Week 3 Notes

By Allen Ge, Meishuai Li, Zeyang Zhu

Big Data

Introduction

- Every object generates data
- Why is this trend emerging now? Because now we have the technology to digitize all existing knowledge, whereas this thing did not exist earlier
- Information is coming from all different kinds of sources, which did not exist before
- A byproduct of our increasing use of technology
- 2020, data volume would be 40 zetabytes
    - 40 zetabyes = grains of sand on earth * 75
    - Data processing in last two years = data proc in last 3000 years

Analogies

Copernicus collected astronomical data – data for the visible world

Microscope opened up the invisible world

Atomic world was opened up by electron microscope

Supervisible world - Big data is a microscope

We collect lots of data and we use powerful algorithms to parse through this data

Emerging Big Data

Before

1. We thought of things
2. Wrote them down = knowledge

Now

1. Bunch of information- not knowledge
2. Play around with it until it becomes knowledge

Every action that we do creates data – that data is mostly meaningless in the raw form until someone contextualizes it and gives meaning to it

Why is Big Data Happening Now?

1. Development of internet
    a. Rapidly increasing the amount of data available
2. Ubiquity of small scale devices that can act as crowdsourced sensors
3. Accelerating data storage capacity and computing power at low cost
    a. Moore's law
    b. Development of cloud computing, e.g., AWS
    c. Development of distributed platforms e.g., Hadoop, Apache Spark
4. Development of ML algorithms to process this data
    a. One has led to the other Big data spurred ML, which in turn led to need for more data

Data Case: Social Media

What data could be collected?

- Interests
- Likes/posts/messages/friends
- Photographs
- Relationship status
- Birthday
- Work and educational history, etc.

Who could use this data?

- Finding potential matches – new feature coming up
- Advertisements
- Cambridge Analytica – Political parties around the world to spread targeted political campaigns
    - American and Indian voters, etc.

Data Case: Vending Machine

What data could be collected?

- What products sell the most
- Payment information about people (if they use electronic payment)
- Time/days when purchases peak or fade
- Locations of vending machines where people buy these products
- Locations of products inside the machines which were picked up

Who could use this data?

- Could help them decide where to place physical stores/distribution centers, etc. so as to maximize their profits
- Could also help in the design of better vending machines

Case Study: Flu Prediction

Old method – Accumulating info submitted by doctors about patient visits

2 weeks to reach CDC

Google uses search results of “flu” to predict outbreak in real-time

Issues:

- Flu mentioned several times in the news…ppl interested in this..drove up Google Search

Unique Case Studies

Big data has allowed us to uncover fairness issues

Big data has been used for disaster response

- Earthquake in Haiti

Big data for uprisings in Tunisia

Target example…Our buying patterns can tell more about our pregnancies than our actual families

Concerns with Big Data

Anything that's going to change the world, by definition should have the ability to change it for the worse

Data revolution also leading to issues

- Somebody is listening in
- Big brother
- My device knowing me

Huge implications for how we interact with machines

Deep Learning

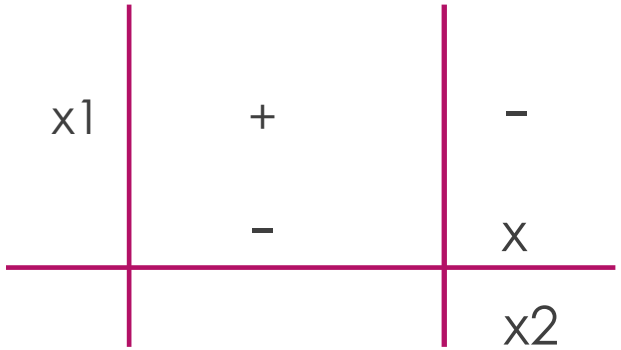
Deep Learning

- Classification:
  1. Classification Problems
  2. Decision Trees
  3. Random Forest
  4. GRT
  5. Neural Network
     - -Lots of temporary data
     - -Lots of computer power

Diff adjust (1920-2010)

