Graduate Course

Statistics 601: Multivariate and categorical data analysis

Instructor: Long Nguyen
Department of Statistics, Univ. of Michigan
Winter 2013
Time: 2:30–4:00 PM MW, B760 East Hall

Course description. This is an advanced introduction to the analysis of multivariate and categorical data. Topics include: (1) dimension reduction techniques, including principal component analysis, multidimensional scaling and extensions; (2) classification, starting with a conceptual framework developed from cost functions, Bayes classifiers, and issues of over-fitting and generalization, and continuing with a discussion of specific classification methods; (3) clustering methods, including hierarchical methods, partitioning methods, K-means, and model-based clustering based on finite mixtures and infinite mixture models; (4) categorical data analysis, starting with estimation and testing with generalized linear models and continuing with a more general treatment of probablistic graphical models.

Prerequisites. The prerequisites are previous coursework in linear algebra, multivariate calculus, and basic probability and statistics. Previous experience in numerical analysis and optimization would be helpful but is not required. Familiarity with R, Matlab, Splus or a related matrix-oriented programming language will be necessary.

Structure/Evaluation. The course will meet twice a week and will follow a regular lecture format. There will be bi-weekly homework assignments, due one week after being passed out. There will be a midterm and a final exam. There will be a team project. Homeworks account for 40% of the grade, while the midterm 20%, the final 20%, the project 20%.

The project consists of an oral presentation of assigned research papers and a summary write-up.

Course homepage. Please use Ctools for all announcements, homeworks, project information and data sets.

Textbook. There is no official textbook for this course. Lecture notes will be given by the instructor. In addition, relevant research papers, book chapters and existing notes will be provided via Ctools. The following references will be useful:

- Perlman's notes (UW) on multivariate gaussian and wishart. Available via Ctools.
- Michael Jordan's selected book chapters on graphical models.
- T. Anderson's book "Multivariate analysis".
- Hastie, Friedman, Tibshirani's book on statistical learning. Available online.
- C. Bishop's book "Pattern recognition and machine learning", Springer 2006.

Tentative outline.

1. Dimensionality reduction methods.
    - introduction: data, problems, approaches
    - multivariate random vectors, multivariate Gaussian
    - model-based linear dimensionality reduction (factor analysis, probabilistic pca)
    - data-driven linear dimensionality reduction (principal component analysis,...)
    - nonlinear dimensionality reduction: MDS and manifold learning
2. Classification methods.
    - model-based classification: linear classification, logistic regression
    - data-driven classification: (pattern recognition/machine learning approaches)
    - the svm, the adaboost, ensemble methods
3. Clustering methods.
    - data-driven clustering: k-means, spectral clustering
    - model-based clustering: mixture of multivariate gaussians
    - EM algorithm for mixture models
    - Sampling methods for mixture models
    - Dirichlet process mixtures and multivariate extensions
4. Graphical/hierarchical modeling for multivariate and categorical data
    - exponential families, generalized linear models
    - graphical models (chains, trees, general graphs): Bayesian networks, Markov random fields, Gaussian fields, hidden Markov models
    - hierarchical models (latent variable models), e.g., Latent Dirichlet Allocation
    - convex duality and variational inference/ sum-product message passing inference algorithms
    - important sampling, MCMC and slice sampling methods for GM/HM