BITS F464: Machine Learning

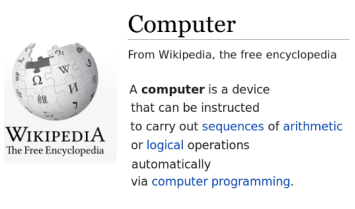
01 Logistics and Introduction

Dr. Kamlesh Tiwari
Assistant Professor, Department of CSIS,
BITS Pilani, Pilani Campus, Rajasthan-333031 INDIA

Jan 18, 2021 ONLINE (Campus @ BITS-Pilani Jan-May 2021)

http://ktiwari.in/ml

Computers

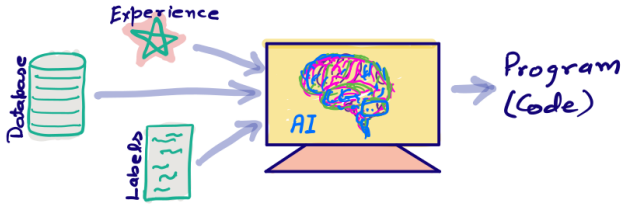


If it behaves smartly

Where is the **Intelligence**? with programmer (Human).

Introduction

We want intelligence in the computer. What does it mean?



Computers should write programs

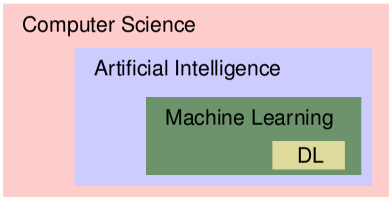
- Intelligence involves performing mundane, formal or expert tasks. Computers wish to use it to pass the **Turing Test**

Note: There are many tasks which computers **cannot** do. Surely!

Artificial Intelligence (AI)

Primary Question: How to make computers do things which at the moment, people do better [1]

AI attempts to build such intelligent entities



[1] There could be some tasks even humans are NOT good at.

Introduction

Computers are used to solve computational problems: Sorting, Searching, Determining the existence of Hamiltonian circuit (traversing every vertex once) or Euler walk (through every edge) in a graph

If we know how to solve the problem, we could write a program

But, for some problems we don't precisely know

- Is there a cat in figure? which cat?
- What is written on the board? Which language it is in?
- How to ask for a help from foreigner *etc.*

The problem is, either 1) we don't know how to solve, or 2) difficult to specify solution procedure

Then we go for **Machine Learning** (ML)

Similar problem was faced by Arthur Samuel

- When in 1956 he wanted to develop a Checkers playing program that could beat him
- Idea was to let the computer play lot of games against itself and learn effective moves
- In 1962 the computer won over human player
- Father of ML

Defined ML as a field of study that gives computers the ability to learn without being explicitly programmed.

Learning is a process, by which, a system improves performance from experience - **Herbert Simon**

Machine Learning: Definition

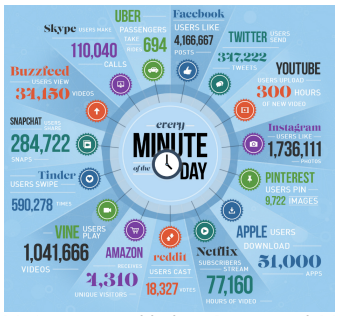
Definition: Tom Mitchell (1998)

a computer program is said to *learn* from experience **E** with respect to some *task* **T** and some performance measure **P**, if its performance on **T**, as measured by **P**, improves with experience **E**.

Performance of an algorithm does not solely depends on itself instead it takes into account of 1) training data, and 2) training methodology.

Data + Compute-Power

We are drowning in data.



What we do with the data

- Classification
- Regression
- Clustering
- Association Rule Mining
- Sequence Discovery

90% generated in last 2 year and 80% of it is unstructured [2].

[2] http://wersm.com/how-much-data-is-generated-every-minute-on-social-media/

Machine Learning: Tasks

Two broad categories of ML models are *Predictive* and *Descriptive*.

