Machine Learning

Introduction

Dan Goldwasser

dgoldwas@purdue.edu

Goal for Today's class

Machine Learning?

What is it?

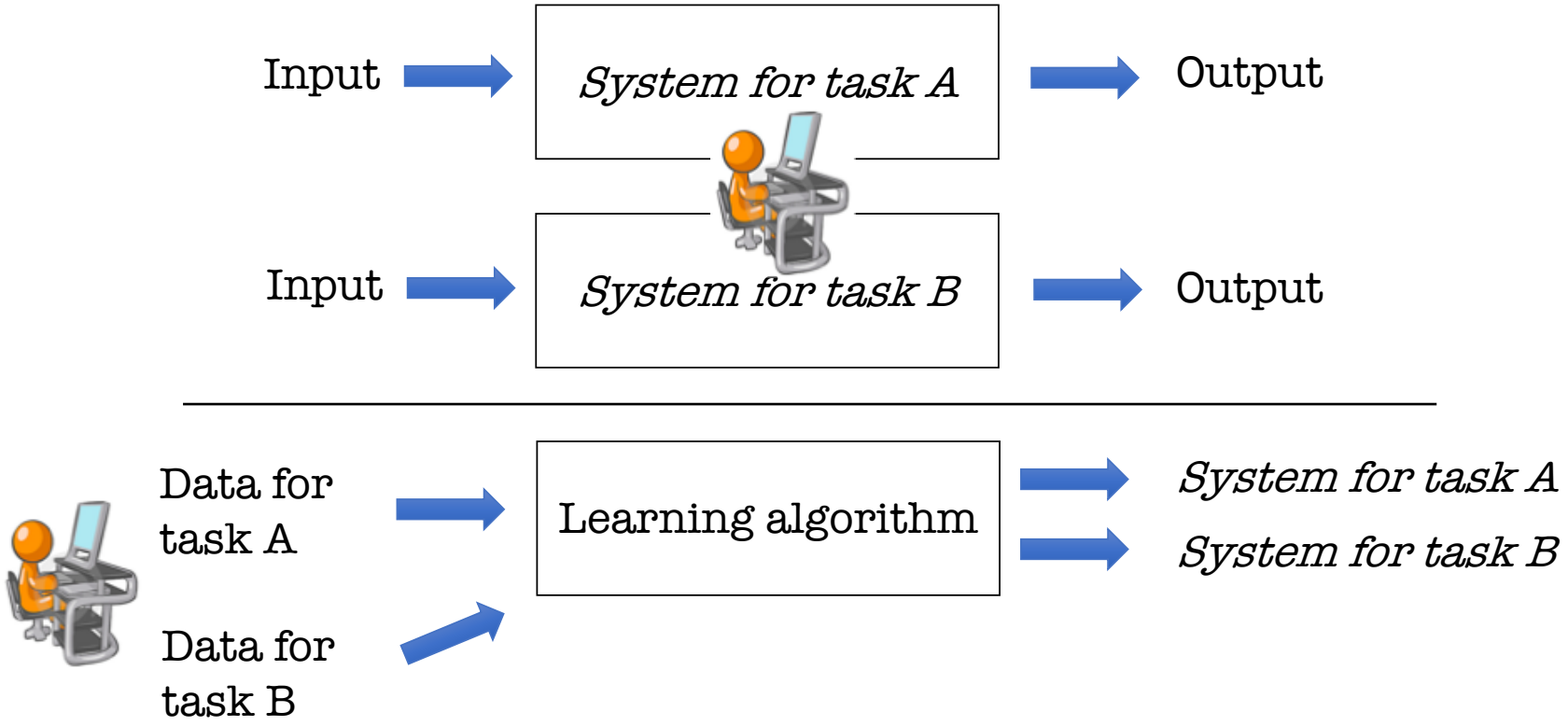
why should I care about it?

Also – grading, policies, etc.

Machine Learning in a nutshell

- A new paradigm for telling a computer what to do!
- **Traditionally:**
  - Write lines of code for different tasks
- **Machine learning**
  - Write code once (learning algorithm)
  - Supply data for different tasks
- Key issue: **Generalization**
- *How to generalize from training examples to new data*
  - Identify patterns in the data and abstract from training examples to testing examples

Machine Learning in a nutshell



Let's travel back in time.. 90's!

- The year was 1994, the conference was CoLT
  - CoLT = Computational Learning Theory

- Participants got a badge with a +/- label
- Only organizer knew the function generating the labels
- Label is determined only by the participant's name

- **Task**: *look at many examples as you want in the conference, and induce the unknown function*
- A great ice breaker for parties!

The badges game

Can you figure out the function?

