1.1 What is Machine Learning?

Arthur Samuel (1959): "Field of study that gives computers the ability to learn without being explicitly programmed.

Tom Michel (1999): "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."

The checkers example, E = 10000s games, T is playing checkers, P if you win or not.

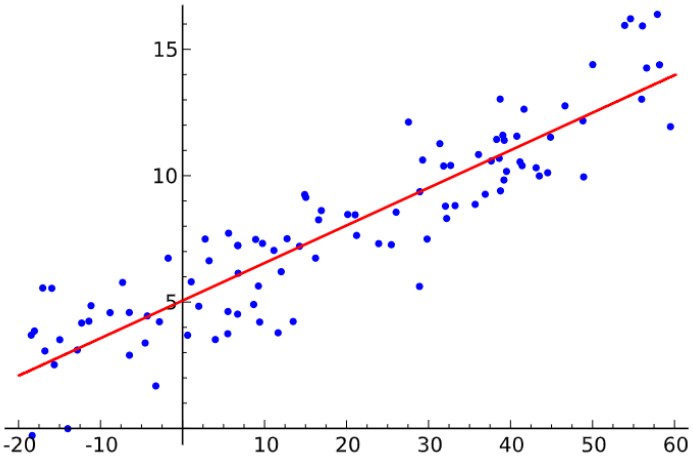
1.2 Several types of learning algorithms

1.2.1 Supervised learning

You first have a data set and **know what the correct output is**. You can consider it as supervised learning if its corresponding dataset has the exact labels. The supervised learning problem is classified into a Regression problem and a Classification problem.

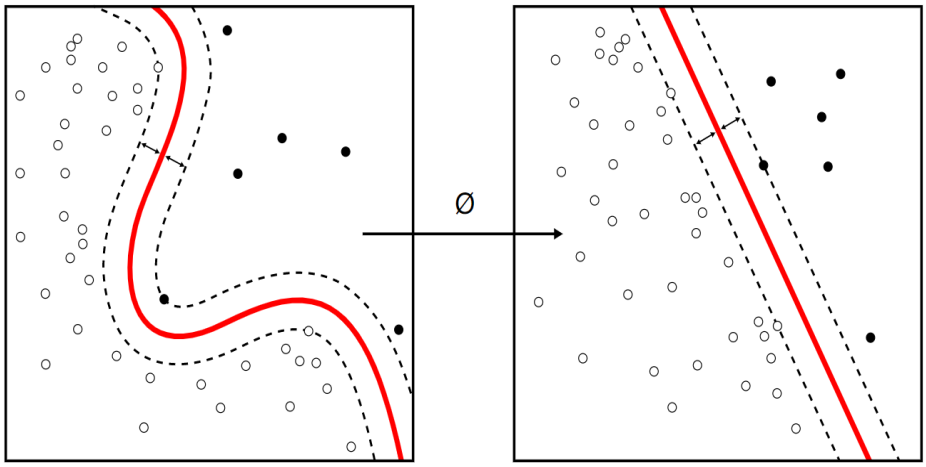
1.2.1.1 Regression problem

In a regression problem, we try to predict the outcome in a continuous output, which means we try to map the input variable to a continuous function. For example, given a photo of a person, predicting age based on the photo, this is a regression problem.



1.2.1.2 Classification problem

In the classification problem, we try to predict the outcome in the discrete output which means we try to map input variables into discrete categories. For example, given to a patient with a tumor, we have to predict whether the tumor is malignant or benign.