- Tasks
  - A. Predictive
    - 1. Classification
    - 2. Regression
    - 3. Time Series Analysis
    - 4. Prediction
  - B. Descriptive
    - 1. Clustering
    - 2. Summarization
    - 3. Association Rules
    - 4. Sequence Discovery

Some of the related tasks are

Classification

Classification maps data into *predefined* labels.

Example: Lots of mails are there in my mail box. Can you tell me which are SPAM?

- Task of supervised learning
- Often based on some patterns or characteristics
- We can use the frequency of words
- Assumption is that some words appears more or less frequently in a SPAM mail

Regression

Regression is used to map data into *real valued* variable.

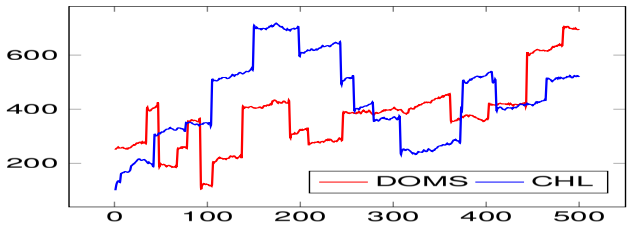
Example: What is the cost of my house?

- Task of supervised learning
- We have data about the cost of house based on features such as
  - location
  - Plot area
  - number of rooms
  - garden available or not
  - how old it is
- Current economical conditions can also matter
- Dimensionality is high

Time Series Analysis

In time series analysis the value of attribute is examined over time.

Example: Which stock is better?



- The values are obtained as evenly spaced time points (daily, weekly, hourly, etc.)
- Distance measures are used to find similarity
- Structural analysis is done

Prediction

Predicting future data states based on current or historical data.

Example: What comes next?

$$2, 4, 6, 8, 10, ...?...$$

$$2, 3, 5, 7, ...?..., 13$$

$$(10jul, rain), (11jul, rain), (12jul, no - rain), (13jul, ...?...)$$

- Predication can sometimes be seen as classification
- Application includes weather, flood, pattern recognition.

Clustering

Clustering is similar to classification except the groups are not pre-defined.

Example: How many kind of files are there in my directory?

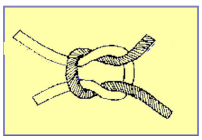
- Unsupervised learning setting
- We can use file name
- Words it has

Example: Who would take my offer?

- The database has information about age, gender, income, location, .. etc.

Summarization

Summarization maps data into subsets with associated simple descriptions (important details of main idea). It is also called characterization or generalization.



Example: How to compare two universities?

- Average JEE rank
- Average number of publication
- Student/Faculty ratio
- Combination

Association Rules

Association rules tries to do linked analysis.

Example: Whether sames products are selling together?

- $I = \{i_1, i_2, i_3, ...i_m\}$, $T = \{t_1, t_2, t_3, ..., t_n\}$ and $t_i \subseteq I$
- Minimum support count should be maintained
- Can you see: Subset of frequent items is also frequent
- Let $t_1 = (1, 3, 4)$, $t_2 = (2, 3, 5)$, $t_3 = (1, 2, 3, 5)$, $t_4 = (2, 5)$, $t_5 = (1, 4, 5)$
- What about $(2, 5)$?
- Apriori analysis can be applied

Sequence Discovery

Sequence Discovery is used to discover sequential patterns is the data.

Example: what is my website access pattern?

- Pattern is based on a time sequence of a action
- It is pattern discovery problem

Types of Learning

- **Supervised:** "right answers" are provided for sufficient training examples. Computer tells "right answers" for new input. Performance measure. (Classification and regression)
- **Unsupervised:** "right answers" are NOT provided and the computer tries to make sense of the data. How good the spread of items is. (clustering and association rule)
- **Semi-supervised:** "right answers" are provided for few training examples only
- **Active:** computer can ask questions. Needs less training. Opposite is passive learning
- **Lazy:** learner do not consolidate the findings.
- **Reinforced:** hit and trial method to minimize cost. (game playing)
- **Transfer:** Learning a task B to do A. (cycle riding for bike riding)
- **Deep:** processing like human brain

Challenges

1. How good is the model
   Accuracy, CRR, EER, FAR, FRR, ROC ...
2. How do I choose a model
   Decision Tree, SVM, Neural Network, ... ?
3. Do I have enough data
   Pre-processing, augmentation, ...?
4. Is data of sufficient quality
   Error/Noise in data, missing values ... ?
5. How confidence the result is
   Significance, Probability ... ?
6. Am I describing the data correctly
   whether features are correct ?
7. How fair the system is
   if it behaves equally to all?

