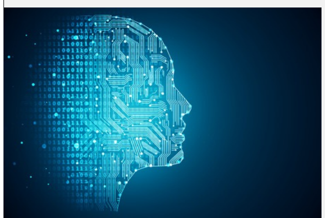


DS 310 Machine Learning

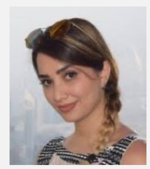
Vasant G. Honavar
Dorothy Foehr Huck and J. Lloyd Huck Chair in Biomedical Data Sciences and Artificial Intelligence
Professor of Data Sciences, Informatics, Computer Science and Engineering, Bioinformatics & Genomics, Public Health Sciences and Neuroscience
Director, Center for Artificial Intelligence Foundations and Scientific Applications
Associate Director, Institute for Computational and Data Sciences
Pennsylvania State University

vhonavar@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 Artificial Intelligence Foundations and Scientific Applications (CENSAI)
  - E335 Westgate Building
  - vhonavar@psu.edu
  - http://faculty.ist.psu.edu/vhonavar

- Teaching Assistant
  - Sahar Hanifi
  - PhD Student, Informatics
  - szh6071@psu.edu



- Students?

What I do

- 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: 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
  - Materials Informatics: Predicting material properties from structure and composition
- 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

About the course

- What is Machine Learning?
- What can you to expect to learn in the course?
- Course mechanics
  - Syllabus
  - Prerequisites
  - Expectations
  - Course policies
  - Course materials
  - Grading
  - …

What is this course about?

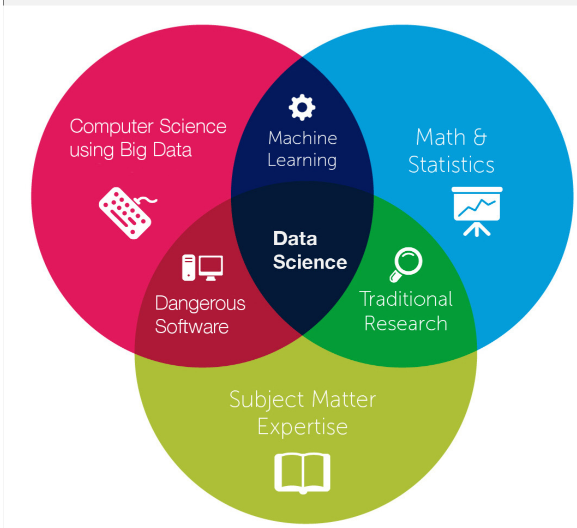
- Learning predictive models from data
  - Why should machines learn?
  - What does it mean for a machine to learn?
  - What can machines learn?
  - How can machines learn?
  - How can we evaluate learned models?
  - How can machines learn better?

Machine learning is a subfield of artificial intelligence

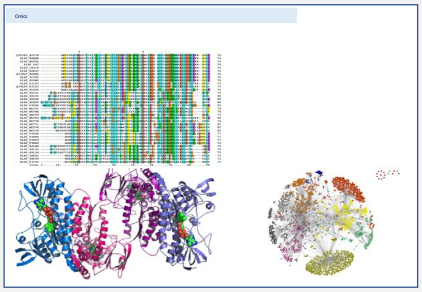
AI is about

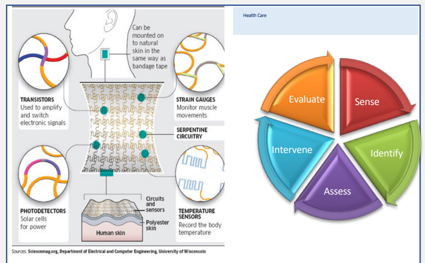
- Study of computational models of intelligence
- Falsifiable hypotheses about intelligent behavior
- Construction of intelligent artifacts
- Mechanization of tasks requiring intelligence
- Exploring the design space of intelligent systems

