CSE 417T
Introduction to Machine Learning

Instructor: Chien-Ju Ho
(the other two sections are taught by Henry Chai and Brad Flynn)
Teaching staff

• Instructors
  • Henry Chai
  • Bradley Flynn
  • Chien-Ju Ho

• TAs
  • Arghya Datta (Graduate Assistant to the Instructors)
    Adam Kern
    Chunyuan Li
    Daniel Teich
    Jack Grundy
    Mingyu Yang
    Ryun Han
    Trevor Larsen
    Zachary Mekus

• Office Hours will be announced by the 2\textsuperscript{nd} week of the semester

• Three sections:
  • (roughly) synchronized course progress
  • shared TA, same homework assignments, same exams
Plan for today

• Welcome and introduction
• What’s the class about?
• Logistics

• Lecture
  • Setting up the learning problem
  • Perceptron learning algorithm

• We’ll announce the first homework this week
What is machine learning?
Learn to tell if this is a tree

A brown trunk coming up from the ground, with branches extending out?
Are these trees?

Hard to define
”know it when I see it”  learned from data
What is machine learning?

“using a set of observations to uncover an underlying pattern”

“learning from data”

Use scenarios of machine learning

• A pattern exists
• No analytical solution: We cannot pin it down mathematically
• We have data on it
Spam detection

pattern

Is the pattern easy to define

data

“The US Department of the Treasury has decided to pay you $250,000.00 USD…”

“… Let’s meet in my office at 5pm…”
Movie recommendation

Netflix challenge:
1 million dollar prize for 10% improvement over their existing recommendation systems
Movie recommendation

viewer: likes comedy? likes action? prefers blockbusters? likes fish?

Match corresponding factors then add their contributions

predicted rating

movie: blockbuster? action content comedy content fishes in it?
Credit card approval

- Using salary, debt, years in residence, etc., approve for credit or not.
- No magic credit approval formula.
- Banks have lots of data.
  - customer information: salary, debt, etc.
  - whether or not they defaulted on their credit.

<table>
<thead>
<tr>
<th>age</th>
<th>32 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>gender</td>
<td>male</td>
</tr>
<tr>
<td>salary</td>
<td>40,000</td>
</tr>
<tr>
<td>debt</td>
<td>26,000</td>
</tr>
<tr>
<td>years in job</td>
<td>1 year</td>
</tr>
<tr>
<td>years at home</td>
<td>3 years</td>
</tr>
</tbody>
</table>

Approve for credit?
More formally for (supervised) learning

• Formulation: (credit card approval example)
  • input: \( x \in X \) (customer’s information)
  • output: \( y \in Y \) (good/bad customer)
  • unknown target function: \( f: X \to Y \) (ideal credit approval formula)
  • data \((x_1, y_1), \ldots, (x_N, y_N)\) (historical records)

• goal: learn a \( g \) close to \( f \) (formula to be used)

• Two central questions
  • How do we learn \( g \)?
  • What can we say about how close \( g \) is to \( f \)?
UNKNOWN TARGET FUNCTION

\[ f : \mathcal{X} \mapsto \mathcal{Y} \]

(ideal credit approval formula)

\[ y_n = f(x_n) \]

TRAINING EXAMPLES

\((x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\)

(historical records of credit customers)

LEARNING ALGORITHM \( A \)

FINAL HYPOTHESIS

\( g \approx f \)

(learned credit approval formula)
Given by the learning problem

UNKNOWN TARGET FUNCTION
\[ f : \mathcal{X} \rightarrow \mathcal{Y} \]

(ideal credit approval formula)

\[ y_n = f(x_n) \]

TRAINING EXAMPLES
\[(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\]

(historical records of credit customers)

LEARNING ALGORITHM
\[ \mathcal{A} \]

(HYPOTHESIS SET)
\[ \mathcal{H} \]

(set of candidate formulas)

FINAL HYPOTHESIS
\[ g \approx f \]

(learned credit approval formula)

learning model
Linear hypothesis space

- Suppose we have data on annual income ($x_1$), debt ($x_2$).
- A possible hypothesis space can be expressed as ($w_1, w_2, t$) where the credit approval function is
  
  Approve if $\sum_{i=1}^{2} w_i x_i \geq t$
  
  Deny if $\sum_{i=1}^{2} w_i x_i < t$

Note that there are infinitely many possible hypothesis!

How can we learn?
What can we say about what we’ve learned?
JELLY BEANS CAUSE ACNE!

SCIENTISTS! INVESTIGATE!

BUT WE'RE PLAYING MINECRAFT!

... FINE.

WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (p > 0.05).

THAT SETTLES THAT. I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.

SCIENTISTS!

BUT MINECRAFT!

From xkcd, by Randall Munroe: http://xkcd.com/882
From xkcd, by Randall Munroe: http://xkcd.com/882
Types of learning

- Supervised learning (main focus of this course)
  - Given training data (input, correct output),
  - try to predict the output for input not seen before

- Unsupervised learning
  - Find patterns in data given only (input)

- Reinforcement learning
  - Learn how to act, based on rewards for actions
Supervised learning

• Given data with (input, correct output), learn a pattern that can predict previously unseen data
Unsupervised learning (more in 517A)

• Suppose you only have the feature vectors but no labels. Still want to describe the data in some useful way.
Unsupervised learning (more in 517A)

N A Rosenberg et al. Science
2002;298:2381-2385
Reinforcement learning (more in 511A)

- Agent interacts with the world by taking actions
- Feedback is in the form of rewards (or costs)
- Agent must learn a policy, which maps from the state of the world to an action
- Major issues:
  - Delayed reward / credit assignment
  - Exploration / exploitation
Course plans (focus on supervised learning)
Course plans

• Foundations
  • What’s machine learning
  • Feasibility of learning
  • Generalization
  • Linear models
  • Non-linear transformations
  • Overfitting and how to avoid it
    • Regularization
    • Validation

• Techniques
  • Nearest neighbors
  • Decision tree
  • Support vector machine
  • Boosting
  • Random forest
  • Neural networks
  • ...

Textbooks

• **Learning From Data.**
  • Y. Abu-Mostafa, M. Magdon-Ismail, and H-T Lin.
  • http://amlbook.com/
  • We will go through this book in the first half of the semester.

• **Computer Age Statistical Inference: Algorithms, Evidence, and Data Science.**
  • B. Efron and T. Hastie.
  • https://web.stanford.edu/~hastie/CASI/
  • We will reference a few sections as the course materials. The PDF file is freely available on their website.
Logistics

• Course websites (check regularly for announcements)
  • http://classes.cec.wustl.edu/~cse417t
  • http://piazza.com/wustl/fall2018/cse417t

• Course grade
  • Homework assignments: 50%
    • Matlab sessions (will post announcements on the website)
    • Worst and second worst assignments will be discounted by 60% and 40%.
  • Exams (exam 1: 25%, exam 2: 25%)
    • 10/4/2018 (Thu) 6:30 PM - 8:30 PM
    • 12/5/2018 (Wed) 6:30 PM - 8:30 PM

• Late day policy for homework assignments
  • 5 late days in total. Maximum 2 late days per assignment
Logistics

• Collaboration and academic integrity policies

• Accommodations and resources
Questions?