The **Best** Machine Learning Model

No free lunch theorem. There is NO best model

that achieves the best test error for every problem

- If model A works better than model B on one dataset, then there is another dataset where model B works better then A

However it should be noted that:

1. The world is very structured, some datasets are more likely than others
2. Model A could be better than model B on a huge variety of practical applications

Applications of ML

In many domains including finance, robotics, bioinformatics, vision, natural language, *etc*.

- Spam filtering
- Speech/handwriting recognition
- Object detection/recognition
- Weather prediction
- Stock market analysis
- Search engines (e.g, Google)
- Ad placement on websites
- Adaptive website design
- Credit-card fraud detection
- Webpage clustering (e.g.,Google News)
- Machine Translation (e.g., Google Translate)
- Recommendation systems (e.g., Netflix, Amazon)
- Classifying DNA sequences
- Automatic vehicle navigation
- Performance tuning of computer systems
- Predicting good compilation flags for programs

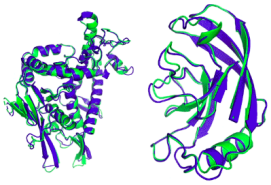
and many more...

Success Stories[3]

AlphaFold-2 solved 50-year-old grand challenge in biology

Determine a proteins 3D shape from its amino-acid sequence

- Typical protein: 10^{300} formations
- Nov 2020.
- Useful to Drug discovery and fundamental biological research
- Global Distance Test (GDT) score improved 40→87.4→92.4
- Use DL, attention, evolutionary sequence alignment



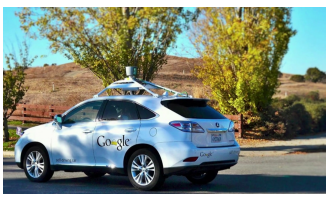
[3] https://www.predictioncenter.org/index.cgi
https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

Success Stories

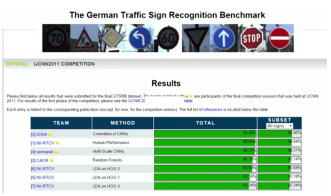
- **Deep Blue** by IBM
- 1997 won over Chess grandmaster Garry Kasparov
- 1996 earlier matches WLLLDD

- **AlphaGo**, RL agent playing GO
- 2016 won by 4-1 over *Lee Sedol* best human player (Why lost?)
- Policy and value N/W
- **AlphaZero** chess, shogi, & GO

Success Stories

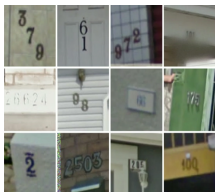


- Waymo: A safer driver that is always alert and never distracted
- First driverless ride on public roads in 2015 giving a ride to a sole blind
- In public: coming soon!



- German Traffic Sign Recognition Benchmark (GTSRB)
- 99.46% against 98.84% of human

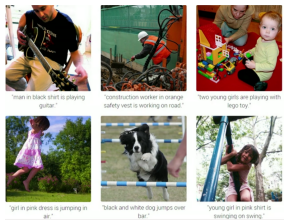
Success Stories



- Google mapped every single location in France in two hour
- Images acquired from Google street view

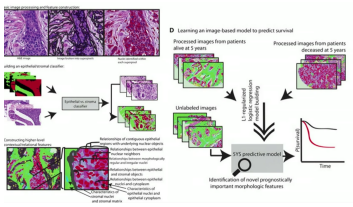
- Example of an image search
- That can take care of color and pose of the object in the image