Machine learning is a subfield of Data Science



Machine learning is essential for extracting knowledge from data

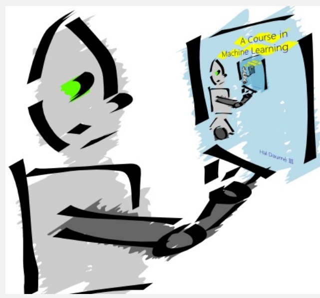




Source: Keith Marzullo

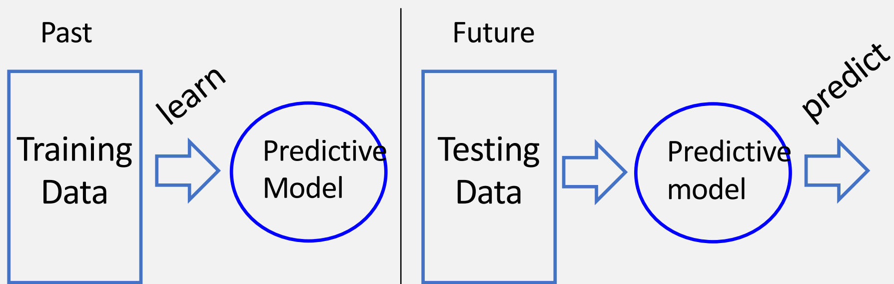
Machine Learning is…

About (computationally) predicting the future based on the past



Machine Learning is…

Machine learning is about (computationally) predicting the future based on the past



Machine Learning is…

- About methods for detecting patterns in data, and using the uncovered patterns to predict the future
- Concerned with methods for extracting actionable knowledge from data

Why should machines learn?

Machine learning is about replacing humans writing code for specific tasks with humans supplying data and 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 she arrives at a diagnosis
- Task parameters often vary across users
  - Detecting SPAM
  - Recommending products
  - Predicting treatment outcomes
  - ….

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 a particular task from observations (data) or interactions (experiments), we can
  - Dramatically reduce the cost of developing AI systems
  - Dramatically accelerate knowledge acquisition from data
  - Dramatically accelerate scientific discovery
  - Dramatically improve healthcare, education, public policy, manufacturing, ….
  - …

Center for Artificial Intelligence Foundations & Scientific Applications
Artificial Intelligence Research Laboratory

Why should machines learn? – Science of learning

Machine learning offers algorithmic models of learning that 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 signal and image processing
- 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 – personalized movie recommendation, 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* – Predicting material properties from material composition or material structure
- *E* – Databases of materials – composition, structure, properties
- *P* – accuracy of material property predictions

What is Machine Learning?

Example 9

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

What is Machine Learning?

Example 9

- *T* – driving a car
- *E* – Observations of driver actions under a broad range of conditions
- *P* – suitable measure of good driving – safety, efficiency, ...

Key requirements

- There are patterns to be learned
- There are data to learn from

Applicant information:

age	23 years
gender	male
annual salary	\$30,000
years in residence	1 year
years in job	1 year
current debt	\$15,000
...	...

Approve credit?

Learning to approve credit

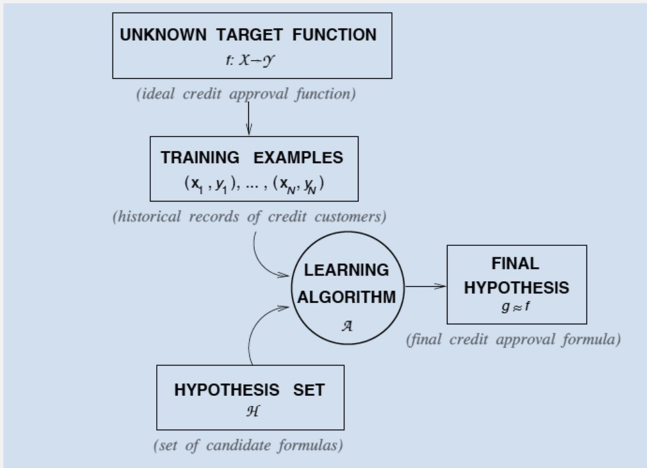
Formalization:

