1 Building Recommendation Systems

Goal: To build a simple recommender system

The focus should be on functionality (not on user interface). The system should include enough content and functionality to be "interesting".

Suggestions for applications:

- "short text" recommendations: jokes, quotes, poetry, baby names, recipes, ...
- "local" recommendations, travel: restaurants, cultural events, places in Brno, holiday locations, countries to visit, tourist attractions, ...
- educational recommendations: courses, foreign language vocabulary, learning materials, ...
- product recommendation (specialized for a particular domain): board games, books for children, wines, beers, specific movie genre, ...
- personalized guides: TV program, museum guide, ...

Essential publications will be found in piazza

2 Final Report Guidelines

Due date: The final report is due the second week of January 2023 - More specific plan will be announced later

3 Project Parts: Proposal, Milestones, Literature Review, Final Report

3.1 Project Proposal (Due November 20, 2023)

In this task you need to specify the domain of application of your RS and the dataset you will use.

Your proposal should be a PDF document and jupyter presentation, giving the title Project of the project, the full names of all of your team members, and Proposal a 300-500 word description of what you plan to do.

3.2 Milestone (Due Dec 6, 2023)

The milestone will help you make sure you're on track, and should describe what you've accomplished so far, and very briefly say what else you plan to do. You should write it as if it's an “early draft" of what will turn into your final project. You can write it as if you're writing the first few pages of your final project report, so that you can re-use most of the milestone

text in your final report. Thus, for example, you should not spend two pages explaining what logistic regression is. Note: We will expect your final writeup to be on the same topic as your milestone.

Include a section that describes what each team member worked on and contributed to the project. This is to make sure team members are carrying a fair share of the work for the project.

Your milestone should be at most 3 pages, excluding references. Similar to the proposal, it should include

- Motivation: What problem are you tackling, and what's the setting you're considering?
- Method: What machine learning techniques have you tried and why?
- Preliminary experiments: Describe the experiments that you've run, the outcomes, and any error analysis that you've done. You should have tried at least one baseline.

4 Final Report Guidelines

4.1 Abstract

It should consist of 2-3 paragraphs consisting of the motivation for your report and a high-level explanation of the methodology you used/results obtained.

4.2 Introduction [~1 page]

Explain the problem and why it is important. Discuss your motivation for pursuing this problem. Give some background if necessary. Clearly state what the input and output is. Be very explicit: "The input to our algorithm is an {image, amplitude, patient age, rainfall measurements, grayscale video, etc. }. We then use a { SVM, neural network, linear regression, etc.} to output a predicted {age, stock price, cancer type, music genre, etc.}." Being explicit about this makes it easier for readers.

4.3 Literature Review [~1 page]

You should find existing papers, group them into categories based on their approaches, and discuss their strengths and weaknesses, as well as how they are similar to and differ from your work. In your opinion, which approaches were clever/good? What is the state-of-the-art? You should aim to have at least 5 references in the related work. Include previous attempts by others at your problem, previous technical methods, or previous learning algorithms. Google Scholar is very useful for this: https://scholar (https://scholar). google.com/ (you can click "cite" and it generates MLA, APA, BibTeX, etc.)

4.4 Dataset and Features [$\approx 0.5 - 1$ pages]

Describe your dataset: how many training/validation/test examples do you have? Is there any preprocessing you did? What about normalization or data augmentation? What is the resolution of your images? How is your time-series data discretized? Include a citation on where you obtained your dataset from. Depending on available space, show some examples from your dataset. You should also talk about the features you used. If you extracted features using Fourier transforms, word2vec, histogram of oriented gradients (HOG), PCA, ICA, etc. make sure to talk about it. Try to include examples of your data in the report (e.g. include an image, show a waveform, etc.).

4.5 Methods [$\approx 1 - 1.5$ pages]

Describe your learning algorithms, proposed algorithm(s), or theoretical proof(s). Make sure to include relevant mathematical notation. For example, you can briefly include the SVM optimization objective/formula or say what the softmax function is. It is okay to use formulas from the lecture notes. For each algorithm, give a short description (≈ 1 paragraph) of how it works. Again, we are looking for your understanding of how these machine learning algorithms work. Although the teaching staff probably know the algorithms, future readers may not (reports will be posted on the class website). Additionally, if you are using a niche or cutting-edge algorithm (e.g. long short-term memory, SURF features, or anything else not covered in the class), you may want to explain your algorithm using $1/2$ paragraphs. Note: Theory/algorithms projects may have an appendix showing extended proofs (see Appendix section below).

4.6 Experiments/Results/Discussion [$\approx 1 - 3$ pages]

You should also give details about what (hyper)parameters you chose (e.g. why did you use X learning rate for gradient descent, what was your mini-batch size and why) and how you chose them. Did you do cross-validation, if so, how many folds? Before you list your results, make sure to list and explain what your primary metrics are: accuracy, precision, AUC, etc. Provide equations for the metrics if necessary. For results, you want to have a mixture of tables and plots. If you are solving a classification problem, you should include a confusion matrix or AUC/AUPRC curves. Include performance metrics such as precision, recall, and accuracy. For regression problems, state the average error. You should have both quantitative and qualitative results. To reiterate, you must have both quantitative and qualitative results! This includes unsupervised learning (talk with your project TA on how to quantify unsupervised methods). Include visualizations of results, heatmaps, examples of where your algorithm failed and a discussion of why certain algorithms failed or succeeded. In addition, explain whether you think you have overfit to your training set and what, if anything, you did to mitigate that. Make sure to discuss the figures/tables in your main text throughout this section. Your plots should include legends, axis labels, and have font sizes that are legible when printed.

