CS 419: Intro to ML (for Non-CS students)

Lecture 0. Motivation and Logistics

Spring 2023

Logistics

(To be updated at Moodle and Team)

Course grading scheme (5–10% movement possible):

MidSem	25 %
EndSem	40 %
Scribe	5 %
Assignment	20%
Paper reviewing	10%

Scribes

- Two sets of Three-Four students are assigned to write notes of a class happened in date X in 2–4 pages using latex typesetting. Deadline is X+4 days.
- TAs will moderate/edit the scribe notes and will release a final scribe in X+8 days.
- One student gets only one chance of writing scribe notes. So, negligence in writing will result in zero credit out of the 5% quota.
- TAs will prepare scribe from the first week to set an example.
- scribe assignment across students will be released this weekend.

Personnel & Attendance Policy

Course TAs: Indra, Pritish, Jeel, more to come.
Communication: Apart from class, we will also use MSTeam mostly (and Moodle) to make course announcements.
Attendance: I want participation rather than mere attendance. Classes will be interactive. Constructive discussions will earn credits.
Doubt clearing three days before exam: No chance. Prepare early.

Course Material

- Slides as well as some detailed course notes will be posted on team later on.
- The following books are recommended:
    - Last year course materials: `https://rebrand.ly/cs419-2021`
    - **Elements of Statistical Learning**, by Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer, 2009. *Available online*.
    - **Pattern Recognition and Machine Learning**, by Christopher Bishop, Springer, 2006.
    - **Artificial Intelligence: A Modern Approach**, by Stuart J. Russell and Peter Norvig, 3rd edition, Pearson, 2010.
    - **Deep Learning**, by Ian Goodfellow and Yoshua Bengio and Aaron Courville
    - **Machine Learning**, by Tom Mitchell. McGraw-Hill, 1997.
    - **Understanding Machine Learning**, Shai Shalev-Shwartz and Shai Ben-David, Cambridge University Press. 2017. *Available online*.

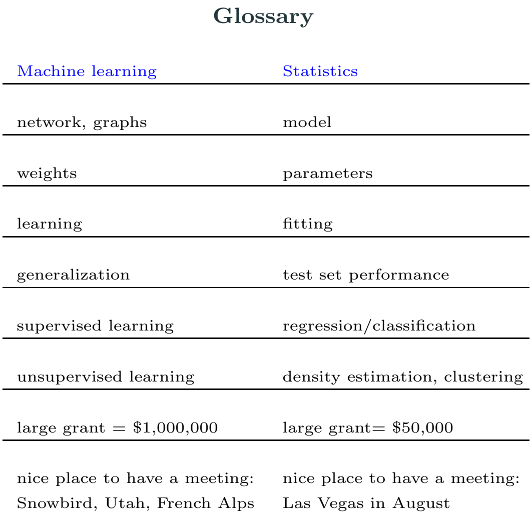
Introduction: What is Machine Learning?

- Machine learning (ML) is a sub-field of computer science that evolved from the study of **pattern recognition** and **computational learning theory** in artificial intelligence.

  In simpler terms:

- Using *algorithms* that iteratively learn from *data*
- Allowing computers to discover *patterns* without being explicitly programmed where to look

ML ↔ Statistics



Glossary

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = \$1,000,000	large grant= \$50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

Glossary from: http://statweb.stanford.edu/ tibs/stat315a/glossary.pdf

Relationship between AI, ML, DL

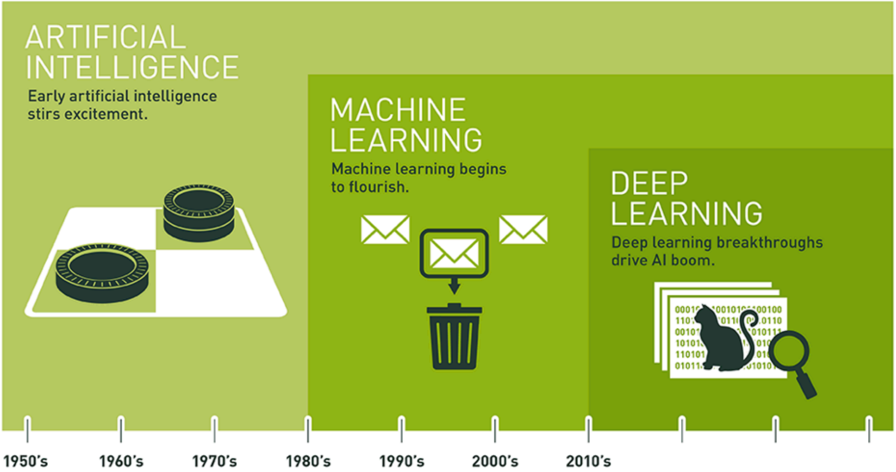

ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.
MACHINE LEARNING
Machine learning begins to flourish.
DEEP LEARNING
Deep learning breakthroughs drive AI boom.
1950's
1960's
1970's
1980's
1990's
2000's
2010's

