Machine Learning - Michaelmas Term 2017
Lecture 1 : Introduction to Machine Learning

Lecturers: Christoph Haase & Varun Kanade

1 What is machine learning?

Machine learning techniques lie at the heart of many technological applications that we use daily. When using a digital camera, the boxes that appear around faces are produced using a machine learning algorithm. When websites such as BBC iplayer or Netflix suggest what a user might like to watch next, they are using machine learning algorithms to provide these recommendations.

Broadly speaking, "*the goal of machine learning is to develop methods that can automatically detect patterns in data, and then to use the uncovered patterns to predict future data or other outcomes of interest* " (see Murphy (2012, page xxvii)). Machine learning means very different things to different people. Rather than restrict ourselves to any particular definition, we'll look at a few different definitions and see how they fit the applications considered.

Mitchell (1997) defines what it means for a computer program to *learn* as follows:

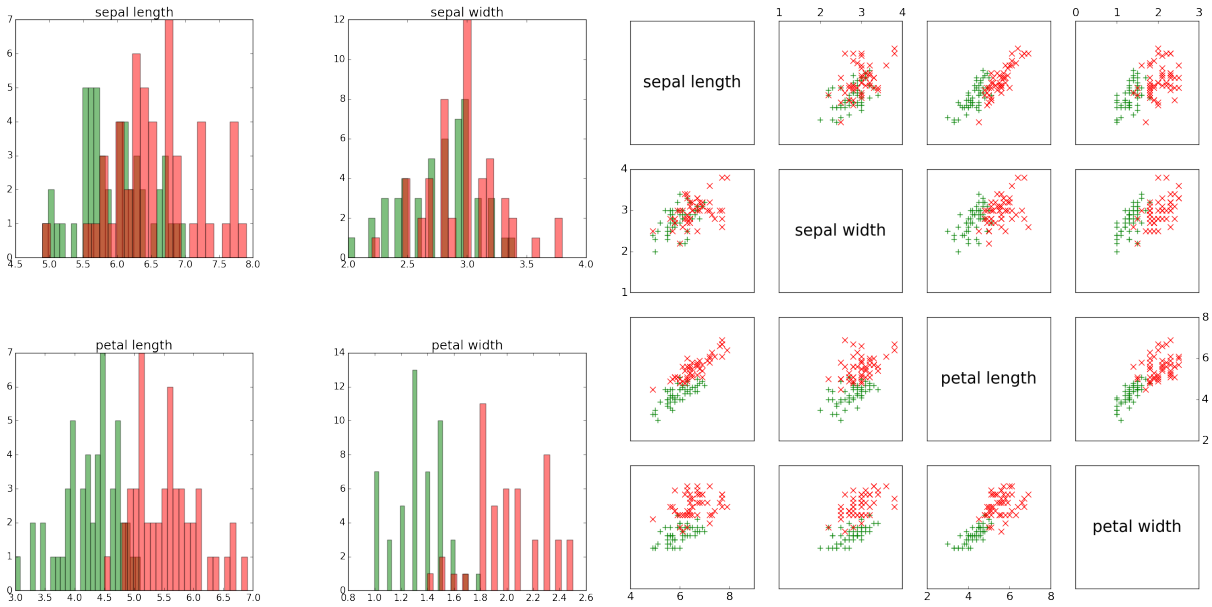
Definition 1 ((Mitchell, 1997))**.** *A computer program is said to* learn *from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.*

In the case of face detection, what would constitute E, T and P? For experiences E, the computer program needs to be input several images containing faces, with the faces clearly marked by bounding boxes. The task T is: given a new image, identify faces by putting boxes around them. Finally, the measure of performance P could be the fraction of faces correctly identified. One may want to consider more stringent notions, by measuring what percentage of each face lies inside the box and outside and so on. Thus, according to Mitchell's definition we would say that a computer program *learns* "face detection", if its performance improves if it is provided with more experience, *i.e.,* more images with faces marked by bounding boxes.

If having a precise definition of machine learning is difficult, it is even harder in the case of *artificial intelligence*. It is increasingly the view that learning will have to form an integral component of any truly intelligent system. For much of the past 60 years, research in artificial intelligence has focused on tasks such as planning, reasoning, deduction, *etc.* Machine learning, which emerged partially as a sub-area of AI, but also through other disciplines outside computer science, such as engineering, control, statistics and neuroscience, often took the more practical view of designing systems that were useful, whether or not they could be deemed intelligent. However, it is worth noting that in arguably the first paper on AI, Turing (1950) had already pointed out that learning would have to be an integral component of any intelligent system– "*Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain*".

