EE2211 Tutorial 1

Dr Feng LIN

Email: feng_lin@nus.edu.sg

01

Q1

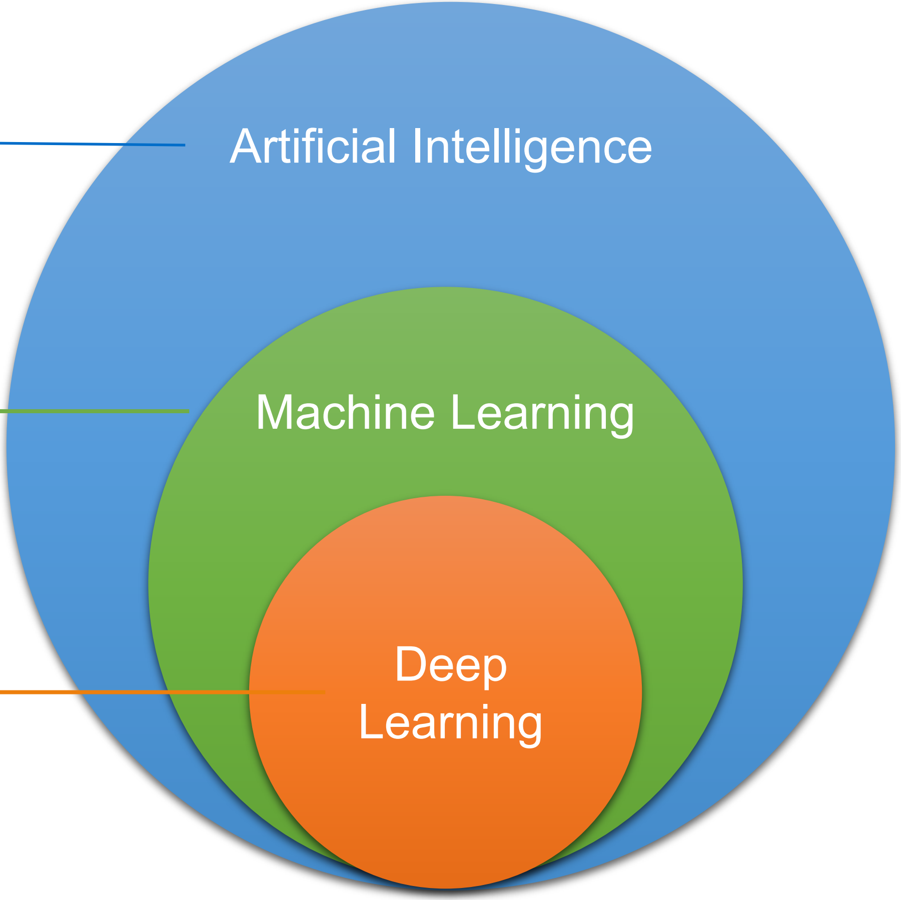
- What is the difference between ML (Machine Learning) and AI (Artificial Intelligence)?

01 Q1

Program with the ability to learn and reason **like humans**

Algorithms with the ability to learn without being explicitly programmed

Subset of machine learning in which artificial neural networks adapt and learn from vast amount of data



Q1

- Artificial Intelligence is the broader concept of machines being able to carry out tasks in a way that we would consider “smart”.
- Machine Learning is a current application of AI based around the idea that we should really just be able to give machines access to data and let them learn for themselves.
- https://www.forbes.com/sites/bernardmarr/2016/12/06/what-is-the-difference-between-artificial-intelligence-and-machine-learning/# 741adc6b2742

02 Q2

Which of the following is the most reasonable definition of machine learning?

a) Machine learning is the field of allowing robots to act intelligently.

b) Machine learning is the science of programming computers.

c) Machine learning only learn from unlabeled data.

Unsupervised learning

d) Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.

Ans: d)

03

Q3

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. Suppose we feed a learning algorithm a lot of historical weather data, and have it learn to predict weather. In this setting what is T?

a) The historical weather data.

b) The probability of it correctly predicting a future data's weather.

c) The weather prediction task.

d) None of these.

Q3

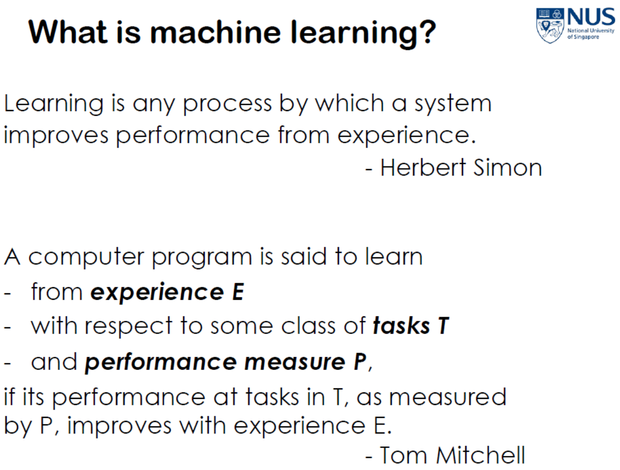
T	P	E
Task	Performance	Experience
Weather prediction	Accuracy of prediction	• Historical data • Weather record

$$l(\text{actual railfall}, \quad \text{predicted rainfall})$$

$$= (\text{actual} \ - \text{predicted})^2$$

$(99mm \ - 100mm)^2$ Small loss

$(0mm \ - 100mm)^2$ Large loss



What is machine learning?

