**Machine Learning for Personalization**
Project

# 1 PROJECT DESCRIPTION

Try to make an *innovative contribution* to the field of *machine learning and personalization*. Unlike the assignments, for the project there is no fixed recipe to follow. Rather, you are free to pick a topic and direction that you find motivating and to leverage the tools covered in class.

One place to start is to consider a state-of-the-art paper in the field and explore implementing one and perhaps extending it. In particular, consider papers that have been published in top machine learning conferences and personalization conferences in the past 10 years. There are already dozens of papers linked on the class home-page

`http://www.cs.columbia.edu/~jebara/6998`

that we have discussed during the lectures. Those are all good candidates for a project. Alternatively, consider the following recent papers and simply search for their titles in Google to get the PDF

```
DropoutNet: addressing coldstart in recommender systems (NIPS 2017)
Recursive partitioning for personalization using observational data
On the optimization landscape of tensor decomposition (NIPS 2017)
Understanding how people use natural language to ask for recommendations (RECSYS 2017)
Translation-based recommendations (RECSYS 2017)
Contextual RNNs for recommendation (RECSYS 2017)
MPR: Multi-Objective Pairwise Ranking (RECSYS 2017)
Folding: Why good models sometimes make spurious recommendations (RECSYS 2017)
Neural attentive session-based recommendation (CIKM 2017)
Joint representation learning for top-N recommendation ... (CIKM 2017)
Deep IV: A flexible approach for counterfactual prediction (ICML 2017)
Collaborative recurrent autoencoder: recommend while learning to fill in the blanks (NIPS 2016)
Deconvolving feedback loops in recommender systems (NIPS 2016)
Optimal tagging with Markov chaing optimization (NIPS 2016)
Scaled least-squares estimator for GLMs in large-scale problems (NIPS 2016)
Optimistic bandit convex optimization (NIPS 2016)
Learning representations for counterfactual inference (ICML 2016)
Recommendations as treatments: debiasing learning and evaluation (ICML 2016)
A neural autoregressive approach to collaborative filtering (ICML 2016)
Adaptive, personalized diversity for visual discovery (RECSYS 2016)
A scalable approach to periodical personalized recommendations (RECSYS 2016)
Fifty shades of ratings: how to benefit from a negative feedback in top-N ... (RECSYS 2016)
Local item-item models for top-N recommendation (RECSYS 2016)
Self-normalized estimator for counterfactual learning (NIPS 2015)
Efficient Thompson sampling for online matrix-factorization recommendation (NIPS 2015)
Context-aware event recommendation in event-based social networks (RECSYS 2015)
Dynamic Poisson factorization (RECSYS 2015)
Top-N recommendation with missing implicit feedback (RECSYS 2015)
Explore-exploit in top-N recommender systems via Gaussian processes (RECSYS 2014)
```

You can find additional papers in leading conference and journal articles such as the following list.

Search for the conference name and the year to find a list of the papers that were published as well as links to their PDFs:

World Wide Web Conference (WWW)
International Confence on Web Search and Data Mining (WSDM)
ACM Recommender Systems (RecSys)
International Conference on Information and Knowledge Management (CIKM)
Neural Information Processing Systems (NIPS)
Uncertainty in Artificial Intelligence (UAI)
International Conference on Machine Learning (ICML)
Journal of Machine Learning Research (JMLR)
Journal of Artificial Intelligence Research (JAIR)
Machine Learning Journal (MLJ)

## 2 TEAM

These are 3-person team projects. It is up to you to form a team of people to explore this effort and to produce a co-authored paper and presentation representing the entire team's work. Machine learning is increasingly a multi-person effort at companies (and in academia) so collaboration is a crucial skill. Use the TA and Professor's office hours to discuss your project ideas and to solicit feedback or guidance. Be prepared to discuss broadly what you plan to do and are doing, what results you expect, etc. Please form your team and have a title describing your project ready by April 8 before midnight. Send the title and the team list to the Professor and the TAs by then.

## 3 PRESENTATION

The final presentations should be in powerpoint or pdf files and must be no more than 6 minutes long. Since you are in a group of 3 people, you don't all have to present, one person could be the designated speaker for the group and everyone will get the same grade for the presentation and the writeup. We strongly suggest that your team does not have more than 8 slides total. Time yourself to make sure you do not ramble on for more than your allowed time. We will deduct points if you exceed your allotted time and we will also stop you if you go over your allowed time (that is how things work at real conferences). The presentations are not intended to be a description of your final outcome but a snapshot of where you currently are in the project and your plan, your data-set and any results you currently have. The final report that is due in May is the final outcome. The presentations primarily serve to inform your colleagues of your ideas and get feedback and input from them (as well as from the TAs and the instructor).

For examples of previous years' projects (some links may be broken, please just try to follow the ones that work), take a look at:

http://www1.cs.columbia.edu/~jebara/6772/proj/

# 4 REPORT

After presentations, submit a write-up in a two-column conference paper-style document as a PDF file. Please do not send your work as a Microsoft Office document, LaTex source code, or something more exotic. Include images within your document as figures. Keep your total write-up no longer than 6 pages (two-column). If you go over the page limits, you will be deducted points (that's how conferences enforce limits). To see how to write a good paper and present it, check out this link:

```
http://www.cs.iastate.edu/~honavar/grad-advice.html
```

In particular see Simon Peyton Jones on *How to Write a Good Research Paper*. We recommend using Latex to write up your report:

```
http://www.latex-project.org
```

Submit your report via Courseworks. If unable to, please email it to both the TAs and Instructor. Please tar.gz everything in your current directory and then send it to us. Make sure you send us a write up of your results as a postscript or pdf file containing any figures, tables and equations as well as your Matlab or C code and scripts as separate files.

# 5 DATASETS

In addition to the personalization data-sets we have discussed in class such as the Million Song Dataset, the MovieLens Dataset, etc. you can find additional data-sets here:

```
Million Song Dataset
https://labrosa.ee.columbia.edu/millionsong/

Spotify's RecSys Challenge Dataset
http://www.recsyschallenge.com/2018/

MovieLens Dataset
https://grouplens.org/datasets/movielens/

Criteo Datasets
http://labs.criteo.com/category/dataset/

Kaggle Datasets
https://www.kaggle.com/datasets

Social Curation Network Datasets (Pinterest)
https://nms.kcl.ac.uk/nishanth.sastry/projects/cd-gain/dataset.html

Social Structure of Facebook Networks
http://sociograph.blogspot.com/2011/03/facebook100-data-and-parser-for-it.html

Stanford Large Network Dataset Collection
http://snap.stanford.edu/data/
```

Sentiment Labeled Sentences Dataset
https://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences

General Social Dataset Repository from Data.Gov
https://catalog.data.gov/dataset

BlogFeedback Dataset
https://archive.ics.uci.edu/ml/datasets/BlogFeedback

UCI Machine Learning Repository
http://www1.ics.uci.edu/~mlearn/MLRepository.html

Data for Evaluating Learning
http://www.cs.toronto.edu/~delve/

Another alternative dataset, see the section "Public Available Datasets."
https://github.com/chihming/competitive-recsys