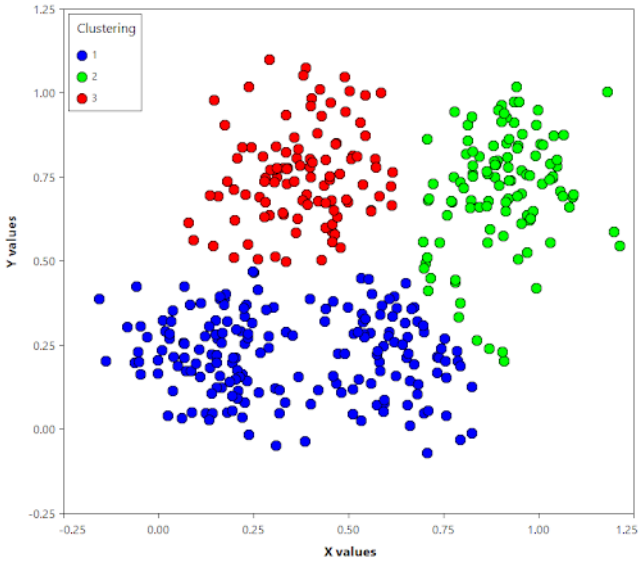


1.2.2 Unsupervised learning

Unsupervised learning allows us to **have little or no idea what our results should look like**. You can consider it as supervised learning if its corresponding dataset has the exact labels. In the unsupervised learning, there is no feedback based on predicted results. Unsupervised learning can be divided into "clustering" and "non-clustering".

1.2.2.1 Clustering

Clustering: taking a collection of 1,000,000 different genes and finding a way to automatically group them into similar or related groups of different variables. In this situation, you have no idea about labels in the dataset.



1.2.2.2 Non-clustering

"cocktail party algorithm" that allows you to find results in a chaotic environment. (that is, identifying individual voices and music from the mixed sound at cocktail parties).

1.2.3 Reinforcement learning

.....

1.2.4 Recommender systems

.....

1.3 Review

Question 1:
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 are E, P, T?
Answer:
T: The weather prediction task.
P: The probability of it correctly predicting a future date's weather.
E: The process of the algorithm examining a large amount of historical weather data.

Question 2:
Suppose you are working on weather prediction, and you would like to predict whether or not it will be raining at 5 p.m. tomorrow. You want to use a learning algorithm for this. Would you treat this as a classification or a regression problem?
Answer:
Classification probelm, because algorithm has to predict whether or not it will be raining at 5 p.m. tomorrow and show the exact and discrete outcome, like raining or not raining (Dichotomous).

Question 3:
Suppose you are working on stock market prediction, and you would like to predict the price of a particular stock tomorrow (measured in dollars). You want to use a learning algorithm for this. Would you treat this as a classification or a regression problem?
Answer:

Regression problem, because algorithm has to predict the price of a particular stock tomorrow and show the exact and continuous outcome, like $7, $8.

Question 4:
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.) In each case, assume some appropriate dataset is available for your algorithm to learn from.
Answer:

[1] Take a collection of 1000 essays written on the US Economy, and find a way to automatically **group these essays into a small number of groups** of essays that are somehow "similar" or "related".
   a. 无监督的学习/聚类问题（类似于讲座中的 Google 新闻示例）
[2] Given a large dataset of medical records from patients suffering from heart disease, try to learn whether there might be **different clusters** of such patients for which we might tailor separate treatements.
   a. 无监督学习，聚类问题，将患者分组到不同的群中。
[3] Given genetic (DNA) data from a person, **predict the odds** of him/her developing diabetes over the next 10 years.
   a. 监督学习，回归问题，从不同人的遗传数据的标记数据集中学习，并输出患糖尿病几率。
[4] Given 50 articles written by male authors, and 50 articles written by female authors, learn to **predict the gender** of a new manuscript's author (when the identity of this author is unknown).
   a. 监督学习，分类问题，从标记数据中学习以预测性别。
[5] In farming, given data on crop yields over the last 50 years, learn to **predict next year's crop yields**.
   a. 监督学习，回归问题，从历史数据中学习（用历史作物产量标记）以预测未来作物产量。
[6] Examine a large collection of emails that are known to be spam email, to discover if there are **sub-types of spam mail**.
   a. 无监督学习，聚类问题，以将垃圾邮件聚类为子类型。
[7] Examine a web page, and **classify** whether the content on the web page should be considered "child friendly" (e.g., non-pornographic, etc.) or "adult."
   a. 监督学习，分类问题，从已被标记为“儿童友好”或“成人”的网页数据集中学习。
[8] Examine the statistics of two football teams, and predicting which team will win tomorrow's match (given historical data of teams' wins/losses to learn from).
   a. 监督学习，分类问题，从历史记录中学习如何进行赢/输预测。

Question 5：
Which of these is a reasonable definition of machine learning?
Answer:
Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.