Learning is any process by which a system improves performance from experience.

- Herbert Simon

A computer program is said to learn

- from ***experience E***
- with respect to some class of ***tasks T***
- and ***performance measure P***,

if its performance at tasks in T, as measured by P, improves with experience E.

- Tom Mitchell

04

Q4

Suppose you are working on weather prediction and use a learning algorithm to predict tomorrow's temperature (in degrees Centigrade/Fahrenheit).

(i) Would you treat this as a classification or a regression problem?

(a) Regression.

(b) Classification.

(c) Clustering.

(d) None of these.

(ii) What kind of data should you gather?

Q4

(i) Predict tomorrow’s temperature,

The output temperature measured by the thermometer is a real number, with an infinite label space, making it a definite regression problem.

Ans: a)

Q4

(ii) Quantitative data

x_i:	x_{i1}:	**Pressure**: Atmospheric pressure readings.
	x_{i2}:	**Humidity**: Historical humidity levels
	x_{i3}:	**Cloud Cover:** Types and amounts of clouds.
	$x_{i4:}$	**Solar Radiation:** The amount of sunlight received.
y_i		**Temperature**

05

Q5

You want to develop learning algorithms to address each of the following two problems.

P1: You'd like the software to examine your email accounts, and decide whether each email is a spam or not.

P2: You have a large quantity of green tea (e.g., 1000kg) with a record of previous sales. You want to predict how much of it will sell over the next 6 months.

Should you treat these as classification or as regression problems?

(a) Treat both P1, P2 → regression problems.

(b) Treat both P1, P2 → classification problems.

(c) Treat P1 → regression problem, P2 → classification problem.

(d) Treat P1 → classification problem, P2 → regression problem.

Q5

P1: Spam detection is a classification problem

P2: Label space of sell is real number, making it a regression problem

Ans: d)

06 Q6

Suppose you are working on stock market prediction. Typically tens of millions of shares of a company’s stock are traded each day. You would like to predict the number of shares that will be traded tomorrow.

(i) Would you treat this as a classification or a regression problem?

(a) Regression.

(b) Classification.

(c) Clustering.

(d) None of these.

(ii) If the data you have collected involved millions of attributes, what would you do?

Q6

(i) Predicting the number of shares involves real numbers, making it a regression problem.

Ans: a)

Q6

features

(ii) If the data you have collected involved millions of attributes, what would you do?

Extract relevant features

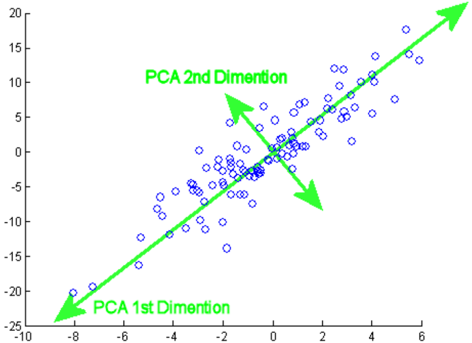
- Remove Redundant Features: Identify and remove features that are highly correlated with each other or provide little information gain.

A small dataset with the following features:

1. Height (in m)
2. Weight (in kg)
3. BMI (Body Mass Index)

$$\text{BMI} = \frac{\text{Weight (kg)}}{\left(\text{Height (m)}\right)^2}$$

- Principal Component Analysis (PCA): Reduce the dimensionality of the data by transforming it into a set of uncorrelated variables (principal components) that capture the most variance.



07

Q7

Some of the problems below are best addressed using a supervised learning algorithm, and the others with an unsupervised learning algorithm. Which of the following would you apply supervised learning to? (Select all that apply) Assume some appropriate dataset is available for your algorithm to learn from.

(a) Determine whether there are vocals (i.e., a human voice singing) in each audio clip extracted from a piece of music, or it is a clip of only musical instruments and no vocals.

(b) Given data on how 1000 medical patients respond to an experimental drug (such as effectiveness of the treatment, side effects, etc.), discover whether there are different categories or “types” of patients in terms of how they respond to the drug, and if so what these categories are.

(c) Given a large dataset of medical records of patients suffering from heart disease, try to learn whether there might be different clusters of such patients for which we might tailor separate treatments.

(d) Given a set of data which contains the diet and the occurrence of diabetes from a population over a 10-year period. Predict the odds of a person developing diabetes over the next 10 years.

