Project – ACM RecSys Challenge 2020

Recommender Systems

VU 194.035, 2020S

Dimitris Sacharidis

TU WIEN Informatics

electronic commerce

ACM RecSys Challenge

- in cooperation with the ACM RecSys Conference
- real-world (industry) recommendation task with real data

- competition with leaderboard (held-out data)
- prizes, invitation to write a report

ACM RecSys Challenge 2020

- Industry Organizer: Twitter
- Predict Tweet Engagement

- Information about the challenge:
  - http://www.recsyschallenge.com/2020/

- Information about the data and leaderboard:
  - https://recsys-twitter.com

ACM RecSys Challenge 2020

- task: predict the **user** engagement with a **tweet**
  - four different types of engagement
    - Likes, Replies, Retweets, and Retweets with comments

- information available about
  - the **tweet**
    - Tweet id, Text tokens, Hashtags, Present media/links/domains, Tweet type, Language, Timestamp
  - the **engaging** user (who interacts-engages with the tweet)
    - User id, Follower count, Following count, Is verified?, Account creation time
  - the **engaged-with** user (who is the owner of the tweet)
    - same as for engaging user
  - the **engagement**
    - Engaged-with follows engaging?, Timestamps

Data Fields

- each training example is about an **engaging user** (*user*) interacting with a **tweet** (*item*)

- contains additional information about
  - the **engaging user** (user profile)
  - the metadata of the **tweet** (item content), including the user profile of the **engaged-with user**
  - the follow relationship between **engaged-with** and **engaging user**

- label is the timestamp of the **engagement** (implicit feedback)
  - for: like, reply, retweet, retweet with comment

Data Fields

tweet | engaged user

	text_tokens	hashtags	tweet_id	present_media	present_links	present_domains	tweet_type	language	tweet_timestamp	engaged_with_user_id	engaged_with_user_follower_count
0	[101, 56898, 137, 174, 63247, 10526, 131, 3197...	NaN	3C21DCFB8E3FEC1CB3D2BFB413A78220	[Video]	NaN	NaN	Retweet	76B8A9C3013AE6414A3E6012413CDC3B	2020-02-12 00:28:43	D1AA2C85FA644D64346EDD88470525F2	737
1	[101, 102463, 10230, 10105, 21040, 10169, 1281...	NaN	3D87CC3655C276F1771752081423B405	NaN	[BB422AA00380E45F312FD2CAA75F4960]	[92D397F8E0F1E77B36B8C612C2C51E23]	TopLevel	D3164C7FBCF2565DDF915B1B3AEFB1DC	2020-02-06 07:49:51	4DC65AC7BD963DE1F7617C047C33DE99	52366425
2	[101, 56898, 137, 11255, 22037, 10263, 168, 11...	[DB32BD91C2F1B37BE700F374A07FBC61]	3701848B96AA740528A2B0E247777D7D	NaN	[2423BA02A75DB2189335DDC3FB6B74A1]	[6D323BE93766E79BE423FAC5C28BE39B]	Retweet	22C448FF81263D4BAF2A176145EE9EAD	2020-02-09 14:07:12	5C671539CB41B9807E209349[illegible]	988
3	[101, 13073, 28757, 106, 100, 14120, 131, 120,...	NaN	18176C6AD2871729384062F073CCE94D	[Video]	NaN	NaN	TopLevel	D3164C7FBCF2565DDF915B1B3AEFB1DC	2020-02-08 12:18:12	70B900BE17416923D1E236A38798F202	1228134
4	[101, 3460, 1923, 6632, 2824, 30368, 2179, 188...	NaN	AF11AF01F842E7F120667B7B0B38676D	NaN	NaN	NaN	Quote	22C448FF81263D4BAF2A176145EE9EAD	2020-02-09 07:34:10	E94C0E9E8494F3D603F9D1A5C5242E3D	73

