Thesis defense

MACHINE LEARNING DEPARTMENT

Diversity-Promoting and Large-Scale Machine Learning for Healthcare

Abstract:

In healthcare, a tsunami of medical data has emerged, including electronic health records, images, literature, etc. These data can be heterogeneous and noisy, which renders clinical decision-making time-consuming, error-prone and suboptimal. In this thesis, we develop machine learning (ML) models and systems for distilling high-value patterns from unstructured clinical data and making informed and real-time medical predictions and recommendations, to aid physicians in improving the efficiency of workflow and quality of patient care. When developing these models, we encounter several challenges: (1) How to better capture infrequent clinical patterns, such as rare subtypes of diseases; (2) How to make the models generalize well on unseen patients? (3) How to promote the interpretability of the decisions? (4) How to improve the timeliness of decision-making without sacrificing its quality? (5) How to efficiently discover massive clinical patterns from large-scale data?

To address challenges (1-4), we systematically study diversity-promoting learning, which encourages the components in ML models (1) to diversely spread out to give infrequent patterns a broader coverage, (2) to be imposed with structured constraints for better generalization performance, (3) to be mutually complementary for more compact representation of information, and (4) to be less redundant for better interpretation. The study is performed in the context of both frequentist statistics and Bayesian statistics. In the former, we develop diversity-promoting regularizers that are empirically effective, theoretically analyzable and computationally efficient. In the latter, we develop Bayesian priors that effectively entail an inductive bias of "diversity" among a finite or infinite number of components and facilitate the development of efficient posterior inference algorithms. To address challenge (5), we study large-scale learning. Specifically, we design efficient distributed ML systems by exploiting a system-algorithm co-design approach. Inspired by a sufficient factor property of many ML models, we design a peer-to-peer system -- Orpheus -- that significantly reduces communication and fault tolerance costs.

We apply the proposed diversity-promoting learning (DPL) techniques and distributed ML systems to address several critical issues in healthcare, including discharge medication prediction, automatic ICD code filling, automatic generalization of medical-imaging reports, similar-patient retrieval, hierarchical multi-label tagging of medical images, and large-scale medical-topic discovery. Evaluations on various clinical datasets demonstrate the effectiveness of the DPL methods and efficiency of the Orpheus system.

Speaker:

Pengtao Xie

Thesis Committee:

Eric P. Xing, Chair
Pradeep Ravikumar
Ruslan Salakhutdinov
Ryan Adams (Princeton)
David Sontag (MIT)

May 21, 2018
11:00am
8102 GHC

Link to draft document:
http://www.cs.cmu.edu/~pengtaox/thesis_draft_pengtaoxie.pdf