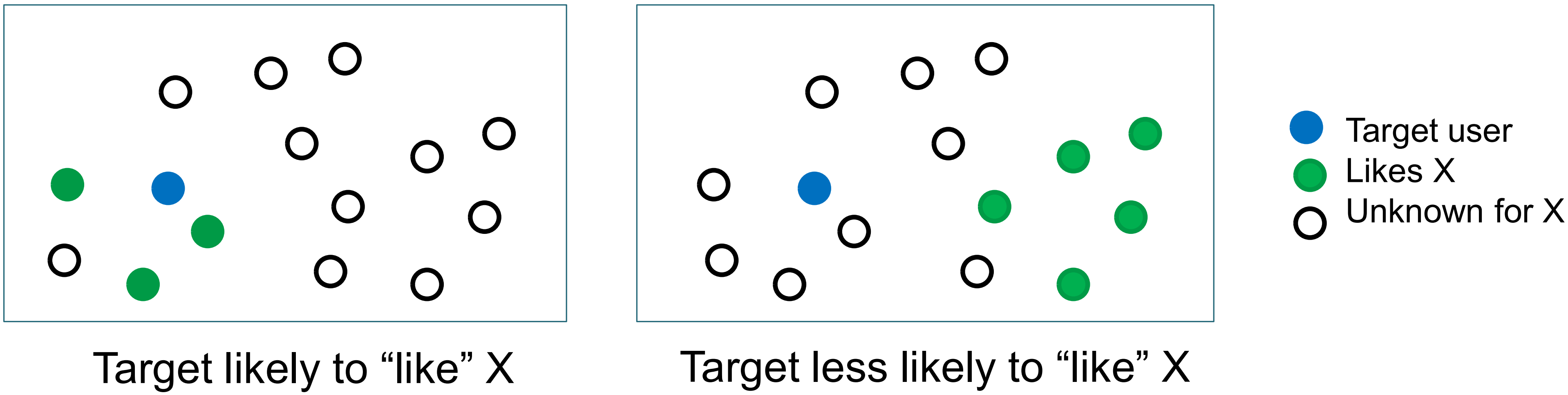
User	b1	b2	b3	b4	b5	b6	b7	b8	b9	b10
U1	1	1	1				1			
U2		1			1	1	1	1		

What is the Jaccard similarity for (u1,u2)?

Binary Ratings: User (or Item) based CF?

- *Problem:* For an unseen item (X), the weighted average rating of the target user's neighbor's on that item is always 1 (the weighted average of many 1's is 1) . So how do we rank all of the unseen items in order to make a recommendation?

- *Possible Solution:* rank unseen items using the average similarity of the target to the users who liked the item*
  - Rational: high similarity suggests that the target user may also like X



An alternative/better approach is to use a MF algorithm customised to impact feedback…

Working with Integer Implicit Ratings

- Deskdrop is an internal communications platform that allows companies employees to share relevant articles with their peers, and collaborate around them.
- The logged data includes the user interactions with the platform, including the type of interaction

The eventType values are:

- **VIEW**: The user has opened the article. A page view in a content site can mean many things. It can mean that the user is interested, or maybe user is just lost or clicking randomly.
- **LIKE**: The user has liked the article.
- **BOOKMARK**: The user has bookmarked the article for easy return in the future. This is a strong indication that the user finds something of interest.
- **COMMENT CREATED**: The user left a comment on the article.
- **FOLLOW**: The user chose to be notified on any new comment about the article.

Create an integer implicit ratings variable:

```
event_type_strength = {
    'VIEW': 1.0,
    'LIKE': 2.0,
    'BOOKMARK': 3.0,
    'FOLLOW': 4.0,
    'COMMENT CREATED': 5.0,
}
```

We will explore this in workshop4

ALS Matrix Factorisation with Implicit Feedback

- Recall an implicit like ~ viewing a product page, clicking an item, purchase etc.
- BUT… there is no dislike signal (unlike explicit ratings)

Implicit Ratings matrix

	i1	i2	i3	i4	i5	i6	i7	i8
u1		4			5			
u2	1				2		5	
u3		3	4	3		5		
u4	2	2			5	5		3
u5			1			4		2

Convert to preferences
P =1 if implicit rating >0 else 0

Preference matrix

	i1	i2	i3	i4	i5	i6	i7	i8
u1	0	1	0	0	1	0	0	0
u2	1	0	0	0	1	0	1	0
u3	0	1	1	1	0	1	0	0
u4	1	1	0	0	1	1	0	1
u5	0	0	1	0	0	1	0	1