4.7 Conclusion/Future Work $[\approx 1-2$ paragraphs $]$

Summarize your report and reiterate key points. Which algorithms were the highest performing? Why do you think that some algorithms worked better than others? For future work, if you had more time, more team members, or more computational resources, what would you explore?

4.8 Contributions

The contributions section is not included in the 5 page limit. This section should describe what each team member worked on and contributed to the project.

4.9 References/Bibliography (No page limit)

This section should include citations for: (1) Any papers mentioned in the related work section. (2) Papers describing algorithms that you used which were not covered in class. (3) Code or libraries you downloaded and used. This includes libraries such as scikit-learn, Matlab toolboxes, Tensorflow, etc. Acceptable formats include: MLA, APA, IEEE.

5 Formatting

Your report will be written in Latex. Feel free to adjust the specific sections according to your needs (e.g. combine introduction and related work or separate the experiments from the discussion. You are free to use single-column or two-column layouts. The paper size is standard A4 or 8.5x11 inches. Your font size must be greater than or equal to 10pt. Do not use less than 0.5 inch margins. If you use latex, we highly recommend using a conference/journal template (e.g. NIPS, IEEE, ICML). They generally provide .tex templates. When you submit your final report, it must be in PDF format.

We look forward to reading about your project!

6 Sources of Data Sets

Movies Recommendation:

- *MovieLens* - Movie Recommendation Data Sets http://www.grouplens.org/node/73 (http://www.grouplens.org/node/73)
- *Yahoo!* - Movie, Music, and Images Ratings Data Sets http://webscope.sandbox.yahoo.com/catalog.php?datatype=r (http://webscope.sandbox.yahoo.com/catalog.php?datatype=r)
- *Jester* - Movie Ratings Data Sets (Collaborative Filtering Dataset) http://www.ieor.berkeley.edu/~goldberg/jester-data/

(http://www.ieor.berkeley.edu/~goldberg/jester-data/)

- *Cornell University* - Movie-review data for use in sentiment-analysis experiments http://www.cs.cornell.edu/people/pabo/movie-review-data/ (http://www.cs.cornell.edu/people/pabo/movie-review-data/)

Music Recommendation:

- *Last.fm* - Music Recommendation Data Sets http://www.dtic.upf.edu/~ocelma/MusicRecommendationDataset/index.html (http://www.dtic.upf.edu/~ocelma/MusicRecommendationDataset/index.html)
- *Yahoo!* - Movie, Music, and Images Ratings Data Sets http://webscope.sandbox.yahoo.com/catalog.php?datatype=r (http://webscope.sandbox.yahoo.com/catalog.php?datatype=r)
- *Audioscrobbler* - Music Recommendation Data Sets http://www-etud.iro.umontreal.ca/~bergstrj/audioscrobbler_data.html (http://www-etud.iro.umontreal.ca/~bergstrj/audioscrobbler_data.html)
- *Amazon* - Audio CD recommendations http://131.193.40.52/data/ (http://131.193.40.52/data/)

Books Recommendation:

- *Institut für Informatik, Universität Freiburg* - Book Ratings Data Sets http://www.informatik.uni-freiburg.de/~cziegler/BX/ (http://www.informatik.uni-freiburg.de/~cziegler/BX/)

Food Recommendation:

- *Chicago Entree* - Food Ratings Data Sets http://archive.ics.uci.edu/ml/datasets/Entree+Chicago+Recommendation+Data (http://archive.ics.uci.edu/ml/datasets/Entree+Chicago+Recommendation+Data)

Merchandise Recommendation:

- *Amazon* - Product Recommendation Data Sets http://131.193.40.52/data/ (http://131.193.40.52/data/)

Healthcare Recommendation:

- *Nursing Home* - Provider Ratings Data Set http://data.medicare.gov/dataset/Nursing-Home-Compare-Provider-Ratings/mufm-vy8d (http://data.medicare.gov/dataset/Nursing-Home-Compare-Provider-Ratings/mufm-vy8d)
- *Hospital Ratings* - Survey of Patients Hospital Experiences http://data.medicare.gov/dataset/Survey-of-Patients-Hospital-Experiences-HCAHPS-/rj76-22dk (http://data.medicare.gov/dataset/Survey-of-Patients-Hospital-Experiences-HCAHPS-/rj76-22dk)

Dating Recommendation:

- *www.libimseti.cz (http://www.libimseti.cz)* - Dating website recommendation (collaborative filtering) http://www.occamslab.com/petricek/data/ (http://www.occamslab.com/petricek/data/)

Scholarly Paper Recommendation:

- *National University of Singapore* - Scholarly Paper Recommendation http://www.comp.nus.edu.sg/~sugiyama/SchPaperRecData.html