Q7

(a) Deterring it there are vocals

Label space = {vocals, no vocals}, classification

(b) and (c)

$\mathrm{x}_{i,}, \ldots, \mathrm{x}_n$

y_i is not available

Clustering problem, Unsupervised learning

Q7

(d) x_i: diet

y_i: probability of a person developing diabetes over the next 10 years

$$\mathrm{x}_{new} \rightarrow y_{new}$$

Regression problem → supervised learning

Ans: (a) and (d)

08

Q8

Suppose you are working on a machine learning algorithm to predict if a patient is COVID-19 infected according to the patient’s particulars such as age and health conditions, symptomatic data, such as fever, dry cough, tiredness, aches and pains, sore throat, diarrhoea, conjunctivitis, and headache etc. What are the Task, Performance, and Experience involved according to the definition of machine learning?

T	P	E
Task	Performance	Experience
Patient classification into {‘infected’, ‘uninfected’}	Accuracy of classification	Age and health condition and Patient’s symptomatic data

09

Q9

We use labelled data for supervised learning, where the labels are used as the desired target of prediction for classifiers. Which of the next data are the useful labelled data?

(a) To build an image object classifier to discriminate between apple and orange, we have many fruit images labelled with the country of origin.

(b) To build a system to predict the number of COVID cases for tomorrow given the past daily record, we have a collection of daily data for a period of 12 months.

(c) To build a classifier to automatically evaluate student essays, we have collected a set of student essays that have not been graded by teachers.

Q9

Ans:

(a) The useful fruit images should be labelled with apple or orange. Country of origin doesn’t tell apple or orange. Therefore, the data is not useful.

(b) We can use n days of historical data as the input, and n+1th day’s data as the target. This dataset is useful.

(c) The useful dataset should include student essays and the grades. Student essays are the input, and the grades are the desired target of prediction. This dataset is not useful.

Ans: (b)

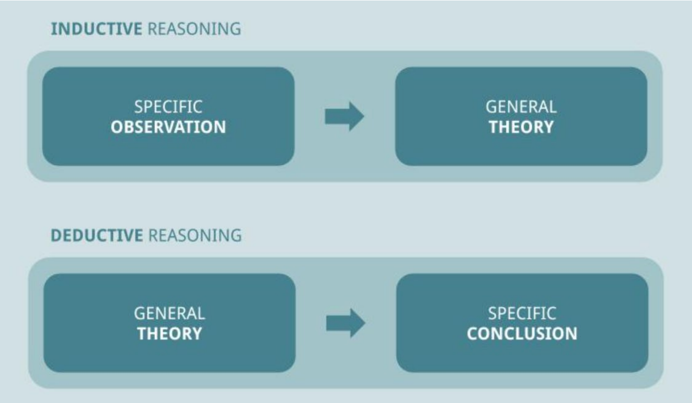
10

Q10

Determine whether each of the following is “inductive” or “deductive” reasoning?

(a) The first coin I pulled from the bag is a penny. The second and the third coins from the bag are also pennies. Therefore, all the coins in the bag are pennies.

(b) All men are mortal. Harold is a man. Therefore, Harold is mortal.



Q10

(a) This statement cannot be true; there is still uncertainty involved. This is an inductive reasoning approach.

(b) It is deterministic; there is no uncertainty in the statement. This is deductive reasoning.

Ans: (a) inductive; (b) deductive

Q11

Find a problem of your interest and formulate it as a machine learning problem. List out the input features and output response and provide your choice regarding the types of learning (such as supervised or unsupervised learning,

Some Python Resource

- Installing scikit-learn (Ref: [Book2] Andreas C. Muller and Sarah Guido, “Introduction to Machine Learning with Python: A Guide for Data Scientists”, O’Reilly Media, Inc., 2017)
- scikit-learn depends on two other Python packages, NumPy and SciPy. For plotting and interactive development, you should also install matplotlib, IPython, and the Jupyter Notebook. We recommend using the following prepackaged Python distributions, which provides the necessary packages:
- *Anaconda*
  - ❖A Python distribution made for large-scale data processing, predictive analytics, and scientific computing.
  - ❖Anaconda comes with NumPy, SciPy, matplotlib, pandas, IPython, Jupyter Notebook, and scikit-learn. Available on Mac OS, Windows, and Linux, it is a very convenient solution and is the one we suggest for people without an existing installation of the scientific Python packages. Anaconda now also includes the commercial Intel
  - ❖MKL library for free. Using MKL (which is done automatically when Anaconda is installed) can give significant speed improvements for many algorithms in scikit-learn.

Some tutorials that might be useful:

- A quick start tutorial on NumPy: https://numpy.org/devdocs/user/quickstart.html
- Some community tutorials on Pandas: https://pandas.pydata.org/pandasdocs/stable/getting_started/tutorials.html
- Scikit-learn tutorials: https://scikit-learn.org/stable/tutorial/index.html

  https://scikit-learn.org/1.4/tutorial/index.html

  https://scikit-learn.org/stable/
- Python Numpy Tutorial (with Jupyter and Colab): https://cs231n.github.io/python-numpy-tutorial/#jupyter-and-colab-notebooks

THANK YOU