CMSC 691
Probabilistic and Statistical Models of Learning
Spring 2020

https://www.csee.umbc.edu/~ferraro/teaching/691-s20/
Welcome!

Administrivia
Frank Ferraro

ITE 358

ferraro@umbc.edu

Monday: 3:45-4:30
Tuesday: 11-11:30
by appointment

Natural language processing:
  Semantics

Vision & language processing

Generative & neural modeling

Learning with low-to-no supervision
Machine Learning: A Terminology Buffet

- **Classification**
- **Regression**
- **Clustering**

**Task:** what kind of problem are you solving?

**Data:** amount of human input/number of labeled examples

**Approach:** how any data are being used

- Fully-supervised
- Semi-supervised
- Un-supervised

- Probabilistic
- Spectral
- Generative
- Memory-based
- Neural
- Exemplar
- Conditional
- ...
Machine Learning: A Terminology Buffet

Classification
Regression
Clustering

Fully-supervised
Semi-supervised
Un-supervised

the task: what kind of problem are you solving?
the data: amount of human input/number of labeled examples
the approach: how any data are being used

Probabilistic
Generative
Conditional
Spectral
Memory-based
Exemplar
...
Course Goals

1. Be introduced to advanced statistical estimation and modeling in ML;
2. Understand how these approaches can be used to develop semi-supervised and unsupervised ML algorithms;
3. Have experience implementing a number of ML programs;
4. Read and analyze research papers;
5. Practice your (written) communication skills.
Incoming Expectations

CMSC 678 is a prereq

- Experience (academic or industry) with ML, probs/stats, and/or any fields that depend on it
  - natural language processing, computer vision, speech recognition, and some applications of data science, econometrics, or bioinformatics

Expected topical familiarity:

- probabilistic notions of
  - independence of two or more random variables
  - conditional independence
  - Marginalization
  - Bayes rule
- computing gradients (multivariate derivatives) and basic matrix manipulation and linear algebra (transpose, matrix multiplication, matrix inversion, determinants, the concept of eigenvectors/eigenvalues)
- formalizing the machine learning task as loss minimization
- implementation, from scratch (not using an existing implementation from, e.g., sklearn), of at least one machine learning algorithm
- reading machine learning papers from top-tier venues

You don’t need extensive familiarity with all of the above
What do you want out of this course?

1. Be introduced to advanced statistical estimation and modeling in ML;
2. Understand how these approaches can be used to develop semi-supervised and unsupervised ML algorithms;
3. have experience implementing a number of ML programs;
4. read and analyze research papers;
5. practice your (written) communication skills.
Course Model

This is *not* a survey course.

We will go deep into the topics.
This is not a survey course.

We will go deep into the topics.

Use of Class Time
• Initially: More lecture-style
• Later: Discussion-based, paper reading
Outline

Welcome!

Administrivia
Syllabus

• https://www.csee.umbc.edu/~ferraro/teaching/691-s20/materials/syllabus.pdf

• Extended detail about:
  – Grading computation
  – Academic honesty policy
  – Accommodations, diversity & inclusion, and Title IX
## Grading

<table>
<thead>
<tr>
<th>Component</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assignments</td>
<td>25%</td>
</tr>
<tr>
<td>Writing Log/Journal Entries</td>
<td>25%</td>
</tr>
<tr>
<td>Course Project</td>
<td>50%</td>
</tr>
</tbody>
</table>
# Grading

<table>
<thead>
<tr>
<th>Component</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assignments</td>
<td>25%</td>
</tr>
<tr>
<td>Writing Log/Journal Entries</td>
<td>25%</td>
</tr>
<tr>
<td>Course Project</td>
<td>50%</td>
</tr>
</tbody>
</table>

Each component is \( \max(\text{micro-average}, \text{macro-average}) \)

