Introduction to this Course

Meelis Kull

meelis.kull@ut.ee

Liivi 2-326
Meelis Kull

• Associate Professor (dotesnt) in Data Mining

• 1998-2011 studied informatics at Tartu:
  – BSc:
    • Supervisor: prof. Mati Tombak
    • Thesis topic: theoretical informatics
  – MSc, PhD:
    • Supervisor: prof. Jaak Vilo
    • Thesis topics: data mining and bioinformatics

• 2013-2017 Postdoctoral researcher at Bristol, UK
  – Project 1: context change and model reuse in machine learning - [http://reframe-d2k.org](http://reframe-d2k.org)
  – Project 2: analysis of data from smart homes - [http://www.irc-sphere.ac.uk](http://www.irc-sphere.ac.uk)
Main scientific interests:

- Machine learning, artificial intelligence, data mining, data science, applications in smart homes, health, smart electric grids, etc.
- Please contact me if looking for thesis topics or supervision

Examples of studied scientific questions:

- Why are predictions from machine learning methods (or AI) often over-confident?
- What can we do about this?
- How to combine data from many sources to improve predictions?
Organisational information about this course
Information about the course

• Course homepage:
  – This is the primary source of all information

• Main material:
This is how the course works

• Week 1:
  – Introductory lecture on ensemble methods

• Week 2:
  – Introductory practice session on ensemble methods

• Weeks 3-14
  – Seminars:
    • Presentation and discussion (60min)
    • Test (15min)
    • Discussion about test (15min)
This is how a seminar works

- Everyone reads the material at home
- 2 students prepare together a presentation
  - Aim at 30min presentation and 30min discussion
  - Both must present at least 10min
- 2 other students prepare together the test
  - Aim at 15min answering time
  - Hand-written or electronic – your choice
  - Can/cannot use materials – your choice
  - Passing threshold is 50% of points over all tests
  - Below 50% - might still pass after talking to me
- Attendance threshold: 75% (or talk to me)
Requirements to pass the course

- Attendance: 75%
- Tests: 50%
- 1 presentation
- 1 test prepared and graded
- If attending <75% or tests <50%:
  - Need to convince me that you have done enough work and I will still give you a pass
- Don’t worry too much when setting a test for others – they will most likely still pass
Amount of work

• 3 ECTS = 3*26=78 hours of intensive work
  – This is an expected average over all students
  – Less background and skills means more hours
• 2 hours per each of 13 seminars (26h)
• 3 hours for continuing after practice (3h)
• 3 hours per each of 11 material readings (33h)
• 12 hours preparing the presentation (12h)
• 4 hours preparing and grading a test (4h)
• Total: 78 hours
How we agree presentation times

• I will do a doodle poll asking you to choose at least 4 weeks when you could present
  – Preferrably a lot more than 4 weeks
• I assume willingness to present implies willingness to make a test that week
• I will prepare a schedule at least 2 weeks ahead
  – Whenever possible I try to give you freedom in choosing the topic and team-mate
Asking for help

• If you are preparing a presentation or a test and have problems understanding:
  – Ask from your team-mate
  – Ask from the other team (presentation/test)
  – Ask from me
# How material is split

## Course plan

<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Seminar</th>
<th>Topic</th>
<th>Sections</th>
<th>Pages</th>
<th>Presenters</th>
<th>Testers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>07.09.2017</td>
<td>no seminar</td>
<td></td>
<td>-</td>
<td>-</td>
<td>Meelis Kull</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>14.09.2017</td>
<td>Lecture</td>
<td>Introduction to Ensemble Methods</td>
<td>-</td>
<td>-</td>
<td>Meelis Kull</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>21.09.2017</td>
<td>Practice session</td>
<td>Introduction to Ensemble Methods</td>
<td>-</td>
<td>-</td>
<td>Meelis Kull</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>28.09.2017</td>
<td>Seminar 1</td>
<td>Combination methods 1</td>
<td>4.1-4.3</td>
<td>67-83</td>
<td>TBD</td>
<td>TBD</td>
</tr>
<tr>
<td>5</td>
<td>05.10.2017</td>
<td>Seminar 2</td>
<td>Combination methods 2</td>
<td>4.4-4.7</td>
<td>83-97</td>
<td>TBD</td>
<td>TBD</td>
</tr>
</tbody>
</table>

From this point on the distribution of topics and pages can change:

<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Seminar</th>
<th>Topic</th>
<th>Sections</th>
<th>Pages</th>
<th>Presenters</th>
<th>Testers</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>12.10.2017</td>
<td>Seminar 3</td>
<td>Diversity 1</td>
<td>5.1-5.3</td>
<td>99-111</td>
<td>TBD</td>
<td>TBD</td>
</tr>
<tr>
<td>7</td>
<td>19.10.2017</td>
<td>Seminar 4</td>
<td>Diversity 2</td>
<td>5.4</td>
<td>111-118</td>
<td>TBD</td>
<td>TBD</td>
</tr>
<tr>
<td>8</td>
<td>26.10.2017</td>
<td>Seminar 5</td>
<td>Ensemble pruning</td>
<td>6</td>
<td>119-133</td>
<td>TBD</td>
<td>TBD</td>
</tr>
<tr>
<td>9</td>
<td>02.11.2017</td>
<td>Seminar 6</td>
<td>Bagging</td>
<td>3</td>
<td>47-66</td>
<td>TBD</td>
<td>TBD</td>
</tr>
<tr>
<td>10</td>
<td>09.11.2017</td>
<td>Seminar 7</td>
<td>Boosting</td>
<td>2</td>
<td>23-46</td>
<td>TBD</td>
<td>TBD</td>
</tr>
<tr>
<td>11</td>
<td>16.11.2017</td>
<td>Seminar 8</td>
<td>Cost-sensitive ensembles</td>
<td>TBD</td>
<td>TBD</td>
<td>TBD</td>
<td>TBD</td>
</tr>
<tr>
<td>12</td>
<td>23.11.2017</td>
<td>Seminar 9</td>
<td>Ensembles and class distribution shift</td>
<td>TBD</td>
<td>TBD</td>
<td>TBD</td>
<td>TBD</td>
</tr>
<tr>
<td>13</td>
<td>30.11.2017</td>
<td>Seminar 10</td>
<td>TBD</td>
<td>TBD</td>
<td>TBD</td>
<td>TBD</td>
<td>TBD</td>
</tr>
<tr>
<td>14</td>
<td>07.12.2017</td>
<td>Seminar 11</td>
<td>TBD</td>
<td>TBD</td>
<td>TBD</td>
<td>TBD</td>
<td>TBD</td>
</tr>
</tbody>
</table>