

**LEARNING MIXTURE MODELS
COURSEWARE FOR FINITE MIXTURE
MODELS OF MULTIVARIATE BERNOULLI
DISTRIBUTIONS**

JAAKKO HOLLMÉN AND TAPANI RAIKO

DEPARTMENT OF INFORMATION AND COMPUTER SCIENCE

HELSINKI UNIVERSITY OF TECHNOLOGY, FINLAND

JAAKKO.HOLLMEN@TKK.FI

HTTP://WWW.CIS.HUT.FI/JHOLLMEN/BERNOULLIMIX

MAY 7, 2008

AIMS OF MACHINE LEARNING TEACHING

- Programming computers to learn from data

STANDARD COURSE FORMAT

- 5 ECTS credit points
- 12 lectures (2h / week)
- Pen-and-paper exercise sessions (2h / week)
- *term project: solve a machine learning problem by programming*

ALTERNATIVE COURSE FORMAT

- 5 ECTS credit points
- 12 lectures (2h / week)
- Pen-and-paper exercise sessions (2h / week)
- *term project: solve a machine learning problem by using a ready-made software package*

THIS TALK

- Term project with BernoulliMix software package
- Experiences on our course: Machine learning:
Advanced probabilistic methods, spring 2008
(part of the Macadamia master's program)

INITIAL IDEA

- Frees the student from the burden of just getting the programs to work
- No auxiliary routines (input/output etc.) needed
- Guided execution

WHY MIXTURE MODELS?

- Material learned in the beginning phases of the course
- Possible to finish the term project within the same semester
- Simplest of the more complex models (very simple Bayesian network with a latent variable)
- Relevance to clustering problems

WHY 0-1 DATA?

- Finite mixture model of multivariate Bernoulli distributions
- Personal research interest
- Shared interest in machine learning and data mining communities
- Simplicity and power combined

THE MODEL

$$P(\mathbf{x}) = \sum_{j=1}^J \pi_j P(\mathbf{x} \mid \boldsymbol{\theta}_j) = \sum_{j=1}^J \pi_j \prod_{i=1}^d \theta_{ji}^{x_i} (1 - \theta_{ji})^{1-x_i}$$

- Finite mixture model of multivariate Bernoulli distributions
- Data \mathbf{x}
- Mixture coefficients π_j
- Parameters of the component distributions θ_{ji}

FIVE PROGRAMS

- command-line tools (Unix / Linux), implemented in C
- `bmix_init`: Initialize a mixture model
- `bmix_like`: Calculate the likelihood of data
- `bmix_train`: Train a model using the EM algorithm
- `bmix_sample`: Sample data from a mixture model
- `bmix_cluster`: Cluster data with the mixture model

DOCUMENTATION

- Written in Texinfo format, possible to generate on-line documentation and printed documentation
- General structure of documentation:
 - Use of programs: command-line options
 - Examples of machine learning tasks
 - Exercises follow the examples

