Project in Probabilistic Models

Course code: 582637
Credit points: 2 cr
Teacher: Dr Antti Honkela (& Prof Petri Myllymäki)
Contact email: antti.honkela@cs.helsinki.fi
Office hours: Please make an appointment by email
Prerequisites: 582636 Probabilistic Models
Graphical model structure learning

- Why do we want to learn the structure
Graphical model structure learning

Why do we want to learn the structure
  ▶ Scientific discovery
  ▶ More efficient density modelling
Graphical model structure learning

- Why do we want to learn the structure
  - Scientific discovery
  - More efficient density modelling
- Potential challenges
  - Uncertainty about the correct structure (weak links, limited data, ...)
  - Learning of correlation instead of causation, equivalent structures
The project task

- Infer the structure and corresponding distributions of a discrete graphical model
- Data: 2500 samples of 21 variables (all discrete with 3 values)
- The network connectivity has been extracted from a real network, but the probability model is synthetic
- Required outputs:
  - Ranked list of all possible arcs in the model
  - Normalised probability distribution over a set of 1500 test vectors
The data

The training and test data sets are available in Moodle

A B C D E F G H I J K L M N O P Q R S T U
1 2 2 3 1 2 2 1 2 3 1 3 3 2 2 2 3 2 1 1 1
2 3 3 3 2 3 2 1 3 3 1 3 3 2 2 3 1 2 1 2 1
1 1 2 3 1 2 2 3 3 1 2 3 3 2 2 1 1 3 1 1 3
3 3 3 1 3 1 2 3 2 1 2 3 3 2 3 3 1 1 3 2 2
3 3 1 1 3 2 1 3 1 1 1 3 3 2 3 3 3 3 1 1 2
3 1 3 1 3 3 2 2 3 1 3 2 3 2 3 3 1 1 1 2 2
3 1 3 1 3 3 2 2 3 1 3 2 3 2 3 3 1 1 1 2 2
...
Ranked arc list

- Return a list of all potential 420 arcs in the model in ranked order with ones you believe to be active in the beginning

- Example:

  A B
  A C
  B D
  C D
  A D
  B A
  B C
  C A
  C B
  D A
  D B
  D C
Test set probabilities

- Return a normalised list of probabilities (must sum to 1) for the test vectors (in order)

- Example:
  3.941543e-01
  1.637665e-02
  3.199843e-01
  1.524029e-02
  7.287055e-04
  1.392376e-03
  3.395783e-09
  2.521234e-01
Evaluation of the predictions

▶ Ranked arc list
  ▶ Evaluated using the area under ROC curve
▶ Predicted probabilities
  ▶ Evaluated using the Kullback–Leibler divergence between the true distribution and the prediction

\[
D_{KL} = \sum_i p_i \log \frac{p_i}{q_i}
\]
Area under ROC example

Here: $AUC=0.8438$
Scoreboard

- All the scores will be published *anonymously* on a scoreboard together with brief descriptions of the methods used.
- You will receive an email notification with your own score.
- *Positions on the scoreboard will not be used as a criterion for course grading!*
Return instructions

- There are four deadlines during the course (always on Tuesdays)
  - 5 April
  - 12 April
  - 19 April
  - 26 April (final DL)
- You must return all your submissions to the course Moodle area
- The return consists of
  - Predictions as specified above
  - 1 line public summary of the methods you have used for the score board
  - 1/2 page diary of your progress
Final return instructions

- The final return (26 April) consists of
  - Your final predictions
  - 1 line summary of the methods
  - A written report of the project containing introduction, methods, results and discussion
    - The weekly diary entries will be included in the report
  - All source code used
Using existing software

- Using existing software in your project is permitted if the software is freely available for academic use
  - Use of commercial packages is not allowed
- Using own code is rewarded in grading but not required
- Remember to give proper credit to packages you use!
Return logistics: Moodle

- All returns must be made to Moodle
  https://moodle.helsinki.fi
- You must log in using your University (non-CS) account
- Please register to the course “Project in Probabilistic Models, spring 2011”
  - The course registration key is “structure”
- For more instructions, please see “Student guide” on Moodle home page
Schedule of the meetings for the rest of the course

- Course meetings on Thursdays at 16-18
- Mandatory attendance on feedback sessions starting 7 April

24 March  Q+A session
31 March   **No meeting**
7 April    First feedback session
14 April   Second feedback session
21 April   **Easter holiday, no meeting**
28 April   Final session
Grading

- The grading will be based on your returned reports and presentations given during course sessions.
- The following will positively influence your grade:
  - Effort put to the problem, innovativeness.
  - Good presentations of your work during the course.
  - Being able to improve your performance during the course and learn from previous results.
  - Use of own software.
- Score board positions will not be used in grading!
Final warning

- In case you are tempted: the test data do not come from the same distribution as the training data. Using them in training the model is not recommended!
Questions?

▶ Any questions?