(exception: you must complete all milestones of the project)
## Grading

<table>
<thead>
<tr>
<th>Component</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Assignments</td>
<td>25%</td>
</tr>
<tr>
<td>Writing Log/Journal Entries</td>
<td>25%</td>
</tr>
<tr>
<td>Course Project</td>
<td>50%</td>
</tr>
</tbody>
</table>

Each component is \( \max(\text{micro-average}, \text{macro-average}) \)

(exception: you must complete all milestones of the project)

See syllabus for example computation
## Final Grades

<table>
<thead>
<tr>
<th>If you get ≥</th>
<th>You get at least a/an</th>
</tr>
</thead>
<tbody>
<tr>
<td>90</td>
<td>A-</td>
</tr>
<tr>
<td>80</td>
<td>B-</td>
</tr>
<tr>
<td>70</td>
<td>C-</td>
</tr>
<tr>
<td>65</td>
<td>D</td>
</tr>
<tr>
<td>0</td>
<td>F</td>
</tr>
</tbody>
</table>
Online Presence

Course webpage:
https://csee.umbc.edu/~ferraro/teaching/691-s20
- Find schedule, readings, course materials
- Summary of notices/due dates

Piazza:
https://piazza.com/umbc/spring2020/cmusc6918
- Announcements of assignments & project
- Post questions (and answers 😊)
- If you’re not enrolled, email me
Running the Assignments

A "standard" x86-64 Linux machine

A passable amount of memory (2GB-4GB)

Modern but not necessarily cutting edge software

If in doubt, ask first
Running the Project

An x86-64 Linux machine

Memory and hardware constraints lifted (somewhat)

If in doubt, ask first
Programming Languages for Assignments

Use the tools you feel comfortable with

Python+numpy, C, C++, Java, Matlab, ...: OK (straight Python may not cut it)

Libraries: Generally OK, as long as you don’t use their implementation of what you need to implement

Math accelerators (blas, numpy, etc.): OK

If in doubt, ask first
Programming Languages for the Project

Use the tools you feel comfortable with

Python+numpy, C, C++, Java, Matlab, ...: OK (straight Python may not cut it)

Libraries: **Use what you want**

Math accelerators (blas, numpy, etc.): OK
Submitting Your Work

https://www.csee.umbc.edu/~ferraro/teaching/691-s20/submit

- Online Google form: must be logged in via UMBC
- Multiple submissions are allowed
- Latest submission will be graded
Late Policy

Everyone has a budget of 10 late days
Late Policy

Everyone has a budget of 10 *late days*

If you have them left: assignments turned in after the deadline will be graded and recorded, no questions asked.
Late Policy

Everyone has a budget of 10 *late days*

If you have them left: assignments turned in after the deadline will be graded and recorded, no questions asked

If you don’t have any left: still turn assignments in. They could count in your favor in borderline cases
Late Policy

Everyone has a budget of 10 *late days*

Use them as needed throughout the course
  They’re meant for personal reasons and emergencies

Do not procrastinate
Late Policy

Everyone has a budget of 10 late days

Contact me privately if an extended absence will occur

You must know how many you’ve used
Main Resource: ITILA

“Information Theory, Inference and Learning Algorithms”
MacKay

http://www.inference.org.uk/mackay/itprnn/ps/

Full book:

Official
Optional Resource: CIML

“A Course in Machine Learning”, v0.99
Hal Daumé III

http://ciml.info/

Full book:
http://ciml.info/dl/v0_99/ciml-v0_99-all.pdf

Unofficial: Recommended
Optional Advanced Resource: ESL

“Elements of Statistical Learning”
Hastie, Tibshirani, Friedman


Full book:

Unofficial: Recommended
Optional Advanced Resource: UML

“Understanding Machine Learning: From Theory to Algorithms”
Shalev-Shwartz, Ben-David

http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/

Full book:

Unofficial: Recommended
Resources #5... ∞

Peer-reviewed articles (journals, conferences & workshops)

ICML

CVF

JMLR