Course code: 582637
Credit points: 2-3 cr
Teacher: Dr Antti Honkela
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

- Scientific discovery
- More efficient density modelling
- ...

- Potential challenges
- Uncertainty about the correct structure (weak links, limited data, ...)
- Equivalent structures
- Learning of correlation instead of causation
- ...
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
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, ...)
  - Equivalent structures
  - Learning of correlation instead of causation
  - ...

The project task

- Infer the structure and corresponding distributions of a discrete graphical model
- Data: 2500 samples of 26 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
Training and test 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 | V | W | X | Y | Z |
| 3 | 1 | 1 | 1 | 3 | 2 | 3 | 1 | 3 | 1 | 3 | 1 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 3 | 2 | 1 |
| 3 | 1 | 2 | 3 | 1 | 3 | 3 | 1 | 1 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 1 | 1 | 1 | 2 | 1 | 3 | 3 | 3 | 2 | 2 | 2 | 1 | 1 |
| 3 | 3 | 1 | 1 | 3 | 2 | 3 | 1 | 3 | 1 | 3 | 2 | 3 | 2 | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 2 | 3 | 2 | 1 |
| 3 | 1 | 1 | 2 | 3 | 2 | 2 | 1 | 2 | 1 | 2 | 1 | 3 | 1 | 3 | 2 | 3 | 1 | 2 | 1 | 3 | 3 | 3 | 3 | 1 | 1 | 1 |
| 2 | 1 | 1 | 3 | 1 | 3 | 3 | 1 | 3 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 1 | 2 | 2 | 3 | 3 | 1 | 2 | 1 | 1 |
| 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 1 | 3 | 2 | 1 | 3 | 3 | 1 | 2 | 2 | 1 | 2 | 1 | 1 | 3 | 1 | 2 | 1 | 1 | 1 |
After the second round returns, there will be an additional “development” data set for testing your probability predictions. This set contains new data vectors and their corresponding probabilities. More information at the second feedback session.
Ranked arc list

- Return a list of all potential 650 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
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!*
In groups of approximately 3 people

1. Suggest (at least) 3 methods you could apply to the task.
2. Which tools do you think are best suited for the task?
Return instructions

- There are four deadlines during the course (always on Monday mornings)
  - 25 March
  - 8 April
  - 15 April
  - 22 April (final predictions 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
Due to popular request from last year, there are two extra submission opportunities

- 2 April
- 19 April

These are optional: use if you wish more feedback

All 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
Final return instructions

- The final report return (23 April) consists of
  - 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
- Additionally: an approximately 5 min presentation at the final meeting
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
- Using significant amount of own code you can get 3 cr instead of 2 cr
- 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 2013”
  - The course registration key is “learning”
- For more instructions, please see “Student guide” on Moodle home page
Schedule of the meetings for the rest of the course

- Course meetings on Tuesdays at 10-12
- Mandatory attendance on feedback sessions starting 26 March

19 March  Q+A session
26 March  First feedback session
2 April    Easter holiday, no meeting
9 April    Second feedback session
16 April   Third feedback session
23 April   Final session with presentations
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?