Machine Learning

Vasant Honavar
Artificial Intelligence Research Laboratory
Informatics Graduate Program
Computer Science and Engineering Graduate Program
Bioinformatics and Genomics Graduate Program
Neuroscience Graduate Program
Data Sciences Undergraduate Program
Center for Big Data Analytics and Discovery Informatics
Huck Institutes of the Life Sciences
Institute for Cyberscience
Clinical and Translational Sciences Institute
Northeast Big Data Hub
Pennsylvania State University

vhonavar@ist.psu.edu
http://faculty.ist.psu.edu/vhonavar
http://ailab.ist.psu.edu
Introductions

• Instructor
  – Dr. Vasant Honavar
  – Professor, IST & CSE, Data Science, BG, Neuroscience
  – Director, Artificial Intelligence Research Lab
  – Director, Center for Big Data Analytics and Discovery Informatics
  – E335 Westgate Bldg
  – vhonavar@ist.psu.edu
  – http://faculty.ist.psu.edu/vhonavar

• Teaching Assistant
  – Aria Khademi
  – PhD Student, Informatics
  – akhademi@ist.psu.edu

• Students?
Research Overview

- **Machine learning**: Statistical, information theoretic, linguistic and structural approaches to machine learning; learning predictive relationships from sequential, graph-structured, multi-relational, multimodal, partially specified, partially labeled, distributed data, linked data
- **Causal Inference**: Causal inference from disparate experimental and observational studies, causal inference from relational data, causal inference from temporal data
- **Knowledge Representation and Inference**: Logical, probabilistic, and decision-theoretic knowledge representation and inference; federated knowledge bases; selective information sharing; federated services; representing and reasoning about qualitative preferences
- **Applied Informatics**
  - **Bioinformatics**: Macromolecular structure and function, analysis, inference, modeling, and prediction of macromolecular (protein-protein, protein-RNA, and protein-DNA) interaction networks, interfaces, and complexes; immune networks; microbiomes etc.
  - **Health Informatics**: Predictive and causal modeling of health outcomes from patient (health records, genomics, socio-economic, environmental) data
  - **Brain Informatics**: Modeling and analysis of structure and dynamics of brain networks
- **Algorithmic Discovery**
  - Algorithmic abstractions of scientific domains
  - Representations of scientific artifacts (experiments, data, models, assumptions, hypotheses, theories ...)
  - Infrastructure for computationally mediated collaborative science
Computing, Artificial Intelligence, and Data Sciences

- Computation is the best formalism we have for describing how information is encoded, stored, communicated and used by natural as well as synthetic systems.

- Computation plays in many sciences a role that is analogous to what calculus played in transforming physics from a descriptive science (pre Newton) into a predictive science (post Newton)
  - Computation: Cognitive sciences / AI : : Calculus : Physics
  - Computation: Life sciences : : Calculus : Physics
  - Computation: Social sciences : : Calculus : Physics

- Algorithms as theories: We understand a phenomenon when we have an algorithm that models it at the desired level of detail.

- Computing offers an exploratory apparatus for science: To the extent that science is about acquiring, organizing, integrating, analyzing, and reasoning with information, computing, science of information processing, provides exploratory apparatus for science.
Transformative role of computation

• Computation offers the best formalism we have for understanding how information is acquired, processed, and used by
  – Computers
  – Brains
  – Genomes
  – Organizations
  – Societies

• Computation : cognitive science :: calculus : physics
• Computation: biology :: calculus : physics
• Computation: social science :: calculus : physics
• Algorithms as theories
  – We will have a theory of intelligence or learning when we have computer programs (information processing models) that display intelligence or learning
About the course

• What is Machine Learning?
• Why are you taking this course?
• What can you expect to learn in the course?
• What are the prerequisites?
• What is expected of you as a student?
• How will you be evaluated?
What is this course about?

• Learning predictive models from data
  – Why should machines learn?
  – What does it mean for a machine to learn?
  – What can machines learn?
    • Classification
    • Function approximation
    • Clustering
  – How can machines learn?
  – How can we evaluate learned models?
  – How can machines learn better?
Machine learning is a subfield of artificial intelligence and data sciences

AI is about

- Study of computational models of intelligence
- Falsifiable hypotheses about intelligent behavior
- Construction of intelligent artifacts
- Mechanization of tasks requiring intelligence
- Exploring the design space of intelligent systems
What is Data Science?

