acdm

association for **clinical data management**

Machine Learning - ACDM Lunch and Learn Summary

Authors: Jennifer Bradford, Tim Armitage and Marcus Schwabedissen
Review: Lauren Alani (Ellis-Hill), Richard Moore

December 2021

www.acdmglobal.org

Machine Learning - ACDM Lunch and Learn Summary

In June-August 2021 members of the ACDM eDigital expert group ran a series of lunch and learn sessions around Artificial intelligence (AI) and machine learning with the aim of introducing data managers to this much talked about area. Each session built upon the next, introducing the overarching concepts and providing a deeper-dive to understand in more detail (without complex maths) the types of machine learning and how it can be applied. The aim of these session was not to make all data managers competent in the application of these techniques but rather for them to appreciate the breadth of approaches available, to understand that the data is key – good quality data and an in-depth knowledge of this data is as important as the method itself and that one size most certainly does not fit all.

If you were not able to attend the sessions then we provide an overview of what was discussed below. There will be more discussion on this topic at the upcoming ACDM conference in Edinburgh on the 13-15th March 2022. In advance of that panel discussion, we would welcome your discussion topics and questions which can be provided in advance to Tim.Armitage@AnjuSoftware.com or alternatively attend the panel discussion in person, where there will be an opportunity to ask the panellists your burning questions.

Introduction to AI and machine learning

There is considerable hype around AI and machine learning across all industries but particularly in healthcare/life sciences in recent years. Various surveys have asked how the world in general feels about AI and the responses are often along the lines of hopeful, excited and curious whilst at the same time concerned. Although AI is reported as the next big thing, the technology has been around a long time, with the first AI boom starting as early as the 1950's and continuing to gain popularity on and off since. Most recently, enablers such as large increases in computing power, growth in the 'datasphere' together with storage capacity and other technological developments have driven this resurgence. AI has widespread application across many industries including finance, transport, education, gaming, media, sports, aerospace and insurance. Healthcare is a key area, with numerous AI companies active across the entire drug development pipeline from discovery through to development and many of the top Pharma companies partnering and/or acquiring AI firms.

So, what is AI? There are many official definitions of AI but essentially it could be described as "how close or how well a computer can imitate or go beyond, when compared to a human being". We can think of AI as weak or strong, with weak AI being reactive, with no or limited memory and the most basic form has no past memory and so cannot use past information to derive information for future actions e.g., the IBM chess program that beat Garry Kasparov in the 1990s. Some weaker AI systems do use past experiences to inform future decisions such as self-driving cars, observations are used to inform actions happening in the immediate future. Strong AI on the other hand is AI that can understand people's emotions, beliefs, thoughts and expectations and be able to interact socially, one step further is an AI that has its own consciousness e.g., a computer with a 'mind' or that is self-aware, and this does not yet exist!

Different methods and approaches come under AI including machine learning, natural language processing, reasoning and planning. We focus on machine learning as these are most commonly seen in the healthcare field. Machine Learning provides a system that has the ability to automatically learn from experience (past data) without being explicitly programmed and there are different types:

- Supervised learning – the system learns from a set of labelled data i.e., each example in the data is tagged with the answer.
- Unsupervised learning – the system is presented with unlabelled data i.e., a set of data where the answer is unknown, and the system has to explore the data and find some structure.
- Deep learning – this can be supervised, unsupervised or semi-supervised and essentially uses multiple layers to progressively extract high level features.
- Reinforcement learning – the system finds the optimal way to accomplish a goal based on rewards or penalties.

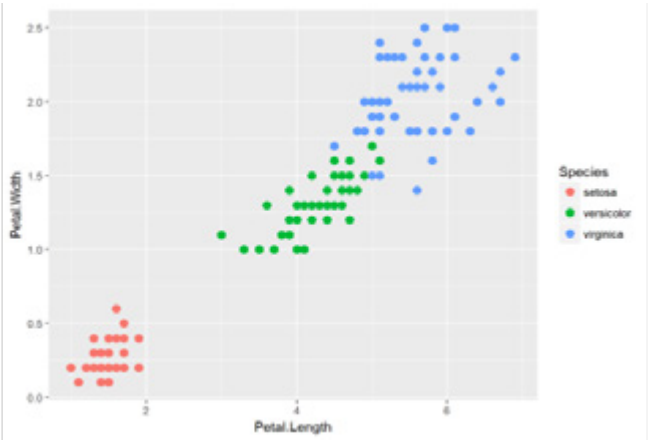
In addition, natural language processing can be used to process and analyse large amount of natural language (unstructured text), which can be used in text mining to extract valuable insights from this text.