Name	Label
Claire Cardie	-
Peter Bartlett	+
Eric Baum	
Haym Hirsh	
Shai Ben-David	
Michael I. Jordan	

Do you need more examples?

The badges game

Can you figure out the function?

Name	Label
Claire Cardie	-
Peter Bartlett	+
Eric Baum	-
Haym Hirsh	+
Shai Ben-David	-
Michael I. Jordan	+

Let's travel back in time.. 90's!

- How did you do it?
  - Not a trivial process, even for humans.
  - Requires raising hypotheses, validating those on data, accepting or rejecting them.

- Can you find an algorithm to do the same process?
  - **A form of search..**
  - *What are you searching over?*
  - *How would you bias the search?*
  - *When would you stop?*

This is machine learning!

What is machine learning?

"Field of study that gives computers the ability to learn without being explicitly programmed"
Arthur Samuel (1959)

"A computer program is said to learn from **experience E** with respect to some **class of tasks T** and **performance measure P**, if its performance at tasks in T, as measured by P, **improves with experience** E."
Tom Mitchell (1999)

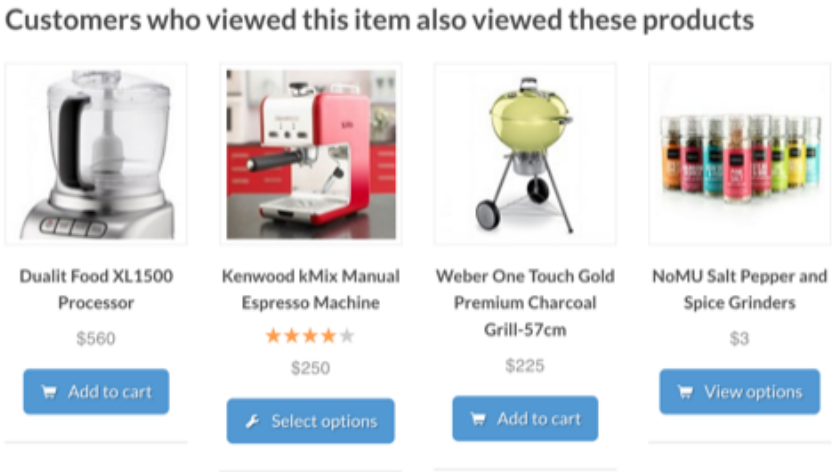
Why Study Machine Learning?

- **Basically** :

  if it works so well, I can probably use it..

- Push the limits of existing systems
  - Flexible, personalized computing
- Systems that can react correctly to previously unseen situations
  - Different paradigm for programming and testing
- Make sense of messy data
  - Traditional methods (coding) cannot handle it

It's everywhere!



Google Translate

Break through language barriers

It's everywhere!

- Spam detection
- Face recognition
- Content recommendation (e.g., movies, books)
- Stock market prediction
- Handwriting recognition
- Speech recognition
- Weather prediction
- Translation
- Automated Personal Assistants
- Self-driving car
- Disease diagnostic

In this class..

- We will look into several learning protocols, using different learning algorithms, for learning different models.
  - Model: function mapping inputs to outputs
  - Algorithm: used for constructing a model, based on data
  - Protocol: the settings in which the algorithm learns.
- Not an introduction to ML tools!
  - Understand the algorithms, their properties and theory behind them.

Learning Model

- We think about learning as producing a function mapping input to outputs, based on data
  - E.g., spam: email → Boolean
- Model is the type of function the learner looks at
  - Linear Classifier
  - Decision Trees
  - Non Linear Classifiers
  - Ensembles of classifiers
- All of these models can be characterized and discussed formally in terms of their expressivity.

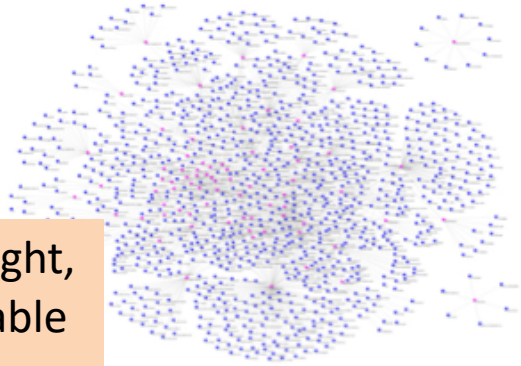
Learning Model Representation

- Learning: producing a function mapping input to outputs, based on data
  - E.g., spam: email → Boolean
- What is the domain and range of that function?
  - Output space defines the learning task
    - Binary, multiclass, continuous, structure,..
  - What is the domain? How is the data represented?

(4.2, 3.5, ...,1.3)

(1,0,0, 2,..0)

I have never had a worst flight,
The seats were uncomfortable
and it was 3 hours late!