Success Stories



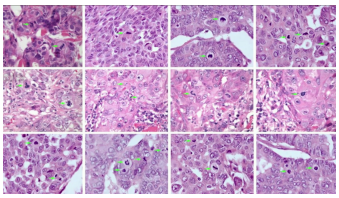
Computers can write

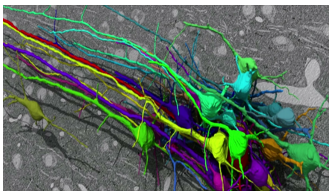
- Man in black shirt is playing guitar
- Two young girls are playing with Lego toy
- Black and white dog jumps over bar



- Tissues in magnification
- Stanford developed a ML algorithm that is better than human pathologist
- In predicting survival rate of cancer suffering

Success Stories

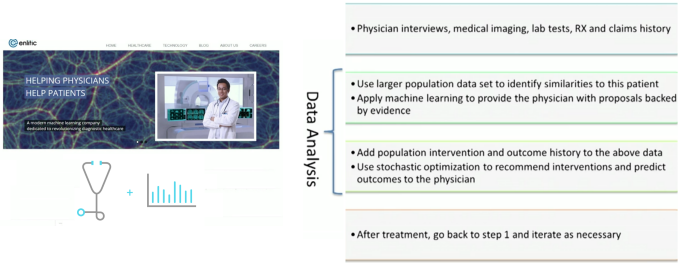




- System were developed by ML experts alone
- Without having any background in chemistry, biology or life sciences
- To identify cancerous tissues under microscope
- To perform neuron segmentation

Success Stories

It is possible to suggest very useful medicines by using just the data analytics techniques

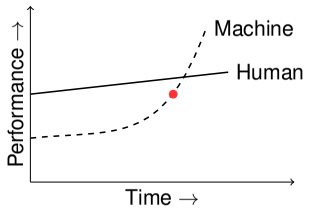


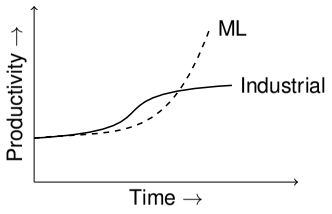
- `Enlitic` uses deep learning to make doctors faster and more accurate
- One have to use the middle path

Success Stories

By using following **four capabilities**, humans can do most of the work ($\sim$ 80%) like driving cars, preparing food, diagnosing diseases, Finding legal precedents, .. *etc.*

1. Reading and writing
2. Speaking and listening
3. Looking at things
4. Integrating knowledge





Some Quotes

- ⋆ `A breakthrough in ML would worth ten Microsoft`
  – ***Bill Gates***, Chairman Microsoft
- ⋆ `ML is new hot thing`
  – ***John Hennesy***, President, Stanford
- ⋆ `Web rankings today are mostly a matter of ML`
  – ***Prabhakar Raghavan***, Director Yahoo Research
- ⋆ `ML is going to result in a real revolution`
  – ***Greg Papadopoulous***, CTO Sun
- ⋆ `ML is today's discontinuity`
  – ***Jerry Yang***, CEO Yahoo
- ⋆ `ML is next Internet`
  – ***Tony Tether***, Director DARPA

Goal and Evaluation

Evaluation Scheme

1.	Mid Semester Test (Open Book)	25%
2.	Lab Submissions. (best 5 out of 7 evaluated)	10%
3.	Assignment (2) latex and notes-scribing	05%
4.	Comprehensive Exam (Open Book) 05 May 2021	35%
5.	Class Project - collect data, run model, write report	15%
6.	Term Project: presentation of research paper	10%

Goal: We expect, at the end, you would be able to get big picture, appreciate how various ML algorithms work, implement them on your own. Identify problems where ML can be applied. Capable enough to apply available tools to solve some of them.

Solution must always be written independently

Thank You!

Thank you very much for your attention!

Queries ?

`Any sufficiently advanced technology is indistinguishable from magic` – Arthur C. Clarke