Proposal for Gaining Access to Datasets in the IMEDS Lab

Mijung Park, Ph.D., Machine Learning Research Associate, University College London
Email: mijung@gatsby.ucl.ac.uk, Cellphone: +44 77 0742 9497

- **Research Objectives:** My research goal is to develop methods for improving the performance of personalised predictions of disease risk using statistical and machine learning tools. The main challenge in disease risk prediction is two-fold: (1) the size of the observational data, typically time-series of health histories, is huge; and (2) there is a myriad of different factors that could possibly be highly non-linearly interacting and considered together as risk factors for a disease. In this light, my aim is to develop tractable and interpretable methods, which will help medical practitioners efficiently handle large-scale data as well as identify important risk factors for each disease.
- **Proposed Approaches:** My project involves sparse coupled latent factor models combined with the logistic regression likelihood. The sparse coupled latent factor models enable us to understand the shared core factors associated with each disease across a certain feature like gender or age. Under the proposed models, posterior inference is analytically intractable and sampling-based inference methods will be difficult to employ due to the large size of data. What I propose is to derive an accurate, fast, and efficient variational inference algorithm by employing sigmoid belief nets to accurately capture the non-standard type (e.g., multi-modal) of posterior distribution over the latent factors and estimating all the parameters simultaneously using recently developed stochastic back-propagation algorithms.
- **Impact:** Existing work often employs black-box types of algorithms to brutally improve the prediction accuracy, without considering interpretability of results. Unlike these black-box methods, it is my hope that my proposed methods will have a high impact in improving disease risk prediction as well as giving medical practitioners insights in each disease's associated core risk factors. The type of datasets I am particularly interested in using is the dataset for stroke prediction task [1]. All drugs and conditions will be considered as features that could affect the stroke occurrence within one year of observation time following the diagnosis of the atrial fibrillation. My proposed method is expected to extract the shared key risk factors of stroke across gender or age among the high dimensional features.
- **Experience:** Mijung Park, Ph.D., is the only person requiring access to the data. Her postdoc adviser and collaborators will be part of the papers she will write using the output generated from the data, but will not require individual access to the data.
- **Timeline:** 36 months from March 2015

References

1. Letham B, Rudin C, McCormick TH, Madigan D (2014) Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. Department of Statistics

Technical Report tr608, University of Washington.