Learning Protocols

- Supervised learning
  - Human (teacher) supplies a labeled examples
  - Learner has to learn a model using this data
- Unsupervised learning
  - No teacher, learner has only unlabeled examples
  - Data mining: finding patterns in unlabeled data
- Semi-supervised learning
  - Learner has access to both labeled and unlabeled examples
- Active learning
  - Learner and teacher interact
  - Learner can ask questions
- Reinforcement learning
  - Learner learns by interacting with the environment
  - Why is it different/similar to Supervised learning? Active learning?

Learning Algorithms

- Learning Algorithms generate a model, they work under the settings of a specific protocol
- Supervised vs. Semi-Supervised vs. Unsupervised
- Online vs. Batch
  - Online algorithm: learning is done one example at a time
    - Winnow, Perceptron,..
  - Batch algorithm: learning is done over entire dataset
    - SVM, Logistic Regression, Decision Trees, …
- How can we compare learning algorithms?

Learning Theory

- Will my algorithm work?
  - Train and then check by testing it
  - What could be a potential problem with this approach?
- Can we reason in a more rigorous, general way?
  - Mathematical definition of learning, guarantees of learning algorithms, and required resources
- Online learning
- PAC

Machine Learning Tasks

- Supervised Classification
  - Most popular scenario
  - Most of this class is about classification
- Regression
- Clustering (unsupervised)
- Structured Prediction
- Reinforcement learning

Classification

- Classification: mapping data into categories
  - Write a face recognition program
  - Determine if an English sentence is grammatical
  - Distinguish between normal and cancerous cells
- Can't we just write code?
- Provide labeled examples and let a classifier distinguish between the two classes
  - What are the labeled examples in each case?

Classification dataset

Class	Outlook	Temperature	Windy?
Play	Sunny	Low	Yes
No play	Sunny	High	Yes
No play	Sunny	High	No
Play	Overcast	Low	Yes
Play	Overcast	High	No
Play	Overcast	Low	No
No play	Rainy	Low	Yes
Play	Rainy	Low	No

- Three components
  - Class label (denoted “y”)
    - Binary, multiclass
  - Features (outlook, temperature, windy)
  - Feature values (denoted “x”)
    - Can be binary, continuous, categories (Sunny, Rainy,…)

Classification Dataset

Class	Outlook	Temperature	Windy?
Play	Sunny	Low	Yes
No play	Sunny	High	Yes
No play	Sunny	High	No
Play	Overcast	Low	Yes
Play	Overcast	High	No
Play	Overcast	Low	No
No play	Rainy	Low	Yes
Play	Rainy	Low	No

- A labeled dataset is a collection of (x,y) pairs
- Task: use dataset to predict new instance class

Class	Outlook	Temperature	Windy?
???	Sunny	Low	No

- Generalization: can we make reliable predictions?

Recipe for building a classifier

- Step 1: Collect Training Data
  - Can be a bottleneck when labeling is difficult
  - Can also come for free!
    - Car detector VS. Cancerous cell identification
- Step 2: Decide Feature representation
  - The classification function domain
  - Captures properties of the input object
    - Trivial decisions in some cases, difficult in others (which?)
  - “Art more than science”
  - Difficult to learn over bad feature representations
    - What is a bad representation for the weather domain?
    - .. And in general?
  - How can we be “more science than art?”

Recipe for building a classifier

- Step 3: Model (representation):
  - What "family" of functions to use?
    - Boolean functions, Linear functions (hyperplanes)
  - Known as the hypothesis space
  - Terminology: Classifier = Function = Hypothesis
- Step 4: Optimization (learning algorithm):
  - Which function to chose?
    - Search over possible functions (i.e., hypothesis space)
  - The learning algorithm uses the training data to choose a specific function
    - Bias the search
    - Evaluate candidate hypothesis

Review Questions

- How can you evaluate a learning system?
- Can learning fail?
- What does it mean and when would it happen?
- What is a hypothesis space?
- Why does choice of hypothesis space matters?
- Why is unsupervised learning important?
- How are cluster assignments different from class labels?
- What is structured learning? Is it the same as running a classifier multiple times?
- Why is reinforcement learning difficult?