Supervised Learning

If we focus a little more on supervised learning, where we have a data set with the answers in it, we would first split the data set into a training and test set. This allows us to develop a machine learning model based on the training set and then use the test data, which the model has never seen previously, to test how well the model performs.

An example of this can be seen based on the Iris data set (a multivariate data set containing 4 features of 50 samples from 3 species of Iris), which is plotted for 2 of the features (petal length and width) and points are coloured according to Iris species in Figure 1(a). If we split this data set into a training and test set and then apply a simple machine learning approach (K-nearest neighbour – k-NN) to the training data, the model generated will enable us to classify the species of Iris from the petal length and width of new data points based on how its nearest neighbours are classified. We can use a confusion matrix to analyse how well the model performs on the test data (see Figure 1(b)):

(a)



(b)

Confusion Matrix - Iris data analysis

Actual \ Predicted	Setosa	Versicolor	Virginica
Setosa	13	0	0
Versicolor	0	19	1
Virginica	0	1	20

Accuracy = 96.3%

Figure 1 (a) A plot to show the petal length and width of the different Iris species in the Iris data set. (b) The confusion matrix from the k-NN model on the training data predicting the species of Iris.

Unsupervised Learning

Unsupervised learning is different to supervised learning in that the data is not labelled, meaning we don't know the answer or how it should be organised. An example of an unsupervised machine learning approach is clustering, one method of which is called K-Means. In this approach we would tell the system how many clusters we are expecting (the k value) and the algorithm would randomly place the k values and then categorise each data point according to how close it was to each of those k values. Once completed for all datapoints, the algorithm would calculate the centre point of all points categorised to a particular k value (centroid) and then recalculate the position of the k values to that centroid and repeat the process. This would continue until the centroids don't move i.e., no points change clusters. A simple example with 2 features is shown in Figure 2.

Figure 2(a) The data has 2 features but it is not known which point is in which cluster (b) the system randomly positions 3 k values (A, B and C) and categorises each point according to how close it is to the k-value (c) The k values are repositioned to the centroid of the categorsied points and this continues until the points are no longer recategorised.

Application of Machine Learning

Now we understand about the different types of machine learning we want to know what is involved in actually applying these approaches to a data set. Of course, it isn't quite as simple as taking some data and pressing a button to run the machine learning. In fact, the machine learning aspect is only a fraction of the time actually spent in a machine learning project, most of the time is spent loading, cleaning and preparing the data prior to analysis. This data preparation is not only necessary but key to a successful analysis. Some of these critical steps are;

- Data cleaning – identify any missing data and understand how to deal with that missing data i.e., delete examples, columns, impute data (of which there are different approaches).
- Feature selection – automatically or manually select the features that will contribute most to the output or prediction.
- to use can have a big impact on performance i.e., can help avoid overfitting, improve accuracy and reduce training time.
- Feature engineering – is there any further processing of the features required, for instance binning (grouping) of ages, this can help prevent over fitting but does remove information. Scaling is another example where features can have vastly different ranges (e.g., age and income), this is necessary for some algorithms. Also, categorical features (non-numerical data) need to be converted to numerical values if they are to be utilised in most machine learning algorithms, deciding the best approach to this can be interesting.

To see a hands on example of a machine learning analysis in practice please visit https://acdmglobal.org/lunch-learn-recordings/ - session 3, where you will also find the recordings for the other 2 lunch and learn sessions on this topic.

A panel discussion around AI and data management will take place at the next ACDM conference, 13-15th March 2022 in Edinburgh.

The ACDM lunch and learn sessions on machine learning were prepared and presented by Tim Armitage, Anju Software, Jennifer Bradford, PHASTAR and Marcus Schwabedissen, QFinity who are members of the ACDM eDigital expert group.