Learning From Data
Lecture 1
The Learning Problem

Introduction
Motivation
Credit Default - A Running Example
Summary of the Learning Problem

M. Magdon-Ismail
CSCI 4100/6100

The Storyline

1. What is Learning?
2. Can We do it?
3. How to do it?
4. How to do it well?
5. General principles?
6. Advanced techniques.
7. Other Learning Paradigms.

our language will be mathematics . . .
our sword will be computer algorithms

Resources

   - slides: www.cs.rpi.edu/~magdon/courses/learn/slides.html
   - assignments: www.cs.rpi.edu/~magdon/courses/learn/assign.html

2. Text Book: Learning From Data
   Abu-Mostafa, Magdon-Ismail, Lin

   - discussion about any material in book including problems and exercises.
   - additional material

4. TA

5. Professor

6. Prerequisites? assignment #0
Let’s Define a Tree?

A brown trunk moving upwards and branching with leaves ...

Are These Trees?

Learning “What are Trees” is ‘Easy’
Defining is Hard; Recognizing is Easy

Hard to give a complete mathematical definition of a tree. Even a 3 year old can tell a tree from a non-tree. The 3 year old has learned from data.

Learning to Rate Movies

- Can we predict how a viewer would rate a movie?
- Why? So that Netflix can make better movie recommendations, and get more rentals.
- $1 million prize for a mere 10% improvement in their recommendation system.

Previous Ratings Reflect Future Ratings

- Viewer taste & movie content imply viewer rating.
- No magical formula to predict viewer rating.
- Netflix has data. We can learn to identify movie “categories” as well as viewer “preferences”

Credit Approval

Let's use a conceptual example to crystallize the issues.

A pattern exists. We don't know it. We have data to learn it.
**Credit Approval**

Let’s use a conceptual example to crystallize the issues.

- 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.

### Age: 32 years
### Gender: male
### Salary: 40,000
### Debt: 26,000
### Years in job: 1 year
### Years at home: 3 years

Approve for credit?

A pattern exists. We don’t know it. We have data to learn it.

**Learning**

- Start with a set of candidate hypotheses $\mathcal{H}$ which you think are likely to represent $f$.
  
  
  $$
  \mathcal{H} = \{ h_1, h_2, \ldots \}
  $$
  
  is called the hypothesis set or model.

- Select a hypothesis $g$ from $\mathcal{H}$. The way we do this is called a learning algorithm.

- Use $g$ for new customers. We hope $g \approx f$.

$\mathcal{X}, \mathcal{Y}$ and $\mathcal{D}$ are given by the learning problem; The target $f$ is fixed but unknown.

We learn the function $f$ from the data $\mathcal{D}$.

**Summary of the Learning Setup**

- UNKNOWN TARGET FUNCTION
  
  $$
  f : \mathcal{X} \mapsto \mathcal{Y}
  $$
  
  (ideal credit approval formula)

- TRAINING EXAMPLES
  
  $$(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$$
  
  (historical records of credit customers)

- LEARNING ALGORITHM
  
  $A$
  
  $g \approx f$
  
  (learned credit approval formula)

- FINAL HYPOTHESIS
  
  $f$

- HYPOTHESIS SET
  
  $\mathcal{H}$
  
  (set of candidate formulas)