Assign a confidence to each preference:
C = 1 + α*rating *(α ~ 40)*
e.g. more views => more confidence

Proceed as with ALS but with a new cost function:

$$\min_{x_\star, y_\star} \sum_{u,i} c_{ui}(p_{ui} - x_u^T y_i)^2 + \lambda \left(\sum_u \|x_u\|^2 + \sum_i \|y_i\|^2 \right)$$

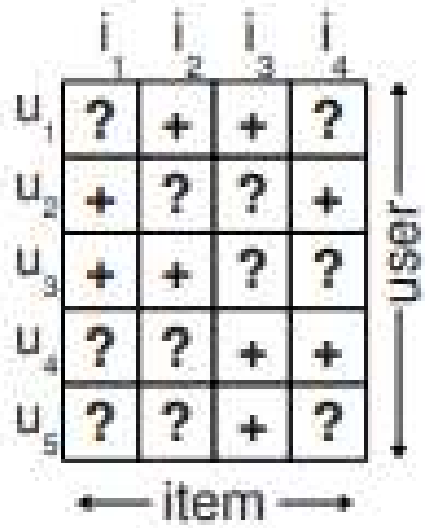
where: u ~ users, i ~ items, x is the user factors matrix, y is the item factors matrix

**Note that the preference matrix has no missing values, so optimisation process is different and more time consuming.*

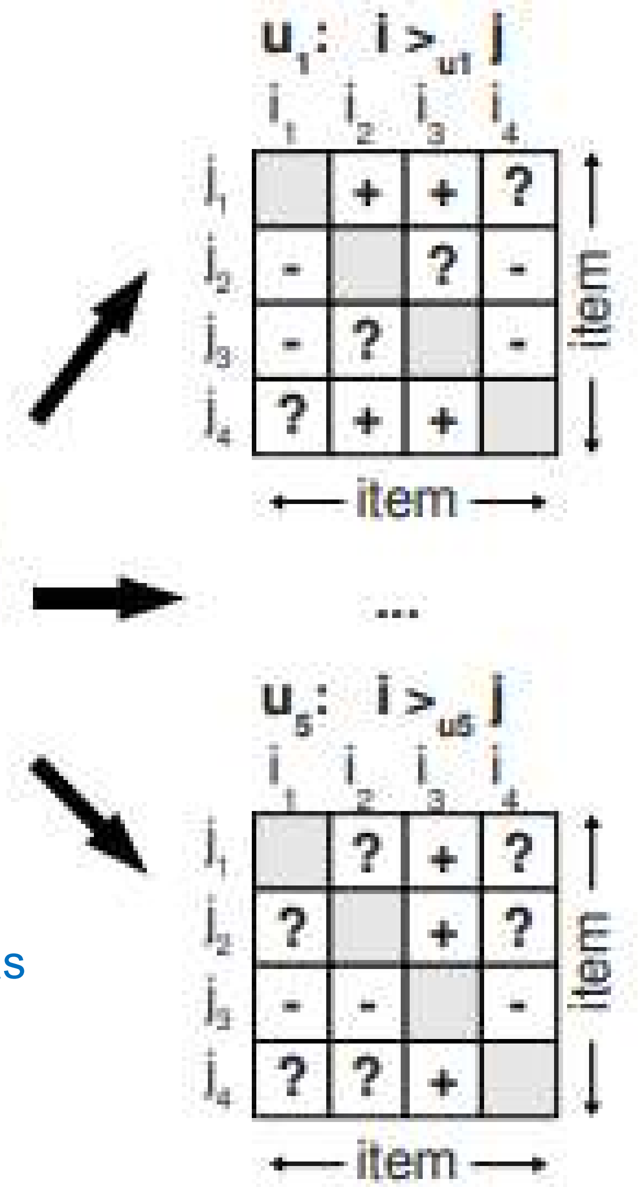
http://yifanhu.net/PUB/cf.pdf

Matrix Factorisation using Bayesian Personalised Ranking (BPR)

- Does not ignore missing values (or treat them as zeros), instead assumes that the user prefers the observed (positive) item over all other non-observed items.
- Converts the observed implicit preference data into a pairwise preference matrix for each user
- Instead of minimising the error for predicted ratings it minimises the ranking error between pairs of items



The observed data.
+ indicates the item was observed (or had other positive feedback)



The pairwise preference matrices
+ indicates user prefers item i over j
- indicates user prefers item j over i

https://arxiv.org/ftp/arxiv/papers/1205/1205.2618.pdf

Workshop4

- Intro to Spark ML library
- Generate an integer implicit rating from individual implicit signals
- Compare explicit and implicit ALS