Computer Science using Big Data

Machine Learning

Math & Statistics

Data Science

Dangerous Software

Traditional Research

Subject Matter Expertise
Machine learning is essential for extracting knowledge from big data.
Why should machines learn?

Machine learning is about replacing humans writing code for specific tasks with humans supplying data and objectives for training machines to perform those tasks.

Machine Learning is most useful when:

- The structure of the task is not well understood but representative data can be obtained.
  - Humans are very good at distinguishing apples from oranges yet terrible at specifying how to do so.
  - An expert physician excels at clinical diagnosis but is often unable to explain how he arrives at a diagnosis.

- Task parameters vary across users.
  - SPAM detection
  - Recommendation of products.
Machine Learning is...

• Programming computers to optimize a performance criterion using data or past experience

• About methods that can automatically detect patterns in data, and use the uncovered patterns to make predictions

• Concerned with the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions
Machine Learning is...

About (computationally) predicting the future based on the past
Machine Learning is about (computationally) predicting the future based on the past.
Why should machines learn?

Practical applications

- Diagnosing diseases from symptoms
- Detecting SPAM
- Determining credit-worthiness
- Recommending products, movies, web pages..
- Targeting advertisements
- Predicting stock prices
- Detecting malware
- Driving cars
- Predicting molecular function from sequence
- Predicting health risks
- Detecting fraud
- Precision farming
- Language translation
Why should machines learn?

Practical

• Explicitly specifying the knowledge needed for specific tasks is hard, and often infeasible
• If we can get machines to acquire the knowledge needed for particular tasks from observations (data), interactions (experiments), we can
  • Dramatically reduce the cost of developing AI systems
  • Automate aspects of scientific discovery
  • ...
Why should machines learn? – Science of learning

Information processing models can provide useful insights into

• How humans and animals learn
• Information requirements of learning tasks
• The precise conditions under which learning is possible
• Inherent difficulty of learning tasks
• How to improve learning – e.g. value of active versus passive learning
• Computational architectures for learning
Machine Learning – related disciplines

- **Applied Statistics**
  - Emphasizes statistical models of data
  - Methods typically applied to small data sets
  - Often done by a statistician increasingly assisted by a computer

- **Data Mining** – roots in databases

- **Pattern recognition** – roots in engineering

- **Machine learning**
  - Relies on (often, but not always statistical) inference from data and knowledge (when available)
  - Emphasizes efficient data structures and algorithms for learning from data
  - Characterizing what can be learned and under what conditions
  - Obtaining guarantees regarding the quality of learned models
  - Scalability to large, complex data sets (big data)
What is Machine Learning?

• A program $M$ is said to learn from experience $E$ with respect to some class of tasks $T$ and performance measure $P$ if its performance as measured by $P$ on tasks in $T$ in an environment $Z$ improves with experience $E$.

Example 1

$T$ – cancer diagnosis
$E$ – a set of diagnosed cases
$P$ – accuracy of diagnosis on new cases
$Z$ – noisy measurements, occasionally misdiagnosed training cases
$M$ – a program that runs on a general purpose computer
What is Machine Learning?

Example 2

$T$ – recommending movies e.g., on Netflix
$E$ – movie ratings data from individuals
$P$ – accuracy of predicted movie ratings

10% improvement in prediction accuracy – $1$ million prize
What is Machine Learning?

Example 3

$T$ – Predicting protein-RNA interactions

$E$ – A data set of known interactions

$P$ – accuracy of predicted interactions
What is Machine Learning?

Example 4

$T$ – Reconstructing functional connectivity of brains from brain activity (e.g., fMRI) data

$E$ – fMRI data

$P$ – accuracy of the reconstructed network
What is Machine Learning?

Example 5

$T$ – solving integral calculus problems, given rules of integral calculus

$E$ – a set of solved problems

$P$ – score on test consisting of problems not in $E$
What is Machine Learning?

Example 6

$T$ – predicting the risk of a disease before the onset of clinical symptoms

$E$ – longitudinal gut microbiome data coupled with diagnostic tests

$P$ – accuracy of predictions
What is Machine Learning?

