204456: Machine Learning

Ch01 - Basic concepts

Jakramate Bootkrajang

Department of Computer Science
Chiang Mai University

Objectives

- Understand what "learning" in machine learning is
- Know when to applied machine learning algorithms
- Be able to differentiate between three majors learning paradigms

Outlines

- What is learning ?
- What is machine learning ?
- When should we use machine learning ?
- Challenges and Related fields

What is learning ?

- Learning is the process of acquiring new, or modifying existing, [wikipedia]
  - knowledge
  - behaviors, skills
  - preferences
- Learning to recognise orchid flowers
- Learning to group similar objects
- Learning to ride a bicycle

Let's learn something

Try to differentiate holy basil from Thai basil

(a) holy basil (horapa)

(b) Thai basil (kraprao)

Let's test ourself

What are these leaves ?

(a) ??

(b) ??

Let's test ourself (again)

What is this leaf ?

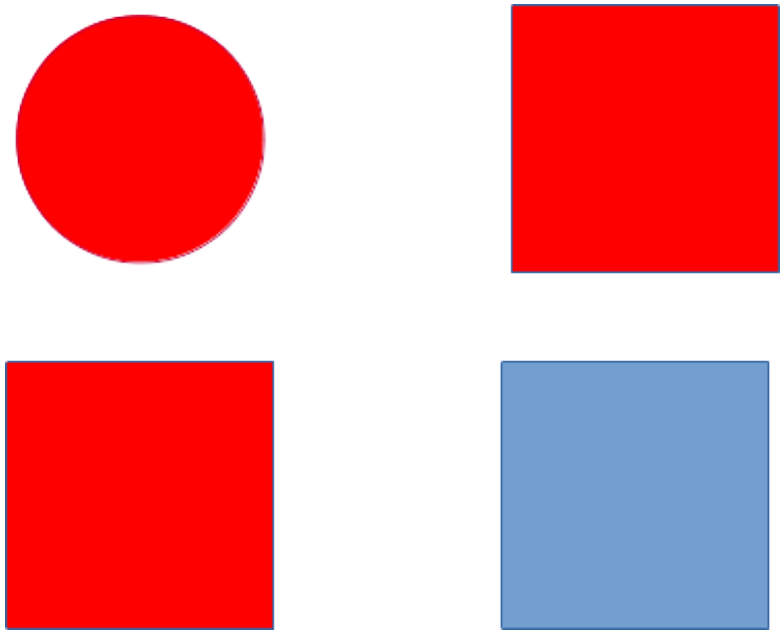
(a) ??

Observations from basil example

- We just did supervised-learning
- Learning is not just remembering but also generalising
- Quality of learning might depends on the quality of training data
  - size
  - correctness

Let's do another task

Organise the following objects into appropriate groups ?