Image from: https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

When do we need AI/ML? (I)

- For tasks that are easily performed by humans but are complex for computer systems to emulate
- Examples include
  - **Vision:** Identify faces in a photograph, objects in a video or still image, etc.
  - **Natural language:** Translate a sentence from Hindi to English, question answering, etc.
  - **Speech:** Recognise spoken words, speaking sentences naturally
  - **Game playing:** Play games like chess
  - **Robotics:** Walking, jumping, displaying emotions, etc.
  - Driving a car, flying a plane, navigating a maze, etc.

When do we need AI/ML? (II)

For tasks that are beyond human capabilities (E.g. IBM Watson’s Jeopardy-playing machine)

Image credit: https://i.ytimg.com/vi/P18EdAKuC1U/maxresdefault.jpg

What is Machine Learning?

- Ability of computers to **learn** from **data** or past experience
- **data:** Comes from various sources such as sensors, domain knowledge, experimental runs, etc.
- **learn:** Make *intelligent* predictions or decisions based on data
  - **Supervised learning**
  - **Unsupervised learning**
  - *Reinforcement learning. Will not be covered in this course.*

Supervised vs. Unsupervised Learning

Task: Suppose you had a basket filled with fresh fruits. Your task is to classify the same type of fruits together.
Suppose the fruits are apple, banana, cherry, grape.

Case 1:

- You already know: Shape (parameterize shape?), Color
- **Training data**: Pre-classified or labeled data
- **Goal:** Learn from the pre-classified or labeled data and predict on new unclassified fruits
- This type of learning is referred to as "**Supervised learning**"

Supervised vs. Unsupervised Learning

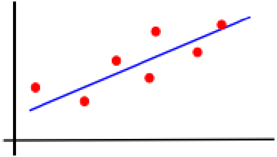
Case 2:

- In this case, you know nothing about the fruits! How will you group fruits of the same type together?
- One approach is to consider various characteristics of a fruit and divide them on the basis of that.
- Suppose you divide the fruits on the basis of *color* first:
  - **Red group**: Apples and cherries
  - **Green group**: Bananas and grapes
- Consider another physical characteristic *size*:
  - **Red and big**: Apple **Red and small**: Cherry
  - **Green and big**: Banana **Green and small**: Grapes
- This type of learning is referred to as **”Unsupervised learning”**

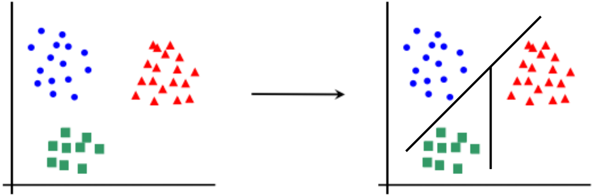
- In supervised learning, the desired outputs are provided which are used to train the machine whereas in unsupervised learning no desired outputs are provided. Instead the data is analysed and studied through clustering, mining associations, reduce dimensionality, *etc.* into different classes.

Three Canonical Learning Problems

1. Regression - Supervised
   - Estimate parameters, e.g. least square fit

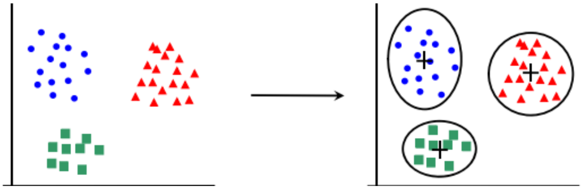


2. Classification - Supervised
   - E.g. Digit recognition

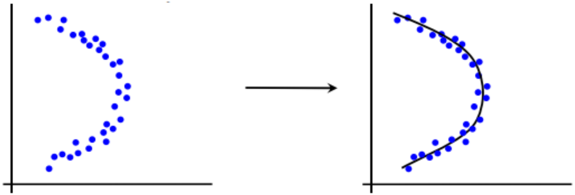


3 Unsupervised Learning - model the data

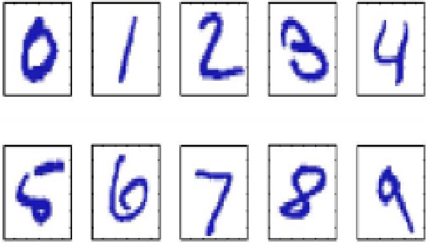
- Clustering



- Dimensionality reduction



Example of supervised learning: Handwritten digit recognition



Digit recognition: Images are $28 * 28$ **pixels**

- Represent input image as a vector x $\in R^{28*28}$
- Learn a classifier $f(x)$ such that,

$$\mathrm{f} : \mathrm{x} \rightarrow \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