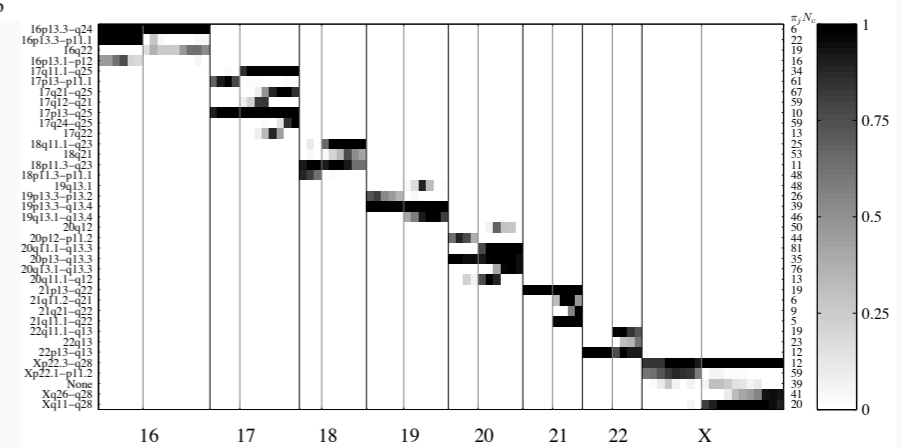
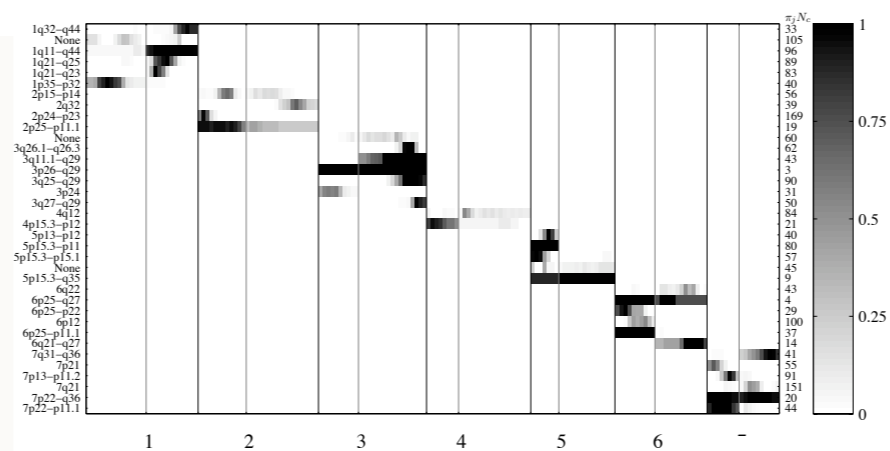
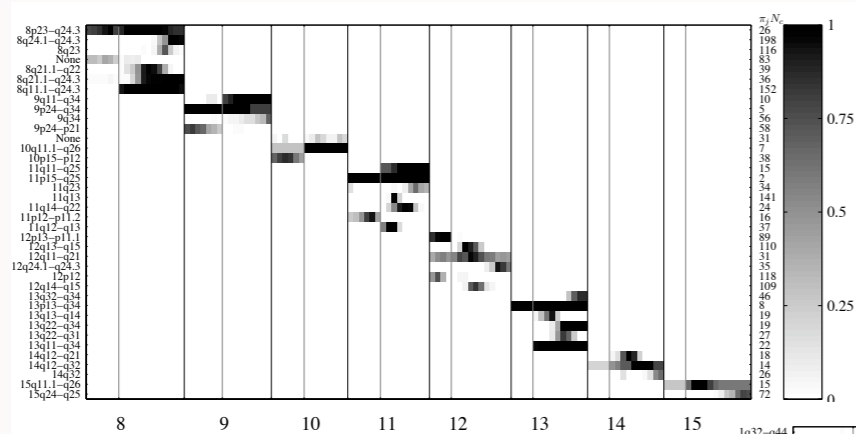
TASKS IN THE EXERCISE

- Tasks are of varying difficulty from repeating the examples (simplest) to a cross-validation based model selection problem (most complex)
- “Explain why...”, “What kind of consequences...”, “Compare the models and explain your findings”
- Concentrate on the machine learning content!

