Machine Learning: CIS 472/572

Introduction

Instructor: Daniel Lowd
Based on slides by Vibhav Gogate, Pedro Domingos, and others.

Logistics

- **Instructor**: Daniel Lowd ([lowd@cs.uoregon.edu](mailto:lowd@cs.uoregon.edu))
  - Office hours: Mon 2-3pm, Wed 1-2pm (Deschutes 262)
- **GE Teaching**: Nisansa de Silva ([nisansa@cs.uoregon.edu](mailto:nisansa@cs.uoregon.edu))
  - Office hours: Tue 1-3pm, Thur 11am-1pm (Deschutes 232)
- **Web**: [http://www.cs.uoregon.edu/Classes/20W/cis472/](http://www.cs.uoregon.edu/Classes/20W/cis472/)
- **Discussion Board**: Piazza (link on web page)
- **Written and Programming Assignments**: Gradescope
  (Good intro, focuses on machine learning concepts before math. Free online. Not finished.)
  - Many supplementary sources – see webpage!
Evaluation

• Homework (30%)
  – 3 written assignments, 3 programming assignments
  – Python template code will be provided
  – Undergraduates may solve programming assignments in pairs
  – All other work must be done independently.

• One midterm (40%)
  – 2/3rds of the way through

• One project (30%)
  – Apply machine learning to a real problem of your choice (Recommended: Participate in a contest on Kaggle.com.)
  – Groups allowed
  – Written report
  – Presentations during final exam time

Kaggle

<table>
<thead>
<tr>
<th>Competition Name</th>
<th>+ Reward</th>
<th># Teams</th>
<th>Deadline</th>
</tr>
</thead>
<tbody>
<tr>
<td>National Data Science Bowl</td>
<td>$175,000</td>
<td>312</td>
<td>2 months</td>
</tr>
<tr>
<td>Driver Telematics Analysis</td>
<td>$30,000</td>
<td>453</td>
<td>2 months</td>
</tr>
<tr>
<td>Helping Santa’s Helpers</td>
<td>$20,000</td>
<td>439</td>
<td>2.1 days</td>
</tr>
<tr>
<td>Click-Through Rate Prediction</td>
<td>$15,000</td>
<td>1203</td>
<td>35 days</td>
</tr>
<tr>
<td>BCI Challenge @ NER 2015</td>
<td>$1,000</td>
<td>161</td>
<td>50 days</td>
</tr>
<tr>
<td>Sentiment Analysis on Movie Reviews</td>
<td>Knowledge</td>
<td>675</td>
<td>54 days</td>
</tr>
<tr>
<td>Finding Elo</td>
<td>Knowledge</td>
<td>100</td>
<td>2 months</td>
</tr>
</tbody>
</table>
Goal: Use course attributes to predict the student's course rating (+2 = loved it, -2 = hated it).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>+2</td>
<td>y</td>
<td>y</td>
<td>n</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>+2</td>
<td>y</td>
<td>y</td>
<td>n</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>+2</td>
<td>n</td>
<td>y</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>+2</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>+2</td>
<td>n</td>
<td>y</td>
<td>y</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>+1</td>
<td>y</td>
<td>y</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>+1</td>
<td>y</td>
<td>y</td>
<td>n</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>+1</td>
<td>n</td>
<td>y</td>
<td>n</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>0</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>0</td>
<td>n</td>
<td>y</td>
<td>n</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>0</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>-1</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>-1</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>-1</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>-2</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>-2</td>
<td>n</td>
<td>y</td>
<td>y</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>-2</td>
<td>y</td>
<td>n</td>
<td>y</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>-2</td>
<td>y</td>
<td>n</td>
<td>y</td>
<td>n</td>
<td>y</td>
</tr>
</tbody>
</table>

So What Is Machine Learning?

- Automating automation
- Getting computers to program themselves
- Writing software is the bottleneck
- Let the data do the work instead!
Another view on machine learning

- Machine learning = automated science (sort of)
- Goal is to go from raw data to useful knowledge
- An ML algorithm finds a theory to fit the data and background knowledge as well as possible.
- A theory is good if it has good predictive accuracy.
Machine Learning: Applications

Examples of what you will study in class in action!

Spam Filtering

Classify as “Spam” or “Not Spam”
Collaborative Filtering
Machine learning has grown in leaps and bounds

- The main approach for
  - Computer vision
  - Speech recognition
  - Natural language processing
  - Computational biology
  - Sensor networks
  - Robotics
  - Web
  - ...and many more each year...
Types of Learning

• **Supervised (inductive) learning**
  – Training data includes desired outputs

• **Unsupervised learning**
  – Training data does not include desired outputs
  – Find hidden structure in data

• **Semi-supervised learning**
  – Training data includes a few desired outputs

• **Reinforcement learning**
  – the learner interacts with the world via “actions” and tries to find an optimal policy of behavior with respect to “rewards” it receives from the environment

Related Fields

• Fields that use machine learning:
  – Artificial intelligence
  – Computer vision
  – Natural language processing
  – Computational biology
  – Robotics
  – ...many more...

• Fields with similar goals to machine learning:
  – Statistics
  – Data mining
  – Data science
  – Psychology (developmental, cognitive)

• Fields used by machine learning:
  – Information theory
  – Numerical optimization
  – Computational complexity
Types of Supervised Learning Problems

- **Classification**: predict a discrete value from a predefined set of values
- **Regression**: predict a continuous/real value
- **Structured prediction**: predict a complex output, such as a sequence or tree

What We’ll Cover

Foundations for the theory and application of machine learning.

- ML Models
- ML Algorithms
- ML Theory
- ML Best Practices

This is not a tutorial on how to use machine learning – you must also understand why.
What We’ll Cover

• **Angles:** Models, Algorithms, Theory, Best Practices
• **Supervised learning:** Decision tree induction, Instance-based learning, Neural networks, Support vector machines, Model ensembles, Learning theory, etc.
• **General machine learning concepts and techniques:** Feature selection, cross-validation, algorithm evaluation, debugging your machine learning system.

**Not covering:**
• Clustering and unsupervised learning (453/553)
• Reinforcement learning (471/571, 410/510 Multiagent Systems)
• Probabilistic graphical models (471/571, 473/573)
• Structured prediction (e.g., machine translation, image segmentation, multi-label classification) (410/510 Natural language processing)

What We’ll Cover: Comparison

**Computer Science**
• **Core concepts:** Variables, conditionals, loops, functions, etc.
• **Key algorithms:** Mergesort, linked lists, binary search trees, breadth-first search, etc.
• **Process:** Debugging, software engineering, etc.

**Machine Learning**
• **Core concepts:** Classification, overfitting, underfitting, training set, etc.
• **Key algorithms:** Decision trees, nearest neighbor, linear models, etc.
• **Process:** Designing and debugging ML systems
Ethics in Machine Learning

• Machine learning has the potential to do great good and great harm.
• We must think carefully about the effects of our systems, both when they work and when they fail.
• I will bring up ethical issues periodically in class.
• I am not an expert in ethics, but will do my best to point out important questions and lead productive discussions.