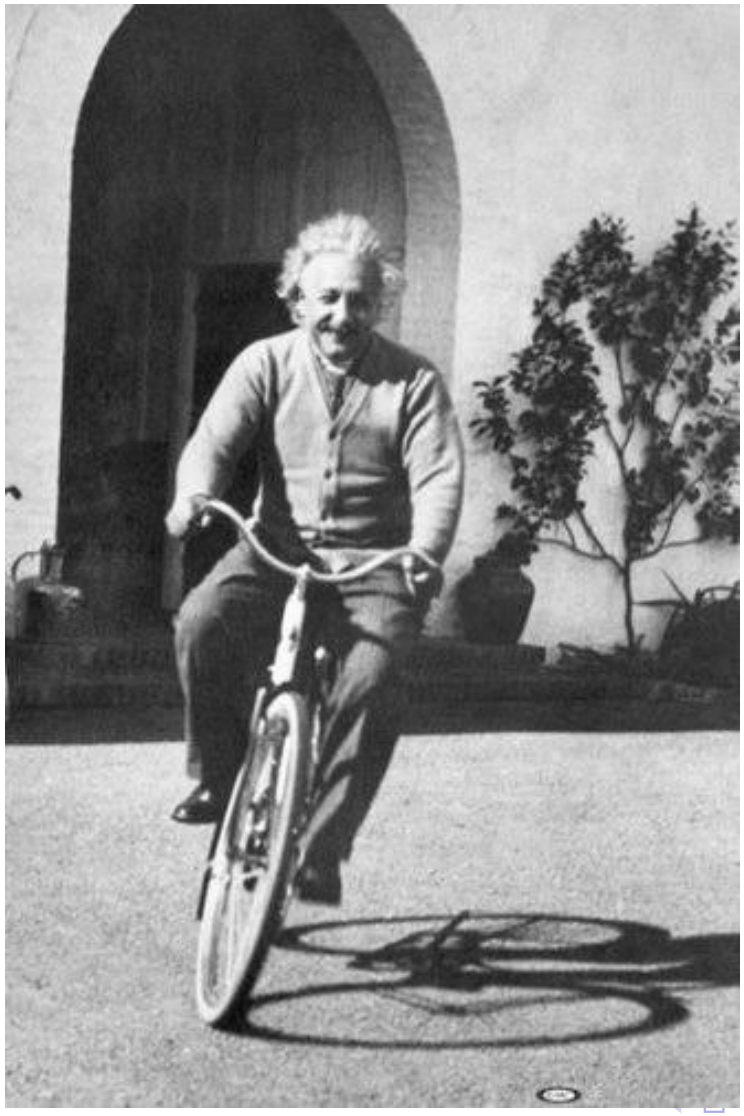


Observations from basil example

- We just did unsupervised-learning
- Results depends on how we define object's properties.
  - color , shape or both color and shape ?
- We will call a property of objects as 'feature'
- The process of defining features is call 'feature extraction'

Yet another example

Let's recall how we learn to ride a bicycle



Observations from bicycle example

- We just did reinforcement-learning
- Nobody can tell you 'precisely' how to ride a bike.
- You need to perform trial and error.
- We you crashed, you know whatever you did before the crashing moment was not right.

From human to machine

- What if we want our computer (machine) to classify basil leaves ?
- Let's tell it what to do step-by-step.
- Turns out that it is not easy to summarise and explicitly program our skills
  - Even if we can the algorithm might be too rigid (does not improve with more data in the future)
- So, let's make the machine learn !!

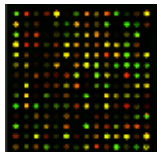
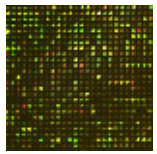
Machine Learning ?

- A study (design and analysis) of algorithms which gains expertise in some specific task using past experience (data)
- Mathematically, ML is to capture a relationship (function) between input data and the desired output.
- Learning is the process of adjusting function's parameters in order to explain the relationship

Input data

Space of inputs: $\mathcal{X}$

$x_1 =$  $x_2 =$  $\dots$

$x_1 =$  $x_2 =$  $\dots$

$$x_1 = \begin{bmatrix} 1 \\ 2 \\ 16 \\ \vdots \end{bmatrix} \quad x_2 = \begin{bmatrix} 0 \\ 5 \\ 19 \\ \vdots \end{bmatrix} \quad \begin{bmatrix} \text{cokes / day} \\ \text{beers / day} \\ \text{BMI} \\ \vdots \end{bmatrix}$$

Output data

Space of outputs: $\mathcal{Y}$

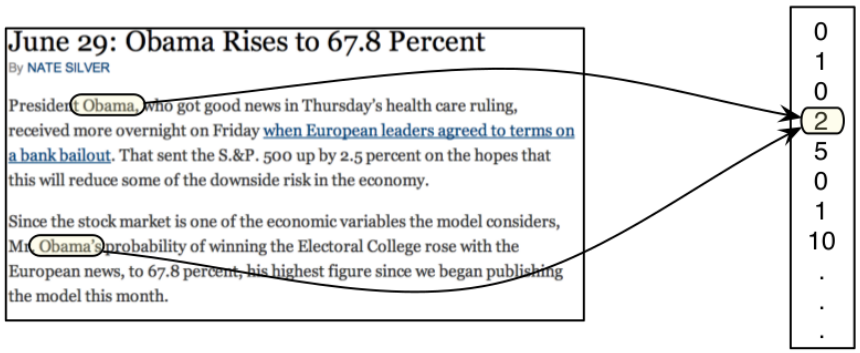
- $y \in \mathcal{Y} = \{0, 1\}$ binary output (0 =‘tiger’ and 1=‘cat’)
- $y \in \mathcal{Y} = \{0, 1, \ldots, K\}$ discrete output
- $y \in \mathcal{Y} = [0, 200]$ real valued output (life expectancy)
- $y \in \mathcal{Y} = \{\text{left}, \text{right}, \text{up}, \text{down}\}$ a set of actions

Paradigms of machine learning

- Supervised learning: learning from labelled data, $\{(x_1, y_1), \ldots, (x_n, y_n)\}$
  - Real valued outputs: regression
  - Discrete valued outputs: classification
- Unsupervised learning: learning from unlabelled data, $\{x_1, \ldots, x_n\}$
- Reinforcement learning: learning from interactions with the world based on awards and penalty. Correct input/output pairs are never presented.