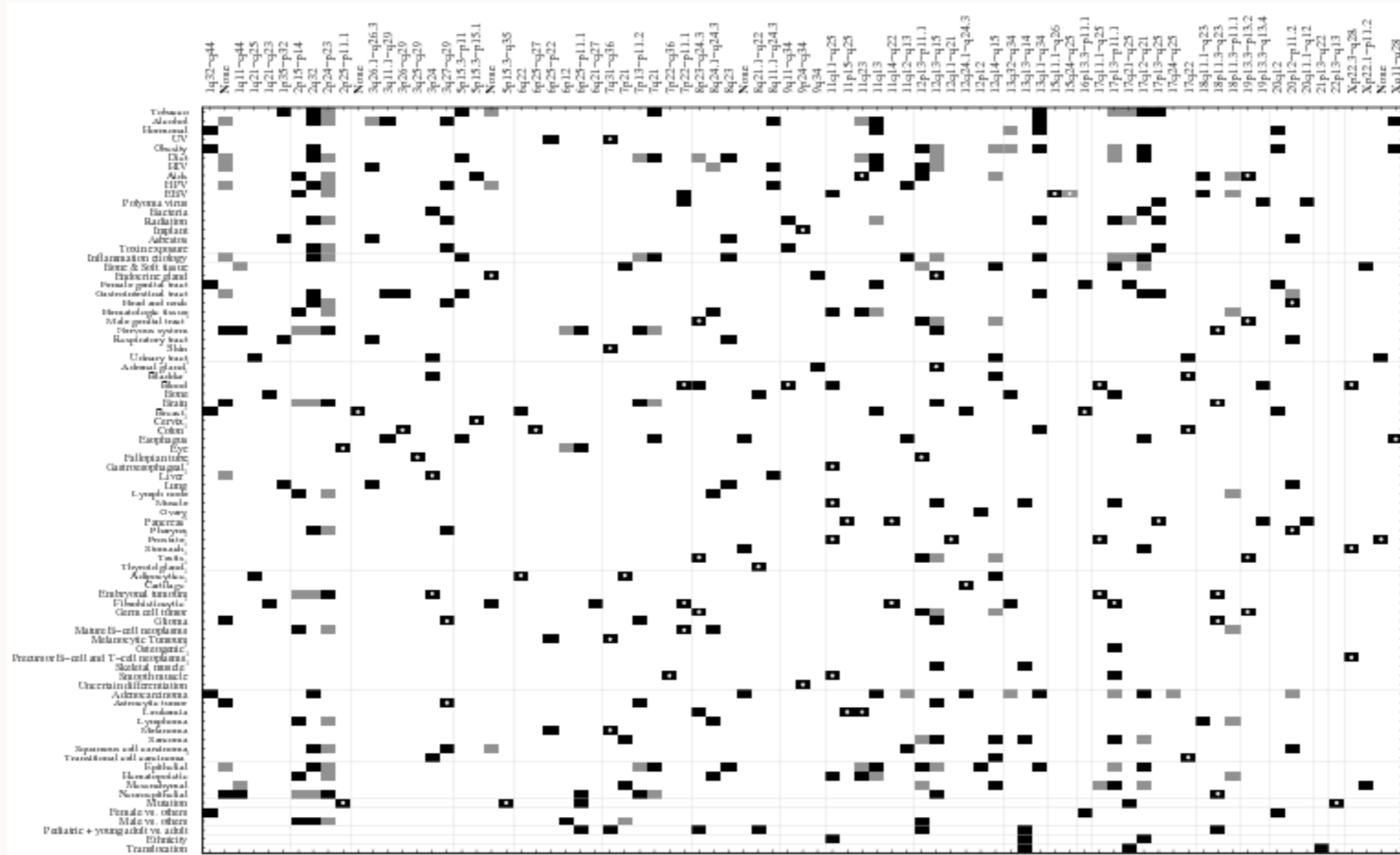
0-1 DATA

- DNA copy number amplification data in gastric cancer (patients $n=35$, genetic markers $d=6$)
- DNA copy number amplification data of chromosome 17 in database of cancers (patients $n=4400$, chromosomal regions $d=12$)
- Some artificial data sets with known structure

DNA COPY NUMBER AMPLIFICATION MODELS



RISK ASSOCIATIONS



PRACTICAL ISSUES DURING THE TERM PROJECT

- Interoperation with other software such as Matlab
- Shell script programming

WHY NOT R, WEKA?

- Here: focused effort on mixture modeling
- R, Weka: overhead in learning to use, not widely known
- Interfacing these systems possible (Matlab, R)
- Aiming at open source teaching material

QUESTIONNAIRE IN RETROSPECT

- I prefer vs. I do not prefer a term project involving programming: 50% vs. 50%
- I am confident in programming machine learning content in following programming languages: Matlab (95%), R and Python (< 25 %), others (< 10 %), C (5 %)
- I am confident in compiling a C program (95 %)
- Shell script programming variably known

PRELIMINARY THOUGHTS

- Feedback on the term project both positive and constructive
- Program code and instructions sufficiently matured
- Of those returned (n=3), all the exercises were completed!
- May 31st is the deadline, the returned term projects (n=40) will help in estimating the difficulty

WORK IN PROGRESS: IMPROVEMENTS

- Support for compiling on a wide number of hosts
- Addition of shell script examples
- Usage improvements
- Source code modifications
- API reference
- Encourage feedback => Get feedback

SUMMARY

- Experiences on our course “Machine learning: Advanced Probabilistic Methods” (Hollmén, Raiko)
- Term project with very little programming
- BernoulliMix: an open source teaching material with programs, documentation, exercises
- <http://www.cis.hut.fi/jhollmen/BernoulliMix>