Professional Certificate in Machine Learning and Artificial Intelligence

Module 2: Introduction to Machine Learning

Quick Reference Guide

Learning outcomes

- Identify the fundamental components of and approaches to machine learning problems.
- Differentiate between machine learning and statistics.
- Classify problems along the major dividing lines of the machine learning landscape.
- Apply the ten steps of a typical machine learning project.
- Identify real-world applications of machine learning in a variety of industries.

What is machine learning?

Simply put, machine learning is about understanding the relationship between several input variables and an output variable.

For example

$Y = \int \left(X_1, \dots, X_p\right) + \in$

where Y is the output variable and $\left(X_1, \dots, X_p\right)$ are the input variables.

Understanding this relationship can be slightly tricky for two reasons:

1. We have limited data points from which to learn about this relationship.
2. The presence of noise (∈), as the relationship between Y and the input variables is a **stochastic** one.

In machine learning, we may term “input variables” as independent variables, predictors, features or fields. Similarly, for “output variable”, we may use the terms dependent variable, response variable, target variable or outcome variable.

We need to learn about the relationship between input variables and output variable for two main reasons:

1. Forecasting
2. Inferencing

Forecasting

In forecasting, we want to be able to predict the value of the outcome variable from the values of the input variable only. In other words, we do not have any data for the output variable.

Examples of real-life application:

1. Determining whether a patient has cancer or not based on collected blood samples. The different characteristics of the blood sample is captured by the input variables.
2. Detecting fraudulent transactions from tax claims or expense claims. The properties of these claims would be captured by the input variables.

The key motivation behind forecasting is to predict the value of the output variable from the available input variables with high accuracy.

The function $\int$ (or the relationship between the input variables and the output variable) that we estimate can be complex. We often use predictors like **neural networks** which work very well but can lead to highly complicated functional relationships.

Inferencing

In inferencing, we want to understand the functional relationship between input variables and the output variable. In other words, we don’t want the relationship to be too complex.

Examples of real-life application:

1. Predicting the sales of different types of marketing campaigns depending on the medium (television, newspaper, radio, etc) we use. We are not merely interested in predicting the increase in sales, but in understanding how the increase can be attributed to the different mediums on which the campaigns were made. This will help in determining the optimal marketing mix
2. Predicting the sales of properties based on different amenities and attributes of houses, such as river-view, etc. Our interest is not in simply predicting the price of the house, but in understanding how a particular attribute contributes to the house price.

The machine learning landscape

There are several major dividing lines in the machine learning landscape, each of which helps classify problems and inform choices on which machine learning method to apply. Some of the major classifications are given below.

Prediction vs classification problem

Prediction problems have an output variable that is a number which is continuous. They are also called regression or estimation problems. In these types of problems, the output variable is always numerical.

For example

Estimating the house price from various input variables. These variables could be categorical (Yes/No) or could be numerical. The output variable that we want to predict, in this case, the house price, is a number.

In **classification problems**, the output variable is not a number. It is a category. Therefore, the output variable is also called categorical.

Example

1. Categorising emails through a spam filter into spam and non-spam mails based on different properties of the email, such as words.
2. Categorising a patient with cancer, as having Type A cancer, Type B cancer, or not having cancer.

We often distinguish between ordinal and nominal categories. A nominal categorical variable does not have any natural ordering (Type A cancer, Type B cancer, No cancer). An ordinal categorical variable has a natural ordering (high, medium, low).

Parametric vs non-parametric approach

In the **parametric approach**, we make an assumption about the function ∫ that we want to estimate.

For example

We may assume that the function which describes the relationship between the input variables and the output variable, is a linear function. With such an assumption, we only need to learn the slopes of that linear function.

In the **non-parametric approach**, we do not make any strong assumptions about the shape of the function ∫ . This approach, in principle, is able to learn any shape. This approach, however, requires more data to learn a reliable functional relationship between the input variables and the output variable.

Supervised vs unsupervised learning

In **supervised learning**, we have a set of input variables X_1 to X_p as well as an output variable Y, and our goal is to study the relationship between these two based on a training data set.

In **unsupervised learning**, we do not have an output variable Y. Our training data comes as a set of records, where each record contains values for the input variables.