→ 1960-1970

Initial Problems



XOR Example

Newsal networks

-Multiple layer perceptron (MLR) ⟷ NN

-Building block

Activation Functions

Non-Linear Activation Functions

1.Sigmoid Function

2. ReLU (Rectified Linear Unit)

3. Leaky ReLU

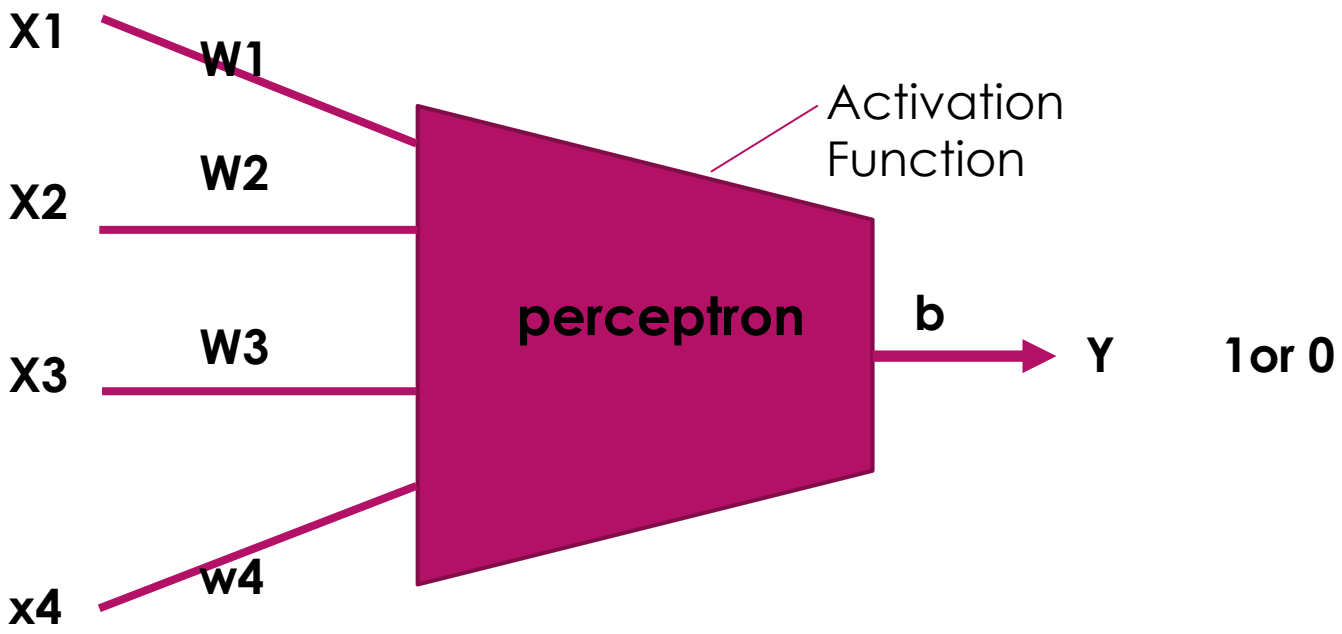
4.Tan H

w1x1+w2x2+w3x3+w4x4 >=b

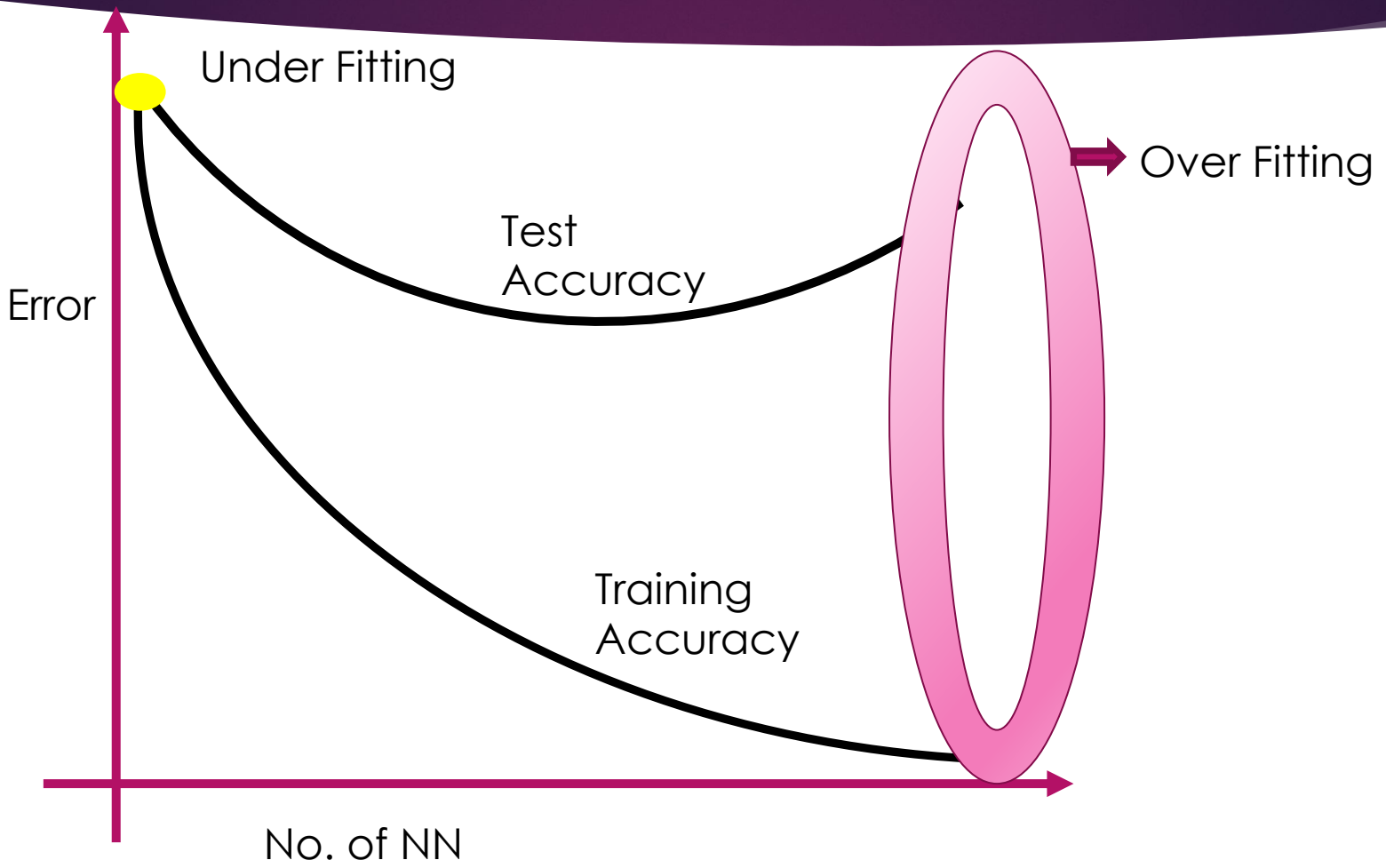