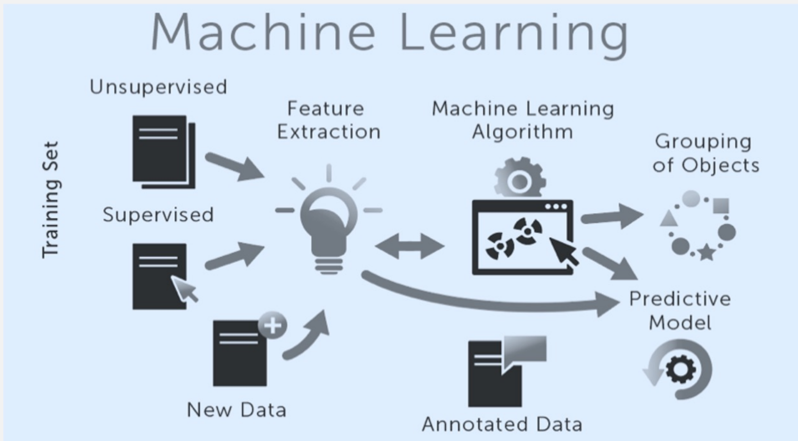
- Input: $\mathbf{x}$ (*customer application*)
- Output: y (*good/bad customer?*)
- Target function: $f : \mathcal{X} \to \mathcal{Y}$ (*ideal credit approval formula*)
- Data: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)$ (*historical records*)

$\downarrow \quad \downarrow \quad \downarrow$

- Hypothesis: $g : \mathcal{X} \to \mathcal{Y}$ (*formula to be used*)

Learning to approve to credit




Machine Learning
Unsupervised
Training Set
Supervised
Feature Extraction
Machine Learning Algorithm
Grouping of Objects
Predictive Model
New Data
Annotated Data

Course mechanics

Course page:

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

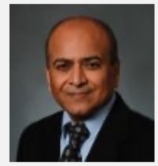
Assignments

- Readings (See study guide)
- Problem sets (Posted on Canvas) 8 – 10
- Lab Assignments (Posted on Canvas) 6 – 8
- Projects (Posted on Kaggle) – 2

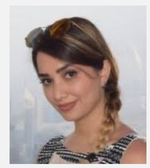
Exams – 2 (midterm, final)

Course staff

- Instructor
  - Dr. Vasant Honavar
  - Professor, IST & CSE, Data Science, BG, Neuroscience
  - Director, Artificial Intelligence Research Lab
  - Director, Center for Artificial Intelligence Foundations and Scientific Applications (CENAI)
  - E335 Westgate Building
  - vhonavar@psu.edu
  - http://faculty.ist.psu.edu/vhonavar



- Teaching Assistant
  - Sahar Hanifi
  - PhD Student, Informatics
  - szh6071@psu.edu



Course Prerequisites

- Algorithmic problem solving
- Programming (Python) and data structures
  - Reading and writing code
- Mathematics
  - Multivariable differential calculus
  - Elementary probability theory
  - Elementary 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 or implement ML algorithms to solve specific ML problems
  - Apply ML algorithms to real-world problems
  - Rigorously evaluate the results
  - Communicate results and any caveats
  - Practice ML responsibly
- In order to get there, you will need to:
  - Work through the relevant mathematics
  - 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

- Daume III, Hal (2017). A course in machine learning. Freely available for download online.

Recommended References

- Watt, J., Borhani, R., Katsagellos, A. (2020). Machine Learning Refined. Cambridge University Press. Available online through Penn State Libraries
- Deisenroth, M.P., Faisal, A., and Ong, C.S. (2018) Math for Machine Learning Cambridge University Press. Available online through Penn State Libraries
- Behrman, K. (2022). Foundational Python for Data Science.
- Vanderplas, J. (2017). Python Data Science Handbook. O'Reilly. Freely available for online reading
- Chen, D. Y. (2018). Pandas for everyone. Pearson.

Labs

- We will use google colab: https://colab.research.google.com
- To access google colab:
  - Sign into your google account using your Penn State email
  - Go to https://colab.research.google.com
  - If you have multiple google accounts, please make sure that you switch to the account associated with your Penn State email address
  - We will share python notebooks on google colab with you using your Penn State email address

What to expect

- Lectures cover concepts, relevant math, algorithms
- Assigned readings and problem sets reinforce the material covered in the class
- Lab assignments will provide hands-on experience with ML algorithms and their applications using Python libraries
- Projects give you experience building, fine-tuning, evaluating, and selecting ML models for real-world problems
- 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%
- Lab Assignments: 20%
- Projects: 30%
- Exams: 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

Please consult course policies regarding late problem sets, assignments, and projects

Other policies

- Academic misconduct
- Copyright
- Disability accommodation
- Educational equity and non-discrimination
- Pandemic guidelines
- Emergency notifications

Resources

- Texts and References
- Study guide
- Resources
  - Tutoring service
  - Counseling
  - Crisis hotline
- ML Resources

Questions?