Basic Concepts Summary

- Basic Concepts:
  - Supervised Learning: Learning with a teacher
  - Unsupervised: Learning without a teacher
- Classification
  - Design a supervised classification system:
    - Annotate data (determine labels)
    - Decide on Feature Representation
    - Choose Model and Algorithm
  - Generalization is the key issue

Course Policies

- When: Tuesday, Thursday 12:00
- Where: Seng-Liang Wang Hall 2599
- Office hours: Thursday after class(email me)
- **TA**:
  - Xiao (Cosmo) Zhang
  - Pradeep Kumar
  - Shamik Roy
  - See Piazza page for office hours (will be updated soon)

Required Background

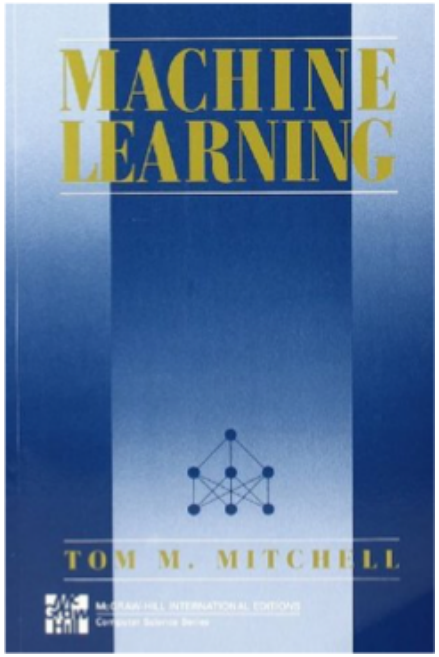
- You need a math background
  - Algebra (describe high dimensional data)
  - Calculus (finding min/max points of functions)
  - Stats and Probability (diggin' in data)
- It is a CS class
  - Algorithm design, complexity analysis, discrete math
    - Should sound familiar – recursion, dynamic programming, hashing function, graphs, trees, BFS, DFS
- It's not a purely theoretical class
  - You need some programming experience

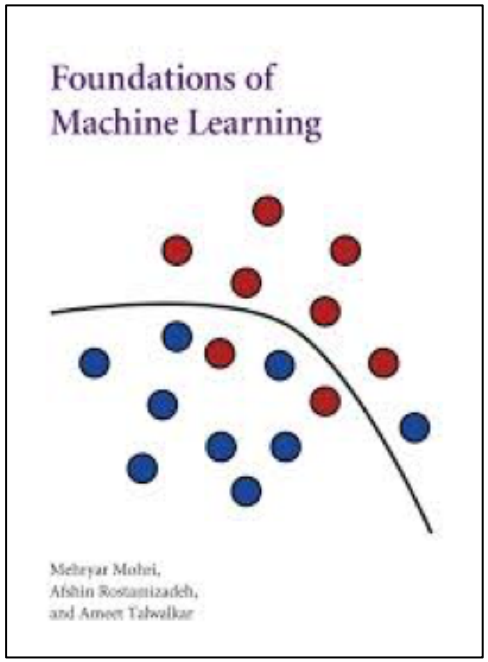
My Expectations

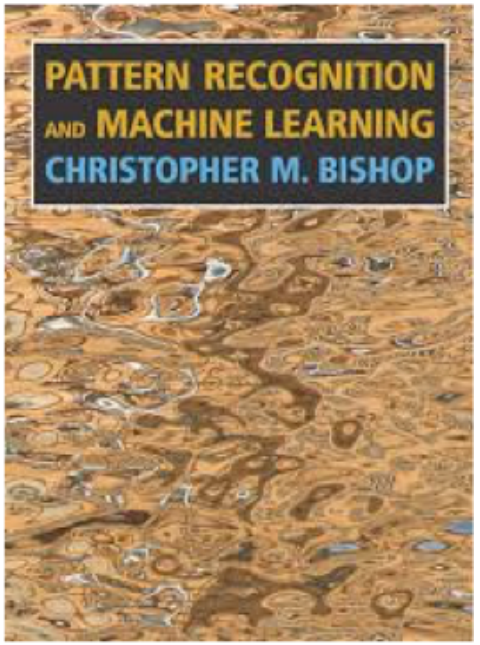
- Academic honesty
  - Submit your own work.
  - Load in group project should be evenly distributed
- Come prepared to class and participate
  - Review previous lecture notes and other materials
  - I will encourage class discussions
- Communicate
  - Inform teaching staff of any problems, or unclear topics
  - Interact with your peers, solve problems together

TextBooks

- No required textbook, several recommended books that I will point to







Communication

- Most of our interaction will be through Piazza
  - Sign up!
- https://piazza.com/purdue/fall2018/cs578/home
- Part of your grade is determined by your Piazza activity
  - Post questions and respond to questions
  - Express opinions, post related material
  - Related cool stuff: https://www.youtube.com/watch?v=DQWI1kvmwRg
- Communicate with teaching staff and come to office hours!

Grading

- Final exam: 40%
- Midterm: 25%
- Homework: 25%
- Attendance and Participation: 10%

Homework

- Written and Machine problems
  - About 5 assignments
  - Machine problems in Python
    - Other language require TA approval
- Late Policy
  - You have 48 hours of late submission for each assignment.
  - 10% will be deducted for each 24 hours.