For example

In a retailer's customer database, the customers can each be described based on their past purchases, which can serve as our input variables. Does this cause our data set to decompose into various groups of customers? This is an example of unsupervised learning, as no customer comes to a shop wearing the label of an 'existing' or 'new' customer.

The machine learning process

In an idealised form, machine learning project can be decomposed into ten different steps.

Ten steps involved in a machine learning project

1. **Define the purpose of the ML project**

   Take decision with a client or a sponsor within the company whether a project method is a one-off effort or an ongoing procedure. This will have implications for users that apply a method or for those who are going to interpret the result of the method.

2. **Obtain the data set for the analysis**

   Data can be sourced internally, such as customer database and purchase databases, and externally, such as credit rating databases.

3. **Explore, clean and pre process the data**

   Few fields in the data can be scaled to be better handled by the ML method. Deal with missing data and outliners by removing the effected records, manually filling in the data and replacing missing data using an algorithm.

4. **Dimension reduction and feature engineering**

   Remove input variables that will not be available while employing the ML technique to forecast or infer data. Also, remove input variables that are irrelevant for the analysis and not correlated to the output variable. Few of the input variables can also be transformed to numeric values or vice versa.

5. **Determine the ML task at hand**

    Decide the type of task you are dealing with. For example, classification task, prediction task, unsupervised learning task etc.

6. **Partition the data (if supervised ML)**

    For a supervised ML problem, partition the data into training data set, validation data set and test data set.

7. **Choose the ML technique(s)**

    Select an appropriate ML technique for the task under consideration – either a single technique or multiple techniques. A few examples of the techniques used are regression and classification techniques, clustering techniques, k – nearest neighbour technique etc.

8. **Use the ML technique(s)**

    Apply the technique(s) to the task in hand and wait for results.

9. **Interpret the results**

    Compare the algorithms against each other and against simple benchmarking strategies.

10. **Deploy the ML technique (optional)**

    Once the results are analysed, use the chosen algorithm, and deploy it in the task.

Machine learning in the real world

Machine learning is a fast-growing field that is being used in many industries for very different purposes.

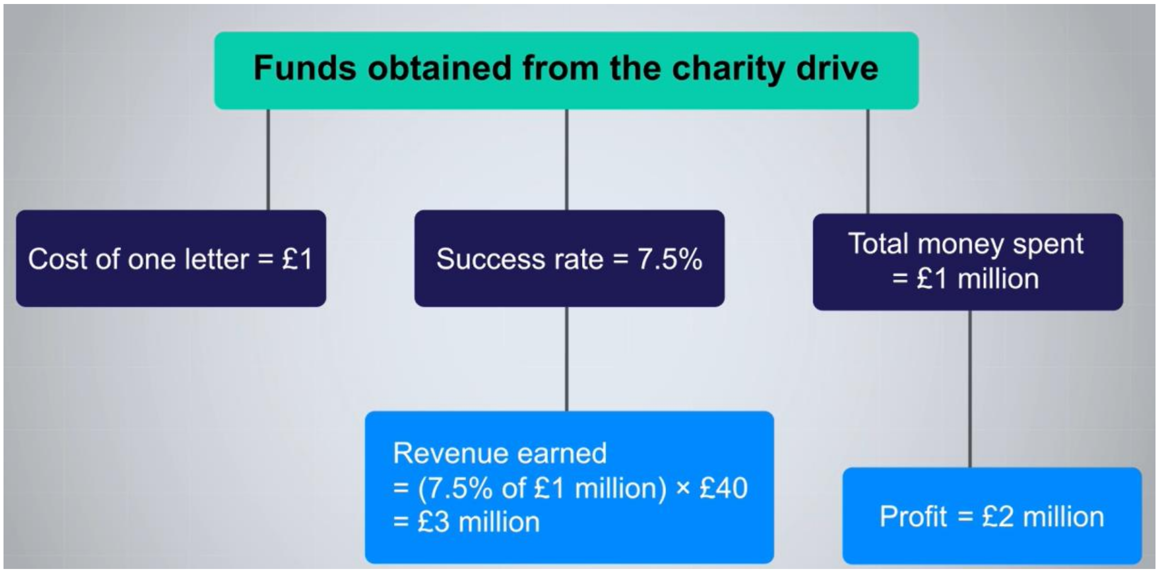
Case study: predicting donor likelihood

The study revolves around an international charity that was piloting an analytics scheme in a country they operate. The pilot country has about one million donors and the charity runs the campaign once a month which repeats annually.

