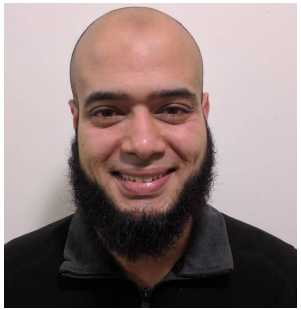
Machine Learning

Machine Learning Algorithms

Mostafa S. Ibrahim
Teaching, Training and Coaching for more than a decade!

Artificial Intelligence & Computer Vision Researcher
PhD from Simon Fraser University - Canada
Bachelor / MSc from Cairo University - Egypt
Ex-(Software Engineer / ICPC World Finalist)



Data is the new Oil

- If there is no data, there is no machine learning
- Services compete on **collecting data** about you
  - Your location, visits, habits, what you buy, what you watch, for how long
- They **use** these services to develop to affect you more
  - Grap more your attention such as reels/shorts, netflix recommendations, products
  - More interesting social media posts to stay more on the platform
    - Then collect more data
    - Show more **ads** that you most probably click!
- They **sell** your data to others (Data Monetization)!
  - Many services share **broad types** of data about us!
- If you're **not paying** for the product, you are the **product**
  - Nothing really for free :)

Question!

- Nowadays, there are many services that offer very short clips for the users
  - Facebook reels, Youtube shorts, Tiktok shorts
- Assume 2 users uploaded the following videos:
  - Video 1: A woman shows how to use a new brand for Women's Electric **Shaver**
  - Video 2: A woman in a bikini behaving in a sexual way
- Ali and Sarah are **twins** at the age 13. Today is their first existence on the internet and they created accounts on a social media providing basic information like age, address and education
  - Which video will Ali most probably see? What about Sarah?

Question!

- While your mum walking in the home, she noticed 2 Facebook reels with almost naked women (or violence)
- Your mum said these things are on your page because you keep looking for such things!
- Is this a correct judgment?
- If not, what to do to build more correct judgement?

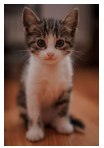
Labeled Data vs Unlabeled data

- **Data** in our context refers to any **raw** input of interest
  - Text, image, video, sound, attributes of [bank user, home, loan,student, etc]
- If we have the have the data and **relevant information** about it, we call this **labeled (annotated) data**
  - We call the label: **ground truth** (GT)
    - It represents our ML algorithms **output**
  - The label must be useful for our task
    - Some input ⇒ output relation
- If we don't have such labels, it is unlabeled data

GT: Cat

GT: Dog

GT: Dog



GT: Cat

Question!

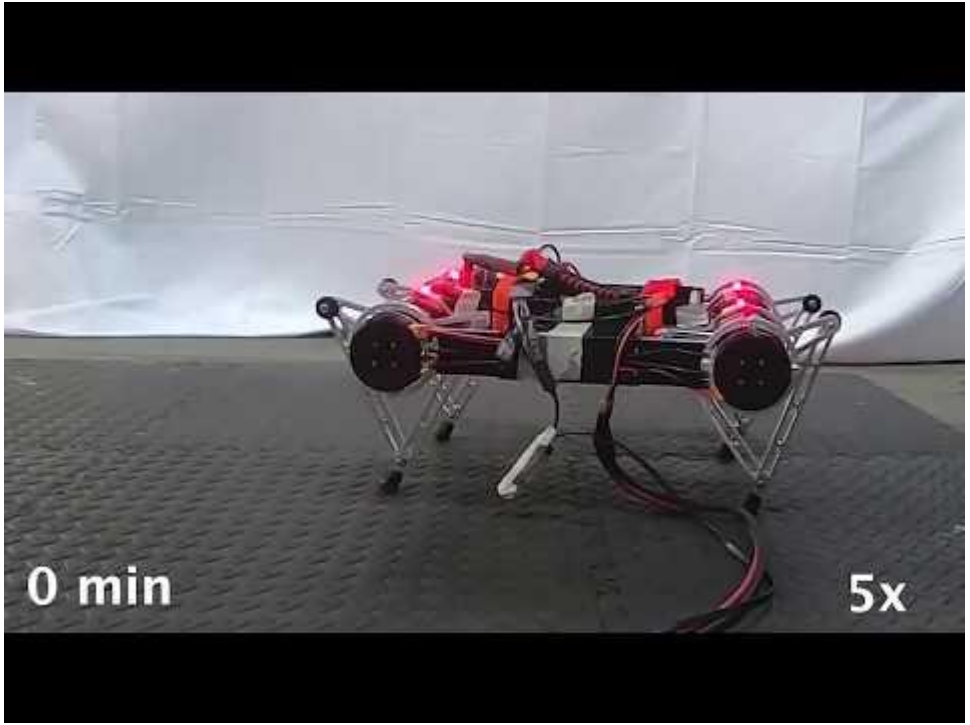
- For the following examples, determine if the data is labeled or not
- 1) We automatically downloaded several random youtube videos
- 2) We crawled the internet and download 100 articles from the following topics: 1) economy, 2) politics, 3) sports
- 3) We downloaded several English articles and asked an Arabic Translator to translate them from English to Arabic
- 4) Our city’s library has 800 face photo of its users and the date of each photo

