CSE 291-D Project

The aim of the project for CSE291-D is to give you hands-on experience in latent variable modeling, and to help synthesize what you are learning in the course. This course is about latent variable models for finding hidden structure in complicated data sets. The art of latent variable modeling in practice is an iterative development process, sometimes referred to as *Box's loop*. An analyst proposes a probabilistic model for their data, develops an algorithm to infer the model from data, and evaluates its performance with respect to the goals of the analysis. This process often needs to be repeated, with the model and algorithm being refined until the performance is satisfactory. The three themes of this course are therefore *models*, *inference*, and *evaluation*. While the exact nature of the project is flexible, your project needs to relate to at least one of these themes, and ideally all three. Note that you may have to read ahead to start your project. All readings are listed on the syllabus, on the course webpage. Some important information:

- The project will be done in groups of 2-4.
- There are three milestones for the project:
  - Project proposal Due 4/25/2017
  - Midterm progress report Due 5/16/2017
  - Final report Due 6/15/2017
- Projects will be assessed based on the midterm and final reports.
- The project may overlap with other research you are doing, but not any other class project. Projects that are related to any domain science research you or your teammates may be working on are, in fact, encouraged – I hope this course helps you solve the data science tasks you are already interested in. However, your work for the project needs to be your own, and not that of your labmates or collaborators.

Project proposal

Project proposals are to be sent to me by email (jfoulds@ucsd.edu), and approved by the deadline. Please send me a short summary (< 1 page) of your proposed study, including the goals of the research, why this is important to tackle, the data you plan to use, the type of models/algorithms you plan to use, and the names of the group members. If you need help selecting a project, please arrange to discuss it with me before the deadline.

Midterm progress report

This a one page report, in the format of the NIPS conference (see `https://nips.cc/Conferences/2015/PaperInformation/StyleFiles` for LaTeX style files and other templates). Please summarize the progress made, results obtained, and methods tried. Include a discussion of any challenges faced, and plans to resolve them.

Final report

Think of the final report as a near-final draft of an academic research paper. The report must be no more than 10 pages, including references, in NIPS format. It must include a discussion of the relevant work in the literature, and how the work goes beyond this. It should include the technical details of the approach, including equations, derivations, graphical model diagrams, and so on. The report will be evaluated on clarity in addition to technical merit. Late final reports will not be accepted.

Project ideas / guidelines

Many types of projects are possible, subject to my approval. Once again, the project should relate to the themes of models, inference, and evaluation. I expect there to be at least a little novelty in the proposed approach, and I expect any algorithms to be well principled, e.g. designed to correctly simulate from a model or to optimize a reasonable objective function. Note that research is challenging and uncertain, and it's ok if the project does not yield positive results. Although this is not at all required, a very typical project might have the following structure:

- The goal is to perform some data mining task or scientific analysis with extrinsic value, e.g. with industrial significance, or motivated from computational biology, sociology, etc.
- The project proposes a probabilistic model for the corresponding data set. The model uses latent variables, such as hidden cluster assignments, to capture important hidden properties of the data.
- The proposed model starts with a standard probabilistic model such as an HMM, mixture model, social network model, topic model, GLM, etc, and builds upon it or specializes it in order to improve performance relative to the task of interest.
- The model is evaluated with respect to some criterion, and compared to simpler approaches, possibly including non-probabilistic ones, and ideally with competing methods from the literature.

Some possible ways to improve on an existing model include leveraging additional data types (e.g. improving a social network model by jointly modeling or conditioning on associated text, audio, etc), incorporating the time-varying/dynamic aspects of the problem into the model, modeling additional latent structure with extra latent variables, and handling missing data. You could also develop an improved inference algorithm.

One possible approach is to use a probabilistic programming language to make model development easier, e.g. WinBugs, Stan, Markov Logic Networks, Probabilistic Soft Logic, FACTORIE, or Infer.NET. However, as this makes things much easier I will expect more in the way of results, such as comparisons of multiple variants of a model. Finally, here are a few possible sources of data sets which might help get you started:

- Yahoo Webscope datasets: `http://webscope.sandbox.yahoo.com/`
- Yelp Dataset Challenge: `https://www.yelp.com/dataset_challenge`
- Stanford Network Analysis Project (SNAP): `http://snap.stanford.edu/index.html`
- GroupLens recommendation system data sets: `http://grouplens.org/datasets/movielens/`
- ImageNet: `http://www.image-net.org/`
- UCI Machine Learning Repository: `http://archive.ics.uci.edu/ml/`
- The Stanford MOOCPosts Data Set: `http://datastage.stanford.edu/StanfordMoocPosts/`
- Microsoft Learning to Rank data sets: `http://research.microsoft.com/en-us/um/beijing/projects/letor/`
- Julian McAuley's webpage: `http://cseweb.ucsd.edu/~jmcauley/`
- Tagged.com social spam data set: `https://obj.umiacs.umd.edu/tagged_social_spam/index.html`