# Chapter 1

# Intro to Machine Learning

What is machine learning? We have used it in many applications nowadays, such as:

- perform web search (search per se, and auto completion your search input)
- recognize human speech (speech recognition, natural language processing)
- diagnose diseases from X-rays (computer vision)
- build a self-driving car (computer vision)
- advertising (recommender system)
- AI for augmented reality
- combating payment fraud (fraud detection)
- inspect and recognize assembly line manufacturing defects (computer vision, etc)
- large-scale agricultural applications
- climate change problem to optimize wind turbine power generation

According to a 2020 study by McKinsey (The state of AI in 2020), AI and machine learning is estimated to create an additional 13 trillion $ dollars of value annually by the year 2030.

We see two formal definitions for machine learning here:

**Definition 1.1** (Machine Learning)**.** *Machine Learning is a field of study that gives computers the ability to learn without being explicitly programmed. —– by Arthur Samuel (1959).*

**Definition 1.2** (Well-posed Learning Problem)**.** *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. —– by Tom Mitchell (1998).*

**Example 1.1** (Recognize T, E, P)**.** *Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to better filter spam. What is the task T, E, P in this setting?*

- Classifying emails as spam or not spam. –> T
- Watching you label emails as spam or not spam. –> E
- The number (or fraction) of emails correctly classified as spam/not spam. –> P

There are **two** basic categories in machine learning algorithms:

1. Supervised Learning
2. Unsupervised Learning

Also some other machine learning algorithms include: reinforcement learning, recommender system, etc.

## 1.1 Supervised Learning

Supervised Learning is when "right answers" are given. Depend on continuous or discrete output, there are two corresponding methodologies: **Regression** (continuous output, or lines) and **Classification** (discrete output, or categories).

- Regression is for predicting continuous valued output. In a house value prediction example, size of the houses in sqft is the feature, and prices of the training data are given as right answers.
- Classification is for discrete valued output, e.g. 0 or 1. In a breast cancer prediction example, the tumor can be either classified as malignant or benign according to its size.

**Example 1.2** (Regression v. Classification Problem)**.** *How to categorize the following two problems?*

- *Problem 1: You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months.*
- *Problem 2: You'd like software to examine individual customer accounts, and for each account decide if it has been hacked/compromised.*

Answer: Problem 1 can be Regression problem, and problem 2 can be Classification problem.

## 1.2 Unsupervised Learning

Unsupervised Learning is when you need to find certain patterns in the data that its value or right answer is not provided. Some example of unsupervised learning including: email spam folder, news categorization, organizing computer clusters, social network analysis, market segmentation, astronomical data analysis.

**Example 1.3** (Cocktail Party Problem)**.** *Two speakers in a cocktail party, each closer to one of the two microphones. We can use unsupervised learning algorithm to find two different sound patterns and thus filtered both speakers' sound.*

The one line algorithm is as follows:

**Algorithm 1** Cocktail Party Problem Algorithm

```
[W,s,v] = svd ((repmat (sum (x.*x,1),size (x,1),1).*x)*x');
```

**Example 1.4** (Supervised v. Unsupervised Learning)**.** *How would you categorize the following problems? Supervised or Unsupervised learning?*

- Given email labeled as spam/not spam, learn a spam filter.
  *A: Supervised learning: Classification (since you are given labeled, which means have "right answer")*
- Given a set of news articles found on the web, group them into set of articles about the same story.
  *A: Unsupervised learning: find patterns and no right answers are given.*
- Given a database of customer data, automatically discover market segments and group customers into different market segments.
  *A: Unsupervised learning.*
- Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as having diabetes or not.
  *A: Supervised learning: classification.*

## 1.3 Jupyter Notebook and Python

Print statements will generally use the python `f` string style for efficiency, here's an example:

```
# f string print statement example
x = 'f string print'
print(f'f strings allow to embed variables {x}.')
```

which will generate the following output:

```
f strings allow to embed variables f string print.
```