Machine Learning Approaches

- There are 3 major machine **learning paradigms**
- **Supervised learning (إشراف)**: Uses labeled data
  - Image #1 is cat. Image#2 is dog
  - Images are the input data. Cat/Dog are labels (ground truth)
  - **Most advances** in the industry is based on it
  - *In **weakly-supervised** learning: the labeled data are noisy or not-specific*
- **Unsupervised learning (بدون إشراف)**: Uses unlabeled data
  - We have 1000 images, but we don't know their labels
  - *In **semi-supervised** learning: some data are labeled and some data are not*
  - *In **self-supervised** learning, we build labels from the data structure **itself***
- **Reinforcement learning**: learning by trial and error (interactive feedback)
  - For example, a chess game AI learner plays millions of games versus itself and from the success/failure (feedback), it learns how to make strong moves

Reinforcement Learning: Learning to Walk

- The robot tries to walk forward
- With every trial, it fails, it learn that its decisions where wrong
- Long time ago, stanford has popular video for Autonomous Helicopters

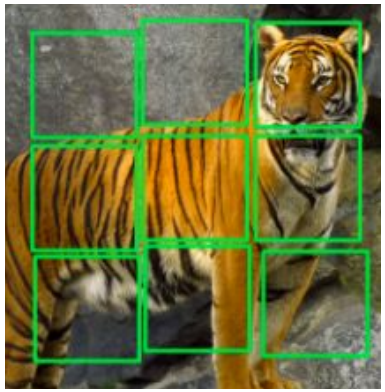


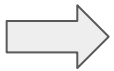
Questions!

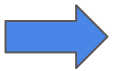
- For each case, identify the machine learning approach
- 1) We have several images downloaded from the internet
- 2) We have several images that may contain several cars. For each image, we know exactly where are the cars in the image (rectangle per car)
- 3) We have several images that may contain several cars. For each image, we know if it has any car in it or not
- 4) We have several images that may contain several cars. For some of them we know exactly where are the cars. For some other images, we know nothing
- 5) We would like our autonomous car to learn **parking** in a simulated program

Questions!

- Given an image *without a label*, we will divide it to 3x3 block
  - Input is a shuffled image and ML learns how to solve the puzzle







Why!

- Why do we have several paradigms for learning from data?
- Because in practice, we can face several scenarios
  - For example, we can have 100 million images from downloaded randomly from the internet. We don’t know what is inside the image
  - It is so costive to ask some people to annotate it
  - Can we do some ML using such data?!

