SySc 431/531: DATA MINING WITH INFORMATION THEORY (DMIT)

MODEL almost ANY DATA!

- Do you have data in spread-sheet (rectangular) format, i.e., where variables are columns and cases* are rows?
- Might the variables have non-linear relations or complex interaction effects?
- Do you want to do exploratory modeling to discover these relations?

For example, if you have a set of input variables that might predict an output variable, where you don’t know what the predictive relations are, do you want to discover these relations? If you answer ‘yes’ to the above questions, take this data mining course.

* Cases could be members of a population where variables are attributes of these members; or cases could be time points (or locations in space) at which the variables are measured.

DMIT is a project-based course that offers you an opportunity to use information theoretic methods to analyze data. These methods are implemented in a software package named OCCAM, developed at PSU, that will be the main analytical tool used in the course. The theory underlying these methods is taught in SySc 551/651 Discrete Multivariate Modeling (DMM), but this course (DMIT) is stand-alone, and does not have DMM as a prerequisite. Only the theory needed to understand the inputs and outputs of OCCAM will be presented, but OCCAM will be treated as a black box, so the algorithms that it implements will not be discussed. The point is to make it possible for you to do exploratory modeling on data of interest to you without having to master the underlying theory first. To understand this theory, you can take DMM later, but this is not required.

Information about OCCAM and papers about these methods and their use – as well as access to OCCAM and a User’s Manual – can be found at the instructor’s DMM web site: http://www.pdx.edu/sysc/research-discrete-multivariate-modeling.

Required readings from this web site are: (i) the OCCAM manual, (ii) the tutorial (“Overview”) paper, (iii) one or more research papers to be used as guidance for the research and selected by the Instructor based on the student’s project. An optional text on these methods is Krippendorff, Klaus (K). Information Theory: Structural Models for Qualitative Data. Series: Quantitative Applications in the Social Sciences, Paper # 62, Sage Publications, Beverly Hills, California, 1986. (ISBN 0-8039-2132-2, paperback)

Prerequisites for 431 section: Upper division standing and completion of one of the SYSC3xxU cluster courses (or permission of instructor).
Recommended preparation: (a) Basic probability and statistics or machine learning (e.g., Math 105 or Stat 243 or equivalent), (b) access to data that you know something about and want to analyze. The instructor will be able to provide data to students who do not have their own data to analyze, but bringing your own data is preferable.

Course requirements & Grades: Graduate students will submit a substantial research paper at the end of the course (80%) & give a class presentation of their results (20%). The research paper should be of the type and quality that could be submitted for presentation at a scientific meeting: introduction (including presentation of the subject problem area), data description, methods (strategy of & steps in the RA analysis), results, discussion. Undergraduates will submit a shorter research paper (100%), focusing on data, methods, and results. Where applicable, graduate projects will use variable-based models without loops, variable-based models with loops, and state-based models, and will utilize both the search and fit actions of OCCAM. Undergraduate projects need use only variable-based models without loops and only the search action of OCCAM.

Course Outline

Class 1,2 Introduction to DMIT: slide presentation on data mining with Reconstructability Analysis (RA) and with OCCAM

Class 3,4 Presentation on the OCCAM software

Class 5 Presentation of a prototype RA research paper by Instructor

Class 6 Class members describe their projects

Class 7-10 Students work in computer lab on their projects, with assistance from Instructor; also extensive interaction via email.

Class 11 Mid-quarter short research reports

Class 11-14 Continued work on research projects

Class 15 Draft papers due for Instructor comments and guidance

Class 15-18 Project work continues

Class 19,20 Class presentations

Class 20 Final projects reports due