Supervised learning: classification

- Text classification: Spam vs Normal, News topics classification
- Bag-of-words representation.

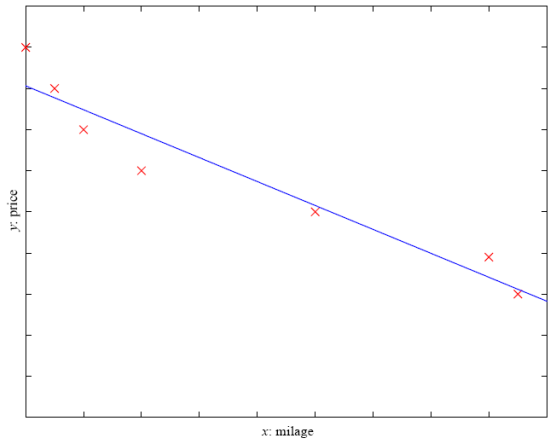


- Let "1" stands for the category "politics", then this example can be represented as

$$\left(\begin{bmatrix} 0 \\ 1 \\ 0 \\ 2 \\ 5 \\ 0 \\ 1 \\ 10 \\ . \\ . \\ . \end{bmatrix}, 1 \right)$$

Regression

- Capturing relationship between mileage and price of a used car
- $y = g(x|\theta)$, x: car's mileage, y: price.
- $g()$ is the model and θ is model's parameter.



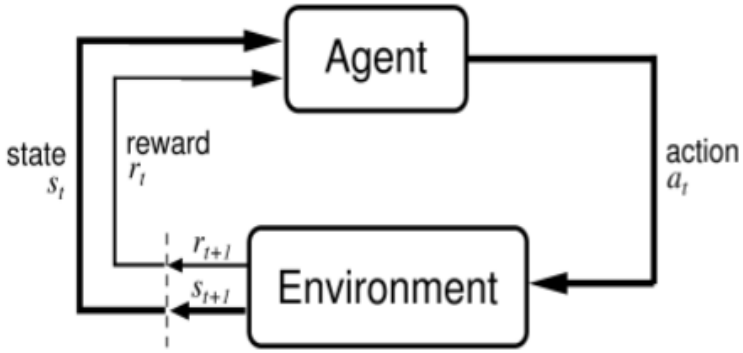
- For navigating a car: Outputs angle of the steering wheel

Unsupervised learning

- Learning "what normally happens"
- Clustering: Grouping similar instances
- Example applications
  - Customer segmentation
  - Image compression: Color quantization

Reinforcement learning

- Learning a policy: A mapping from situation (state) to action.
- Reward action which leads to good state.
- Punish action which leads to bad state.



  - Many control related applications
  - Game playing
  - Robot in a maze

Machine learning steps in general

1. Know what you want to do. Understand your data. (Objective)
2. Build a model that is a good and useful approximation to the data. (Modelling)
3. Devise an algorithm to learn the model: how to adjust model's parameters. (Learning)
4. Test your model using existing data or new unseen data. (Performance measure)
5. Theoretically show that your model will work on any new data of the same kind.

When to use ML ?

- Learning is useful when:
  - Humans are unable to explain their expertise (speech recognition)
  - Human expertise does not exist (Fraud detection)
  - Solution changes in time (online learning, objective function changes)
- Polynomial time problems are better solved using standard algorithms.

Challenges

- Structure of data is often unknown or poorly understood.
  - What is the best model for the data ?
- Data abnomalities
  - Class imbalanceness
  - Noisy feature, noisy label or both
  - Missing or incomplete data
- High dimensionality: e.g., gene expression profiles
- Scaling of algorithms to massive data sets

Machine learning and related fields

- A.I.: Artificial intelligence
- Machine learning: Focuses on learning, one way toward A.I.
- Pattern recognition: ML-based approaches, syntactic PatRec (regular expression).
- Data mining: Originated from database people, focus on extracting useful information from large databases.
- In the end, it all depends on profession and background of user.

Lastly: you should be able to answer these

- What is machine learning ?
- When to applied machine learning algorithms ?
- What are three majors learning paradigms ? How do they differ from each other ?