Problem Types

- There are a few **types** of problems ML tackle
- Regression (التوقع)
- Classification (التصنيف)
- Forecasting (التنبؤ المستقبلي)
- Clustering (التجميع)
- Recommendation (التوصية)

Machine Learning Algorithms

- In each learning paradigm, there are different algorithms that are based on this learning approach
- Supervised learning
  - Regression (Linear/Polynomial, **Neural Network,** Decision Trees, Random Forest, **XGBoost**)
  - Classification (Logistic Regression, **Neural Network,** KNN, Naive Bayes, SVM, Decision Trees, Random Forest, **XGBoost**)
  - Recommender (Collaborative filtering, Content-based filtering: can use **Neural Network**)
- Unsupervised learning
  - Clustering (K-means, GMM, **Neural Network**)
  - Dimensionality Reduction (تخفيض الأبعاد)(**Neural Network**, PCA, t-SNE)
  - Generative modeling (Deep **Neural Network**: GAN, VAE)
- Reinforcement learning
  - State of the Arts (SOTA) uses deep **Neural Network**

One more time!

- **Problem type** such as Classification
- **Learning paradigm** such as Supervised Learning (SL)
- **Learning algorithm** such as Neural Network (NN)
- *Others might call all of therm learning algorithms. The same for me*
- *Don’t memorize the names. You will get them when you study them!*

Classification Task

- Classification is a **supervised** learning task where the output is a class label
- Class/**Discrete** labels examples:
  - 0 or 1 [for spam or not]
    - We call it **binary** classification
  - From 0 to 4 corresponding to [cat, dog, person, car, bicycle]
    - We call it **multiclass** classification
- Number of classes typically is limited, e.g. less than 50
  - However, it still can be thouthands in deep learning applications

Classification Task

- Assume we built a **classifier** to **classify** an image to: cat, dog, person, car, or bicycle

ML Classifier

- This is a **dog** (with 80% certainty)
- This is a cat (with 10% certainty)
- This is a person (with 5% certainty)
- This is a car (with 2% certainty)
- This is a bicycle (with 3% certainty)

- The classifier will assign a probability for each output class/category
  - Observe: probabilities above sum to 1 (100%)
- We can select the class with the highest probability, e.g. the dog for this case

Regression Task

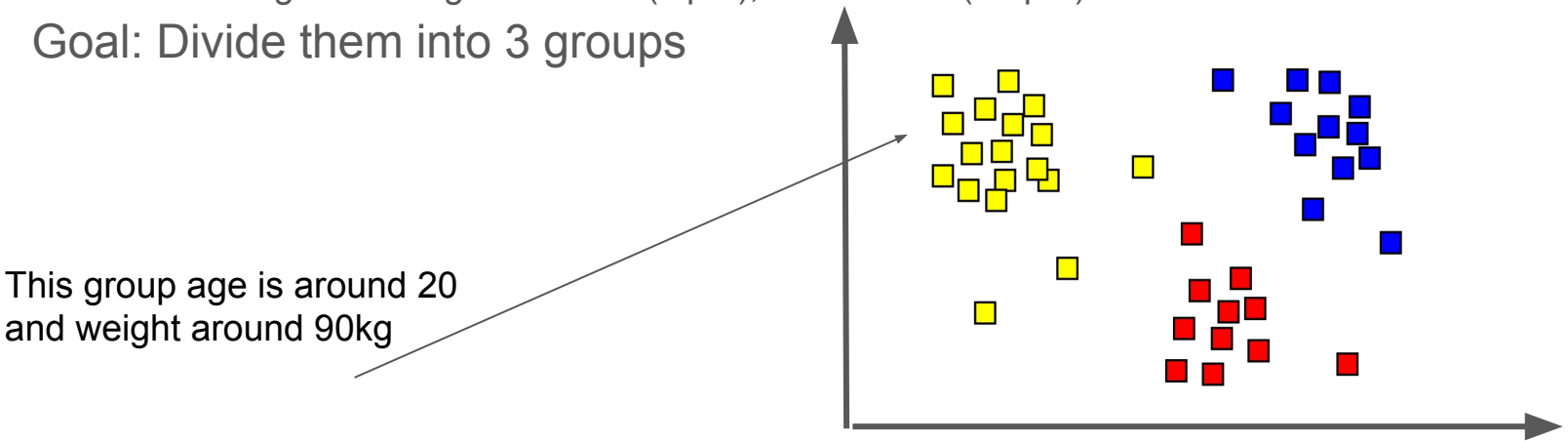
- What if would to predict the housing price? The salary of a person?
  - The range now is huge
  - It is better to think of that as continuous output (vs discrete output in classification)
- Regression is a **supervised** learning task where the output is continuous
  - Typically the outputs are scaled in [0 - 1] range. More later