w1x1+w2x2+w3x3+w4x4 <b

$$\frac{1}{H * e \text{ ^-(w1x1+w2x2+w3x3+w4x4+b)}}$$

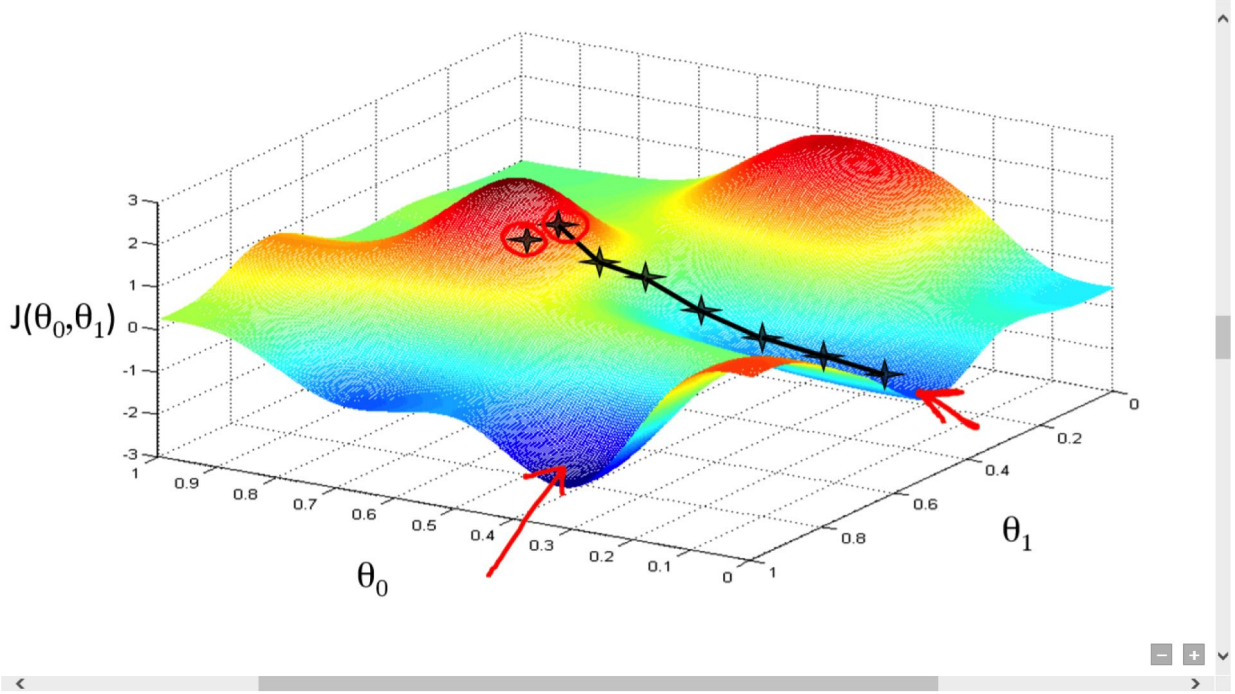
Perceptron


X1
W1
X2
W2
X3
W3
x4
w4
Activation
Function
perceptron
b
Y
1or 0

Bias variance trade off



Gradient Descent



- Deep network are trained using gradient descent
- Back propagation is often how you do gradient descent with neural networks