Welcome to CPSC 340

Instructor: Raymond Ng
Email: rng@cs.ubc.ca
Office: CICS/CS Rm 303
Office Hour: Friday 1:30 pm - 2:45pm
(and before and after midterm)
TAs’ office hours: TBA


Course Material

  - Key chapters: 18, 19, 20 and a bit of 21 and 22
  - Recommend that you own the book if you plan to
take 322 and 422
- References:
  - (mathematical) “The Elements of Statistical Learning,” Hastie, Tibshirani
  and Friedman (2nd edition) [Free download for UBC students]

Who cares about taking 340??

- BIG Data is everywhere
  - Health informatics, Finance, interactive videos, EOS project
  - Next lecture: what Big Data means and applications in personalized medicine
- Key learning objectives: data exploration, data
  quality, unsupervised learning, supervised learning
- Jobs and demand for qualified personnel

<table>
<thead>
<tr>
<th>Planned Schedule</th>
<th>Mon</th>
<th>Wed</th>
<th>Fri</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sep 1</td>
<td>X</td>
<td>outline</td>
<td>intro</td>
</tr>
<tr>
<td>Sep 8</td>
<td>clustering</td>
<td>clustering</td>
<td>clustering</td>
</tr>
<tr>
<td>Sep 15</td>
<td>tutorOw</td>
<td>X</td>
<td>PCA</td>
</tr>
<tr>
<td>Sep 22</td>
<td>outliers</td>
<td>decTrees</td>
<td>decTrees</td>
</tr>
<tr>
<td>Sep 29</td>
<td>decTrees</td>
<td>ROC*</td>
<td>CV</td>
</tr>
<tr>
<td>Oct 6</td>
<td>regress</td>
<td>ridge</td>
<td>lasso</td>
</tr>
<tr>
<td>Oct 13</td>
<td>X</td>
<td>elasticN*</td>
<td>logistic</td>
</tr>
<tr>
<td>Oct 20</td>
<td>neuNet</td>
<td>neuNet</td>
<td>MT</td>
</tr>
<tr>
<td>Oct 27</td>
<td>MTQ/A</td>
<td>navBayes</td>
<td>navBayes</td>
</tr>
<tr>
<td>Nov 3</td>
<td>nonPara</td>
<td>SVM*</td>
<td>SVM</td>
</tr>
<tr>
<td>Nov 10</td>
<td>ensemble</td>
<td>ranForest</td>
<td>EM</td>
</tr>
<tr>
<td>Nov 17</td>
<td>EM</td>
<td>HMM*</td>
<td>NLPApn</td>
</tr>
<tr>
<td>Nov 24</td>
<td>NLPApn</td>
<td>CRF</td>
<td>review</td>
</tr>
</tbody>
</table>
Course Evaluation

<table>
<thead>
<tr>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Final exam</td>
</tr>
<tr>
<td>In-class midterm (Oct 24)</td>
</tr>
<tr>
<td>4 assignments</td>
</tr>
<tr>
<td>8 tutorials</td>
</tr>
</tbody>
</table>

Must pass the final exam and above 50% in overall grade.

Tutorials

- The course is designed to give you lots of hands-on practice (i.e., Rstudio)
  - Tutorials start on Sep 15 till Nov 17 (no tutorial on Thanksgiving Monday)
  - 8 graded tutorials (1% each)
  - Hand in what you have done at the end of the tutorial
  - Each tutorial to be done in pairs (one of you must have a laptop)

Assignments

- You will be writing R scripts. If you know Java and/or python, the R language is easy to learn
- You will be introduced to R scripts starting from the first tutorial
- Beyond that, you will need to pick it up on your own from [http://www.r-project.org/](http://www.r-project.org/)
- Assignments to be done on your own, all due on Wednesdays

At the end of the term

- You will have learned most of the basic unsupervised and supervised learning methods
- You will understand the pros and cons of various methods
- You will be able to conduct your own data analytics tasks
Bigger Picture Wrt CS Courses

- An agent
  - Takes as input: prior knowledge, past experiences, observations, and goals/values
  - Reasons and makes inferences
  - Determines as output: actions that affect the world
  - these topics are covered in 322 and 422 (AI)
- Then what happens in the world change past experiences and observations, and got fed-back to the agent
  - Learning from data are covered in 340

Policies(1): Plagiarism

- Plagiarism will be heavily monitored
- Rule of thumb:
  - Discussion is encouraged; learning from each other is important
  - But you must write your own work
- When in doubt:
  - Ask me
  - Full and voluntary disclosure of the collaboration

Policies(2): Exams

- I rarely re-weighed midterms; better study hard for it
- Level of difficulties: my goal is to put the median grade around B (~75%)
  - Depending on the performance of the midterm, I set the level of difficulty of the final exam to hit the target

Policies(3): Required Reading

- The prescribed chapters of the textbook, including material not covered in lectures
  - Parts of the material is intended to be self-taught
  - I expect you to spend time reading
- Course notes and tutorial notes posted on Connect
- You are also encouraged to read the bulletin board on a regular basis
Policies(4): Attending classes

- You may want to print out the course notes and read it ahead of time (self-reading again)
- I would like to go fast on the course notes to spend more time on examples and questions from you
- You don’t need to attend classes but examples discussed in the lectures may not appear elsewhere

Policies(5): Math background

- This course is designed to cover various basic aspects of data mining and machine learning
- It is intended to get you to a point that you can create your own DM task, knowing your choices and how the various methods work
- Some of the methods require strong math background (e.g., probabilistic models, proofs)
- I assume you have the background (hence, the pre-reqs) and give a math formulation relevant to the topic