Working Group on Statistical Learning Seminar

Title: Uncovering User Roles in Forum Data Using a Variational Bayes Mixed Membership Approximation

Speaker: Arthur White

Date: Fri 13th May 2011 at 1:00PM

Location: Statistics Seminar Room- L550 Library building

Abstract: Discussion forums are a central part of Web 2.0 infrastructure, and are a prominent form of social information exchange. As they have increased in traffic, the need to understand and characterize user behavior has emerged. For instance, forum hosts may wish to provide incentives to posters who regularly instigate discussions, or investigate whether forum discussions are being dominated by highly uncommunicative, non-social members.

We attempt to uncover user behavior from data recorded from Boards.ie over a twelve-month period. The Boards.ie data set is unusual in that it contains the internal reply structure of discussions, allowing us to discern not only the volume of user posts, but in addition the volume of replies, and initiated conversations. Each thread in the Boards.ie dataset can be represented as a weighted, directed graph, with vertices representing users in the forum, a directed edge indicating a reply to a post, and edge weight indicating the number of posts between users. Collapsing these graphs by aggregating edge weights across all threads in each forum then allows us to profile users within the forum.

Decomposing forums using conventional mixture model methods reveals many groups with several overlapping features. We attempt to describe user behaviour in a more concise manner by using a mixed membership model. In a mixture model framework there are assumed to be some number G of underlying groups , with the probability of group membership modelled by a vector τ. For the mixed membership model we instead assume that an underlying number K of extreme profiles generate the data, with each individual i in the dataset having a membership vector τ_i. Each user's behaviour is then treated as some combination of the profiles in the model.

Because direct inference of the model is intractable, a variational approximation is applied. Technical difficulties and methods to identify the correct number of extreme profiles are examined, and some early results from forum data will be discussed.