ML Regressor

Estimated price:
2,500,750 pounds

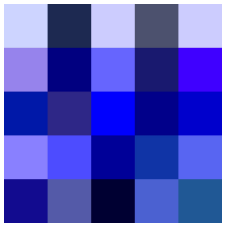
Clustering Task

- Clustering is an **unsupervised** task where we group **similar** data together
  - Each group is called a cluster
  - We need to decide a similarity criteria
- Imagine we have a set of people. For each person we know the age and the height. We can draw/visualize them on X-Y plane
  - We call age and height features (input), NOT labels (output)
- Goal: Divide them into 3 groups

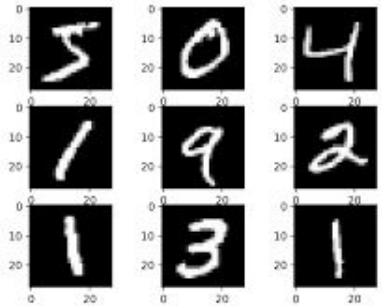


Questions

- What is the **problem type**: classification, regression or clustering
- Given a picture, predict the age of a person
- Predict whether a document is for politics, economy or social studies
- Grouping people in a social network
- Predicting credit approval based on historical data of a person
- Given an image of a **single digit** written by hand. We would like to know the digit value
- Given an image collection of **shades** for 3 colors (Yellow, Blue and Red)
  We would like to separate them into 3 categories.



***Shades** of blue include cyan, navy, turquoise,*
aqua, midnight blue, sky blue, royal blue, and aquamarine



Summary

- Let X be an input (e.g. image) and Y is its relevant output (e.g. cat)
- Dataset: Collection of Pairs of (X, Y): This is a **supervised** problem
  - Is Y a discrete value of limited range? If yes, this is classification task
    - E.g. labels are: 0, 1, 2, 3, 4, ….500
    - Problems: Spam Filter / Breast Tumor Classification: malignant or benign
  - Is Y a continuous/float value? If yes, this is a regression task
    - E.g. labels are: 0.5, -2.5, 1.3, 2.4, and so on
    - Problems: Housing Price Prediction
- Dataset: Collection of X only. This is unsupervised problem
  - More sense later

Studying Machine Learning

- ML has many prerequisites
  - Strong Programming Skills
  - Algorithms, especially Graph Theory + some DP/Greedy
  - Basics of OOP / Inheritance / Operator Overloading for Deep Neural Networks
  - A lot of mathematics: Linear Algebra, Calculus, Probability, Statistics
- The good news:
  - From an algorithm to another, you need only a small subset of these prerequisites
  - Most of the algorithms don’t work well in practice. So take it easy :)
    - However, studying them will build your mentality
  - With a wise roadmap, you can do it!
- It is hard to understand several concepts from the first trial
  - Expect yourself to repeat 2-3 times. Expect yourself to try other resources
  - **Try to prepare before the lecture to boost your overall understanding**

Optional Relevant Resources

- Supervised Learning - Prof Andrew Ng
- Introduction to ML - Prof Hamid Tizhoosh
  - History. Can we measure intelligence?
- Introduction to ML - Prof Eric Grimson

"Acquire knowledge and impart it to the people."

"Seek knowledge from the Cradle to the Grave."