2 A brief history of machine learning

Machine learning as a modern discipline grew out of several different disciplines, notably, statistics, computer science, and neuroscience. This is by no means a complete list—researchers in optimisation, decision theory, information theory, game theory, *etc.* have made and continue to make contributions to the general area of machine learning.



(a) (b)

Figure 1: (a) Histogram plots of different measurements for versicolour and virginica varieties. (b) Scatter plots of pairwise measurements of versicolour and virginica varieties.

2.1 Fisher and the Iris Dataset

Arguably, one of the first instances of an automatic classification program was due to the statistician and biologist Ronald A. Fisher.[1] Fisher (1936) was given data about three kinds of iris flowers—setosa, versicolour and virginica.[2] The data consisted of 4 measurements for 50 flowers of each type—the sepal width and length and the petal width and length. His proposal was to look at a linear function of the four measurements that best discriminates the types. This method may seem quite simple, but it is quite powerful and used in some applications to this date (it is known by the name of Fisher Linear Discriminant Analysis[3]). Also remember that 1936 is at least a decade before the first computers were built!

Figure 1 shows histogram plots of individual measurements (Fig. 1(a)) and scatter plots of pairwise measurements (Fig. 1(b)) for two types of flowers—versicolour and virginica. Looking at the scatterplots one can see how using a linear combination of multiple measurements may provide a better solution than any individual measurement.

2.2 Rosenblatt and the Perceptron

Frank Rosenblatt constructed an electronic device called the Perceptron, which was inspired by biological principles.[4] In machine learning terms, a perceptron is simply viewed as taking multiple inputs, and if the weighted sum of these exceeds a certain threshold, the output is 1, otherwise it is 0 (see Fig. 2). One of the remarkable aspects of the perceptron was that Rosenblatt also designed an algorithm that would adjust the weights if on a new input the prediction was incorrect. Under certain assumptions, it can be shown that this algorithm converges to the correct classifier.[5]

While we won't discuss perceptrons much in this course, we will return to them in the form

[1] `https://en.wikipedia.org/wiki/Ronald_Fisher`
[2] `https://en.wikipedia.org/wiki/Iris_flower_data_set`
[3] `http://www.ics.uci.edu/~welling/classnotes/papers_class/Fisher-LDA.pdf`
[4] `https://en.wikipedia.org/wiki/Frank_Rosenblatt`
[5] See `http://www.cs.columbia.edu/~mcollins/courses/6998-2012/notes/perc.converge.pdf`

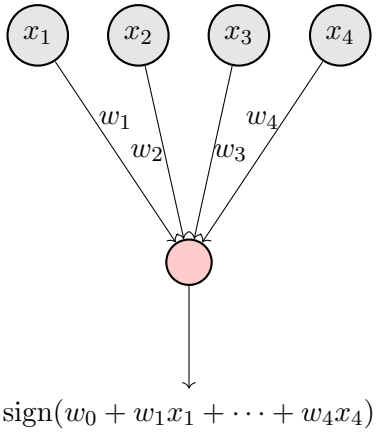


Figure 2: A perceptron with four inputs

of more general artificial neurons, where rather than just thresholding to 0 or 1, the output will be a continuous function of the weighted sum of the inputs. A "deep neural network" is nothing but many artificial neurons stacked together in layers.

3 Machine Learning Paradigms through Applications

3.1 Supervised Learning

In supervised learning, the training data consists of input and output pairs. The goal of the learning algorithm is to *learn* a map from input to output. We look at the examples discussed in the lecture.

Boston Housing Dataset

The data consists of several attributes of a house.[6] Some of these are numerical such as area, crime rate in the neighbourhood, age of the building; others are categorical, *e.g.,* whether or not the house is on the river. The goal is to predict the house price using the attributes.

When the output of the learning algorithm is a real value, as in this case, this is called *regression.*

Imagenet Object Detection

The data consists of images with bounding boxes around objects and the labels for the objects.[7]. The task is to put bounding boxes around objects in certain categories and to label them.

When the output of the learning algorithm is a category, this is called *classification.* In this case, there are multiple categories and it is referred to as *multi-class* classification. When there are only two categories it is referred to as *binary* classification. In fact, in the case of detecting objects in an image, there may be several objects appearing in the same image. This is sometimes referred to as *multi-label multi-class* classification.

