## 0 <br> CDS Education <br> We explore, learn, and educate big minds.

## Updates

CMS will be up tonight.
Work on Project A (due next Tuesday Midnight).
If you joined late, permission numbers will be sent out soon.

+ Extended quiz deadline


## CDS Education

 Introduction to Machine Learning for Python
## Intro to Machine Learning and

 Linear Regression
## Machine Learning

## According to Wikipedia...

"Machine Learning is a subfield of computer science that gives computers the ability to learn without being explicitly programmed."


## Better Definition

## By Professor Kilian Q.

Weinberger
CS 4780/5780


## Traditional Computer Science

Data



Program $\square$


Output

Computer

## Machine Learning


$\longmapsto$ Program
Output $\square$


Computer

## Machine Learning

## Traditional CS



## My Definition

## Let's talk about functions...

$\mathrm{f}(\mathrm{x})$ ? = x ? = $\mathrm{x}^{\wedge}$ ? ?
Vertical line test?

## Functions

Function

## Something

## Something



Input

## ML Setup

Hypothesis: Some speculative relationship between the input space and output space

Input Space: Variable or set of variables(data)
Output Space: Target variable to estimate

## Supervised vs Unsupervised

Supervised learning problems...

- Known target variable info
- Validation examples

Unsupervised learning problems...

- Unknown target variables
- Difficult to validate

- Supervised learning:
given $\left(x_{1}, y_{1}\right), \ldots,\left(x_{n}, y_{n}\right)$, learn $f(x)=y$
- Unsupervised learning: given $x_{1}, \ldots, x_{n}$, learn patterns or structure
- Online learning: for $i=1, \ldots, n$, given $x_{i}$, predict and observe $y_{i}$, learn $f(x)=y$
- Active learning: for $i=1, \ldots, n$, choose $x_{i}$, predict and observe $y_{i}$, learn $f(x)=y$
- Reinforcement learning: for $i=1, \ldots, n$, choose $x_{i}$, predict $y_{i}$, observe reward $r_{i}$, learn $f(x)=y$


## Supervised Learning



## Setup

- Training / Validation split
- Feature variable(s)
- Target variable
- Train and Test

Sent Mail
spam
(372)
fir
Trash

## Machine Learning

## Machine Learning

## Traditional CS



## Training and Testing



## Train/Test Split

## Training

## Testing



## Supervised Learning

## Training

## Testing



## Validation Set

- Split data into two sets
- Train model on one and validate on another
- Advantages / Disadvantages?




## Output Space Properties

- Continuous - e.g. temperature, height, probability
- Discrete - e.g. car brands, race, Pokémon type, diagnosis



## Regression vs Classification



Classification


Regression

## What is Learned

Function

Weighted Sum

INPUT x


$$
y=B_{0}+B_{I} x_{1}+\ldots+B_{p} x_{p}
$$

## Objective function

- All ML problems are optimization problems
- Format: Minimize/Maximize Obj in terms of $x$.
- Subject to set of constraints
- Objective functions represent assumptions
- Value of objective is an estimation of error


## Calculating Error



## Linear Regression





## Linear Regression

$$
y=B_{0}+B_{1} x_{1}+\ldots+B_{p} x_{p}+\varepsilon
$$

What are the assumptions?

- Linear relationship
- B, the coefficient vector, does not depend on $x$
- There is an unremovable noise
- This noise is normally distributed about the line


## Objective: Least Squares Error (L2)

$$
\sum_{i=0}^{n}\left(y_{i}-\left(B_{0}+B_{1} x_{1}+\ldots+B_{p} x_{p}\right)\right)^{2}
$$

What does this minimize?
Why this form?

## Coming Up

Your problem set: Project A
Next week: Introduction to Classification
See you then!


