University of California, Irvine
Statistics Seminar
Obsidian Security Lecture

Bayesian Categorical Matrix Factorization via Double Feature Allocation

Peter Mueller
University of Texas, Austin

Thursday, November 21, 2019
4 p.m., 6011 Bren Hall
(Bldg. #314 on campus map)

We propose a categorical matrix factorization method to infer latent diseases from electronic health records data. A latent disease is defined as an unknown cause that induces a set of common symptoms for a group of patients. The proposed approach is based on a novel double feature allocation model which simultaneously allocates features to the rows and the columns of a categorical matrix. Using a Bayesian approach, available prior information on known diseases greatly improves identifiability of latent diseases. This includes known diagnoses for patients and known association of diseases with symptoms. For application to large data sets, as they naturally arise in electronic health records, we develop a divide-and-conquer Monte Carlo algorithm, which allows inference for the proposed double feature allocation model, and a wide range of related Bayesian nonparametric mixture models and random subsets. We validate the proposed approach by simulation studies including mis-specified models and comparison with sparse latent factor models. In an application to Chinese electronic health records (EHR) data, we find results that agree with related clinical and medical knowledge.

For directions/parking information, please visit https://uci.edu/visit/maps.php and http://www.ics.uci.edu/about/visit/index.php