Example 7

$T$ – predicting sleep quality from actigraphy data

$E$ – actigraphy data with sleep stage labels

$P$ – accuracy of predictions
What is Machine Learning?

Example 8

$T$ – Uncovering the causal relationship between exercise, diet and diabetes

$E$ – Data from observations and interventions (changes in diet, exercise)

$P$ – accuracy of causal predictions
Key requirements

- There is a pattern to be learned
- There are data to learn from

## Applicant information:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>23 years</td>
</tr>
<tr>
<td>gender</td>
<td>male</td>
</tr>
<tr>
<td>annual salary</td>
<td>$30,000</td>
</tr>
<tr>
<td>years in residence</td>
<td>1 year</td>
</tr>
<tr>
<td>years in job</td>
<td>1 year</td>
</tr>
<tr>
<td>current debt</td>
<td>$15,000</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Approve credit?
Learning to approve credit

Formalization:

- Input: \( x \) (customer application)
- Output: \( y \) (good/bad customer?)
- Target function: \( f : \mathcal{X} \rightarrow \mathcal{Y} \) (ideal credit approval formula)
- Data: \( (x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N) \) (historical records)
- Hypothesis: \( g : \mathcal{X} \rightarrow \mathcal{Y} \) (formula to be used)
Learning to approve a credit
Machine Learning

Unsupervised

Feature Extraction

Supervised

Machine Learning Algorithm

Training Set

Grouping of Objects

New Data

Predictive Model

Annotated Data
Course mechanics

Course page:

– http://faculty.ist.psu.edu/vhonavar/Courses/ds310/homepage.html
  • Syllabus
  • Texts
  • Study Guide
  • Course policies – grading, academic misconduct etc.

Assignments

– Readings (See study guide)
– Problem sets (Approximately bi-weekly)
– Lab Assignments, including programming (approximately 4)

Exams – 2 (midterm, final)
Course staff

• Instructor:
  – Dr. Vasant Honavar
  – Professor
  – E335 Westgate Bldg
  – vhonavar@ist.psu.edu
  – http://faculty.ist.psu.edu/vhonavar

• Teaching Assistant
  – Aria Khademi
  – PhD Student, Informatics
  – akhademi@ist.psu.edu
Course Prerequisites

• Problem solving skills

• Programming (Python) and data structures
  – Reading code
  – Designing programs based on specs
  – Writing well-designed, well-documented code

• Mathematics (multivariate differential calculus, basic probability and statistics, basic linear algebra)

• Writing and presentation skills
Course objectives

• Upon successful completion of the course, you should be able to:
  • Look at a problem and identify if ML is an appropriate solution
  • If so, identify what ML algorithms might be applicable
  • Understand why and how ML algorithms work and when and why they might fail
  • Adapt ML algorithms or implement ML algorithms
  • Apply those algorithms
  • Rigorously evaluate the results
  • Communicate results and any caveats

– In order to get there, you will need to:
  • Work through the relevant mathematics (calculus, probability, statistics, linear algebra, optimization)
  • Familiarize yourself with the relevant tools
  • Read, write, and apply ML programs
On a lighter note..

Upon completion of the course, you will be able to laugh at these signs, or at least know why one might...
Textbooks

Required Textbooks


Recommended References

Labs

- Virtual machine with
  - Anaconda preinstalled
  - Python
  - Jupyter notebooks
  - SciPy, NumPy, Pandas, Scikit-Learn packages
- Details to be provided
What to expect

• Lectures cover concepts, relevant math, algorithms
• Assigned readings and problem sets (both completed on your own time) reinforce the material covered in the class
• Lab assignments (completed on your own time) will provide hands-on experience with applying or adapting existing algorithms or writing new algorithms
• Expect to stay busy and learn a lot
  – Rule of thumb: For each hour of class time, expect to spend three hours outside class
Grading

- Problem Sets: 20%
- Exam I: 25%
- Lab Assignment 25%
- Exam II: 25%
- Class participation: 5%

- 93% - 100%   A
- 90% - 93%    A-
- 87% - 90%    B+
- 83% - 87%    B
- 80% - 83%    B-
- 77% - 80%    C+
- 70% - 77%    C
- 60% - 70%    D
- 0% - 60%     F
Questions?