engaged user | engaging user | engagement

engaged_with_user_following_count	engaged_with_user_is_verified	engaged_with_user_account_creation	engaging_user_id	engaging_user_follower_count	engaging_user_following_count	engaging_user_is_verified	engaging_user_account_creation	engaged_follows_engaging	reply_timestamp	retweet_timestamp	retweet_with_comment_timestamp	like_timestamp
706	False	1403069820	000046C8606F1C3F5A7296222C88084B	131	2105	False	2019-11-17 08:11:09	False	NaT	NaT	NaT	NaT
2383	True	1230139136	00006047187D0D18598EF12A650E1DAC	22	50	False	2012-06-26 01:26:02	False	NaT	NaT	NaT	NaT
167	False	1530094483	0000648BAA193AE4C625DDF789B57172	251	[illegible]	False	2016-02-26 08:01:11	False	NaT	NaT	NaT	NaT
5413	False	1378699943	000071667F50BAFEA722A8E8284581E5	18	58	False	2013-09-06 00:32:44	False	NaT	NaT	NaT	2020-02-10 03:29:24
299	False	1549054499	00007745A6EE969F1A0F44B10DC17671	268	526	False	2009-09-07 03:40:00	False	NaT	NaT	NaT	NaT

labels

Prediction Task

- goal is to **predict** for each example the **engagement**

- specifically, you will predict and report the **probability of engagement**
  - separately for each engagement type
  - we do not want to predict the timestamp

- output format:
  - **engaging user id**, **tweet id**, **probability**
    - separately for each engagement type

Evaluation

- for each example:
  - you predict a probability,
  - and the ground truth is 0/1 (whether there is an engagement timestamp)

- evaluate as a binary classification task in two ways

1. Area Under Precision-Recall Curve
   - generate precision-recall pairs for various probability thresholds
     - assumes anything above threshold is predicted as relevant
2. Cross-Entropy Loss = Log-Loss (for binary classification)
   - measure how good the predicted probabilities are

For the Project

Data and Evaluation

- we provide a small sample of the data to work on
  - ~80,000 examples for 10,000 engaging users
  - you may sample it down even further
  - you may work with one engagement type

- we provide some sample code to parse data

- you will split it yourself into train and test subsets

- you will implement the evaluation metrics
  - using the provided code at https://recsys-twitter.com/code/snippets

Expected Work

- implement a couple of different recommendation approaches
  - go beyond pure collaborative filtering approaches (that only work with engaging user id and tweet id)
  - preprocess the data accordingly

- design an evaluation protocol
  - split into train/test (maybe also into a dev set)
  - decide on range of hyperparameter values to explore
  - implement evaluation metrics

What to Submit

- code
  - in any language you wish
- written report
  - detail all your work, thought process, decisions made
  - show evaluation results based on your protocol
  - draw conclusions

Collaboration across groups

- you may optionally share **code** that handles some basic **tasks**
    - make it public for other groups to use

- rewarded with **project points**

- tasks should **not** be recommendation approaches, but rather helper functions

- consider the following sample predefined tasks (T1-T5)
    - or propose other tasks

T1. Split into train, dev, test

- Sub-sample to create **test**, non-test datasets
- Optionally split non-test into **train** and **dev**
  - e.g., to implement k-fold validation

T2. Evaluation

- Parse test to create the ground truth output file
  - **engaging user id**, **tweet id**, **label**

- Implement the `read_predictions` function from https://recsys-twitter.com/code/snippets

T3. Create a Ratings Matrix

- for use in pure collaborative filtering approaches
- a matrix for each engagement type
- implicit feedback: 0 or 1

T4. Extract the Social Network

- Twitter Social Network is directional (follower – following)

- parse the `engaged_follows_engaging` field
  - each example gives you an edge

- Create the **adjacency matrix** representation of the social graph
  - 1 if an edge exists between two users, 0 otherwise

- how can you use this information?

T5. Implement a Baseline

- Implement the neural network approach described in the challenge paper:
  - https://arxiv.org/abs/2004.13715