Another example: Image recognition

Identify the object in the image

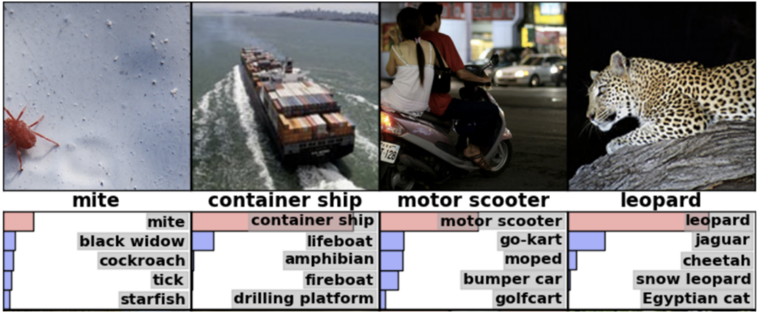


Image credit: "ImageNet classification with deep CNNs", Krizhevsky et al.,2012.

Course Overview

- Supervised Regression
  - Linear regression
    - **Introduction to Loss Functions**: Least squares & Likelihood
    - **Introduction to Prior/Regularization**: **Bayesian Regression**, **Ridge and Lasso**
    - **Optimization of Loss Function** + **Regularizer**: Gradient Descent,
  - Non-linear regression
    - **Introduction to Kernels:** SVR (support vector regression)

Course Overview

- Supervised classification
  - **Linear Classification:** Perceptron & Logistic Regression
  - **Non-Linear Classification:** Neural Networks and Support Vector Classifiers
  - **Deep Learning:** Convolution, Recurrence, LSTMs, etc.
  - **Bagging, Boosting and Feature Selection:** Illustration with Decision Trees
- Unsupervised learning
  - K-Means, K-Mediod and Hierarchical Clustering
  - Mixture of Gaussians and the general EM algorithm

Academic Integrity

We expect you to abide by an honor code, that you will not be involved in any sort of plagiarism.

All the assignments/quizzes and the project will be scrutinized. If caught for copying or plagiarism, the name of *both parties* will be handed over to the **Disciplinary Action Committee (DAC)**[1]

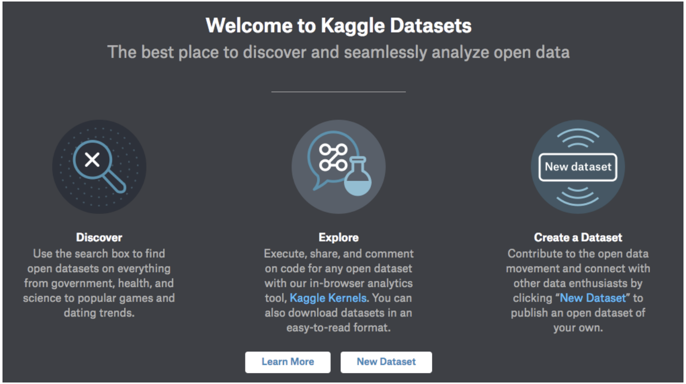
Any sort of plagiarism will be treated very seriously.[2].

[1] http://www1.iitb.ac.in/newacadhome/punishments201521July.pdf

[2] http://www1.iitb.ac.in/newacadhome/procedures201521July.pdf

Datasets abound...

Kaggle: `https://www.kaggle.com/datasets`



Welcome to Kaggle Datasets

The best place to discover and seamlessly analyze open data

Discover
Use the search box to find open datasets on everything from government, health, and science to popular games and dating trends.

Explore
Execute, share, and comment on code for any open dataset with our in-browser analytics tool, Kaggle Kernels. You can also download datasets in an easy-to-read format.

Create a Dataset
Contribute to the open data movement and connect with other data enthusiasts by clicking "New Dataset" to publish an open dataset of your own.

Datasets abound...

- **Kaggle:** `https://www.kaggle.com/datasets`
- Another good resource: `http://deeplearning.net/datasets/`
- Popular resource for ML beginners: `http://archive.ics.uci.edu/ml/index.php`
- Interesting datasets for computational journalists: `http://cjlab.stanford.edu/2015/09/30/lab-launch-and-data-sets/`
- Speech and language resources: `www.openslr.org/`

Typical ML approach

How do we approach an ML problem?

- **Modeling:** Use a model to represent the task
- **Learning:** The model could be parameterized and the parameters are learned using data
- **Decoding/Inference:** Given a model, answer questions with respect to the model

Textbooks

- Understanding Machine Learning by Shalev + Shai Ben David
- Probablistic Machine Learning by Kevin Murphy.
- Course notes