3.2 Unsupervised Learning

In unsupervised learning, we are simply given data without any outputs; the goal is usually to find some interesting structure in the dataset.

[6]The actual dataset is available to explore at `https://archive.ics.uci.edu/ml/datasets/Housing`
[7]`http://image-net.org/`

Movie / User	Alice	Bob	Charlie	Dean	Eve
The Shawshank Redemption	7	9	9	5	2
The Godfather	3	?	10	4	3
The Dark Knight	5	9	?	6	?
Pulp Fiction	?	5	?	?	10
Schindler's List	?	6	?	9	?

Figure 3: A synthetic table of some user ratings for some movies

Clustering Genetic Data

Novembre et al. (2008) look at genetic data of around 3,000 European individuals. They find that this data can be clustered remarkably well in groups that correspond to European geography. Although, they had the information about the individuals' nationalities, the learning algorithm was only given genetic data and asked to detect clusters. *Clustering* is a canonical example of unsupervised learning. Another one, which also appears in the work of Novembre et al. (2008), is *dimensionality reduction*. Reducing the dimension of the data is useful both for learning and for visualisation.

3.3 Other Paradigms

We briefly discuss a few other paradigms that don't neatly fit into either supervised or unsupervised learning, and in some cases are entirely different. We will not discuss these much in this course; however, they are important in many applications in modern machine learning.

3.3.1 Active Learning

In active learning, the task is to predict outputs corresponding to individual instances (as in supervised learning), *e.g.,* labelling an image. However, initially all the data is unlabelled. In active learning, the goal is to design algorithms that try to identify the most useful data which can then be labelled by humans as needed. The need to constantly have a human in the loop is what makes this less practical in many settings. However, many applications do try to get "noisy feedback" from human users, by randomising orders of search results, *etc.*

3.3.2 Semi-Supervised Learning

In semi-supervised learning, the goal is to design learning algorithms that make use of large quantities of unlabelled data that may often be available in conjunction with some labelled data. For example, there are large datasets of images with labels available, but the number of unlabelled images vastly outnumbers them. The question is can this additional unlabelled data be exploited to improve the performance of learning algorithms.

3.3.3 Collaborative Filtering

This is best explained by the example of recommender systems. As discussed in the lectures, a matrix with entries containing ratings of movies by users is quite sparse (Fig. 3). Most users will have seen only a small fraction of the movies, although all but a few movies will be seen by a substantial number of users. The goal is to complete this matrix accurately so that useful predictions can be made to users.

3.3.4 Reinforcement Learning

Reinforcement learning (RL) differs from other paradigms of machine learning in the sense that there need not exist explicit inputs and outputs. A key aspect of settings where RL is applied

is the sequential nature of actions that must be taken. For example when designing systems for self-driving cars or automatically flying-helicopters, whether or not an action was a mistake may only be known several steps after the action was taken. While, it is hard to generate examples of what is correct and incorrect behaviour, it is often possible to define a reward function. There has been a lot of interest in RL recently, and in particular it was an integral component of the system that beat the world Go champion (Silver et al., 2016).

4 Some Practical Concerns

Before applying machine learning techniques, it is important to make sure that the data has a useful representation. For example, in the case of spam detection, rather than provide the raw electronic format of an email directly to a learning algorithm, it might be worthwhile to extract useful features, such as specific words, domain, IP address, *etc.* In medical applications, it may be often necessary to have experts pointing out which test should be conducted on patients and what measurements be made that may be useful to detect predisposition towards certain diseases. For the most part, we will work with the assumption that the data has already been cleaned up and put in a format that is suitable for machine learning algorithms to be applied. Recently, there has been renewed focus on automatically extracting features from data, particularly for images, text and audio using deep neural networks.

One of the things to bear in mind that a machine learning algorithm can only do useful things so long as it is given useful data. If one trains on spam data from the 1990s, this may be useless as most spammers have adapted to the modern spam classification technology and constantly seek to break it. As illustrated by the example in the lecture, if one only trains a classifier on photos of dogs taken outside and cats taken inside, one cannot be certain that even a highly accurate classifier is doing anything more than simply counting green pixels!

References

Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Human Genetics*, 7(2):179–188, 1936.

Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

Kevin P. Murphy. *Machine Learning : A Probabilistic Perspective*. MIT Press, 2012.

John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, et al. Genes mirror geography within europe. *Nature*, 456(7218):98–101, 2008.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529 (7587):484–489, 2016.

Alan M. Turing. Computing machinery and intelligence. *Mind*, 59:433–460, 1950.