For each campaign the charity sends out a letter sometimes with goodies like a calendar, to all active donors (someone who has made donations within the last three years). These are the details:

- The cost of a letter = 75 pence
- The cost of a letter and a calendar = £5
- Success rate = Percentage of donors who donate (typically between 5% – 10%)

Now, if each letter costs £1 and the success rate is fixed at ten percent:



Funds obtained from the charity drive

Cost of one letter = £1

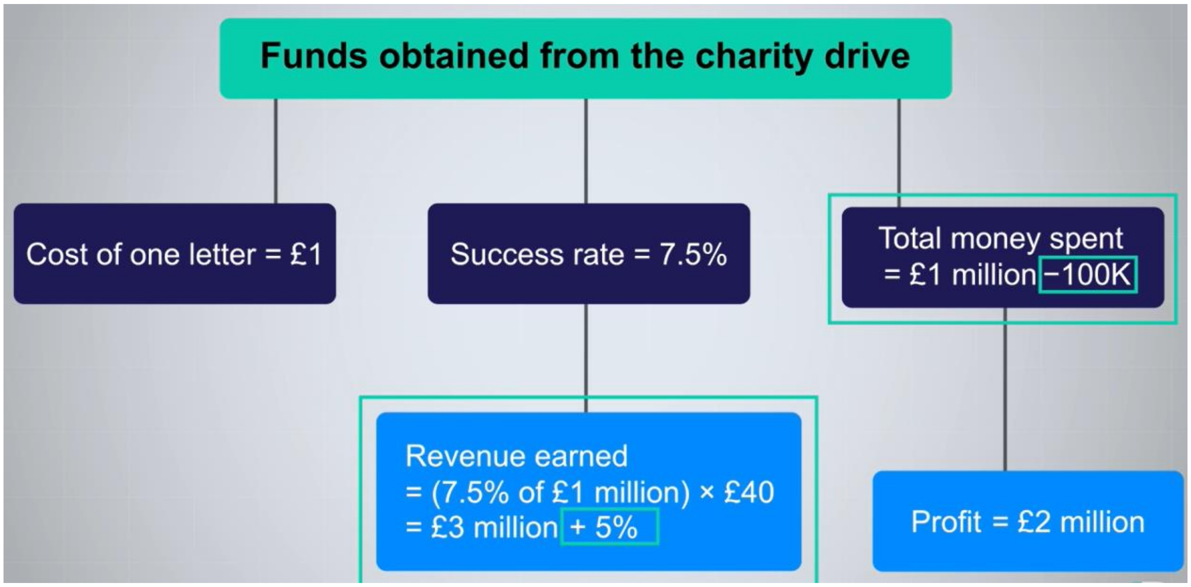
Success rate = 7.5%

Total money spent = £1 million

Revenue earned = (7.5% of £1 million) × £40 = £3 million

Profit = £2 million

Further, if **ten per cent** of the donors that will not contribute are excluded from the list:



Funds obtained from the charity drive

Cost of one letter = £1

Success rate = 7.5%

Total money spent
= £1 million −100K

Revenue earned
= (7.5% of £1 million) × £40
= £3 million + 5%

Profit = £2 million

The revenue earned increased by five per cent along with 100,000 pound cost reduction.

So, the key idea here is to predict the donors who are less likely to donate. A decision tree was later implemented to predict the donors, by selecting parameters such as, age, gender, donation history and location. By implementing this decision tree and analysis results, the charity was able to save tens of thousands of pounds each month.

Real-world applications of machine learning

Yelp: Yelp is a crowd-sourced platform for reviewing local businesses. They use machine learning to identify attractive looking photos that best represent a business to show users first.

Danske Bank: Nordic bank Danske Bank use artificial intelligence and data analytics to detect fraud. By using the deep learning algorithm, the bank saw a 60% reduction in false positives and a 50% increase in true positive detection.

The UK National Health Service (NHS): NHS partnered with AI company Kortical to select the best machine learning models to predict the supply and demand of platelets. Through using AI, they were able to reduce expired (and therefore wasted) platelets by 54% and reduce ad hoc transport by 100%. You can learn more on Kortical’s website.

Blue River Technology: They use machine learning to classify plants to better target herbicides. Their algorithms are built in python using Facebook’s PyTorch.