Course Staff

Instructors:
- Nina Balcan  http://www.cs.cmu.edu/~ninamf
- Avrim Blum  http://www.cs.cmu.edu/~avrim

TAs:
- Sarah Allen  http://www.cs.cmu.edu/~srallen
- Nika Haghtalab  http://www.cs.cmu.edu/~nhaghtal
Lectures in general

On the board

Occasionally, will use slides
Machine Learning

Image Classification

Document Categorization

Speech Recognition  Protein Classification

Branch Prediction  Fraud Detection  Spam Detection

Playing Games  Computational Advertising
Machine Learning is Changing the World

“Machine learning is the hot new thing”
(John Hennessy, President, Stanford)

“A breakthrough in machine learning would be worth ten Microsofts” (Bill Gates, Microsoft)

“Web rankings today are mostly a matter of machine learning” (Prabhakar Raghavan, VP Engineering at Google)
The COOLEST TOPIC IN SCIENCE

• “A breakthrough in machine learning would be worth ten Microsofts” (Bill Gates, Chairman, Microsoft)

• “Machine learning is the next Internet” (Tony Tether, Director, DARPA)

• Machine learning is the hot new thing” (John Hennessy, President, Stanford)

• “Web rankings today are mostly a matter of machine learning” (Prabhakar Raghavan, Dir. Research, Yahoo)

• “Machine learning is going to result in a real revolution” (Greg Papadopoulos, CTO, Sun)

• “Machine learning is today's discontinuity” (Jerry Yang, CEO, Yahoo)
This course: foundations of Machine Learning and Data Science
Goals of Machine Learning Theory

Develop and analyze models to understand:

- what kinds of tasks we can hope to learn, and from what kind of data
- what types of guarantees might we hope to achieve
- prove guarantees for practically successful algos (when will they succeed, how long will they take?)
- develop new algos that provably meet desired criteria (potentially within new learning paradigms)

Interesting connections to other areas including:

- Algorithms
- Probability & Statistics
- Complexity Theory
- Optimization
- Game Theory
- Information Theory
Example: Supervised Classification

Decide which emails are spam and which are important.

Goal: use emails seen so far to produce good prediction rule for future data.
Example: Supervised Classification

Represent each message by features. (e.g., keywords, spelling, etc.)

<table>
<thead>
<tr>
<th>“money”</th>
<th>“pills”</th>
<th>“Mr.”</th>
<th>bad spelling</th>
<th>known-sender</th>
<th>spam?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>

Example label

Reasonable RULES:

Predict SPAM if unknown AND (money OR pills)

Predict SPAM if 2money + 3pills - 5 known > 0

Linearly separable
Two Main Aspects of Supervised Learning

Algorithm Design. How to optimize?
Automatically generate rules that do well on observed data.

Confidence Bounds, Generalization Guarantees, Sample Complexity
Confidence for rule effectiveness on future data.

Well understood for passive supervised learning.
Using Unlabeled Data and Interaction for Learning

Application Areas

Search/Information Retrieval

Computer Vision Spam Detection

Computational Biology

Robotics Medical Diagnosis
Massive Amounts of Raw Data

Only a tiny fraction can be annotated by human experts.

Protein sequences  Billions of webpages  Images
Semi-Supervised Learning

Learning Algorithm

Unlabeled data

face

not face

Labeled data

Classifier

raw data

Expert Labeler
Active Learning
Other Protocols for Supervised Learning

• **Semi-Supervised Learning**
  Using cheap unlabeled data in addition to labeled data.

• **Active Learning**
  The algorithm interactively asks for labels of informative examples.

Theoretical understanding entirely lacking 10 years ago. Lots of progress recently. We will cover some of these.
Distributed Learning

Many ML problems today involve massive amounts of data distributed across multiple locations.

Often would like low error hypothesis wrt the overall distrib.
Distributed Learning

Data distributed across multiple locations.

E.g., medical data
Distributed Learning

Data distributed across multiple locations.

E.g., scientific data
Distributed Learning

- Data distributed across multiple locations.
- Each has a piece of the overall data pie.
- To learn over the combined $D$, must communicate.

Important question: how much communication? Plus, privacy & incentives.
The World is Changing Machine Learning

New approaches. E.g.,

- Semi-supervised learning
- Interactive learning
- Distributed learning
- Multi-task/transfer learning
- Deep Learning
- Never ending learning

Many competing resources & constraints. E.g.,

- Computational efficiency (noise tolerant algos)
- Human labeling effort
- Statistical efficiency
- Communication
- Privacy/Incentives
Structure of the Class

Basic Learning Paradigm: Passive Supervised Learning

• Basic models: PAC, SLT.
• Simple \textit{algos} and \textit{hardness} results for supervised learning.
• Standard Sample Complexity Results (VC dimension)
• Modern Sample Complexity Results
  • Rademacher Complexity; localization
• Weak-learning vs. Strong-learning
• Classic, state of the art algorithms: AdaBoost and SVM (kernel based methods).
  • Margin analysis of Boosting and SVM
Structure of the Class

Other Learning Paradigms

• Incorporating Unlabeled Data in the Learning Process.

• Incorporating Interaction in the Learning Process:
  • Active Learning
  • More general types of Interaction

• Distributed Learning.

• Transfer learning/Multi-task learning/Life-long learning.

• Deep Learning.

• Foundations and algorithms for constraints/externalities. E.g., privacy, limited memory, and communication.
Structure of the Class

Other Topics.

- Methods for summarizing and making sense of massive datasets including:
  - unsupervised learning.
  - spectral, combinatorial techniques.
  - streaming algorithms.

- Online Learning, Optimization, and Game Theory
  - connections to Boosting
Admin

- Course web page:
  http://www.cs.cmu.edu/~ninamf/courses/806/10-806-index.html

Two grading schemes:

1) Project Oriented.
   - Project [60%]
   - Take-home final [10%]
   - Hwks + grading [30%]

2) Homework Oriented.
   - Hwk +grading [60%]
   - Take-home final [10%]
   - Project [30%]
1) Project Oriented.

- Project [60%]
  - explore a theoretical or empirical question;
  - write-up --- ideally aim for a conference submission!
  - Small groups OK.

- Take-home final [10%]

- Hwks + grading [30%]
2) Homework Oriented.
   - Hwk +grading [60%]
   - Take-home final [10%]
   - Project [30%]
   - read a couple of papers and explain the idea.

Course web page:
http://www.cs.cmu.edu/~ninamf/courses/806/10-806-index.html
Lectures in general

On the board

Occasionally, will use slides