An approach towards determining value through the application of Machine Learning

Christopher Hutchings

Director, Galorath Incorporated.

Email – chutchings@galorath.com

Tel – 310 906 6320

Abstract

Making a distinction between two or more alternatives based on 'value' can be stymied by subjectivity, personal motive and the myopic reluctance to consider characteristics other than cost. It is my intention to illustrate, using a machine learning mechanism for evaluation of related information, an approach towards the discernment of other value characteristics.

The most prevalent use of this technique is in the real estate market, where the value of a home can be determined by such exemplar characteristics as location, square footage and the supply / demand paradigm. I will demonstrate this application and then propose applications closer to home.

Introduction

Machine learning or automated learning has emerged at the forefront edge of Information technology over the last few years and is expanding at an extraordinary rate. Today we are surrounded by an immense number of machine learning applications across a broader spectrum of common-place uses such as email spam detentions systems, credit card fraud systems and internet search mechanisms that adapt to our ever-changing and often nuanced needs.

As programmers our task job is to determine a series of rules that inform a computer how to precisely resolve a specific problem. Machine learning is not that. Machine learning is where the computer determines the rules to solve a problem without being explicitly or prescriptively directed and without bias.

Protection from Spam emails is without question the most prevalent use of machine learning as such a relatable example. Using traditional methods, a programmer would have to develop code that filters out junk email using complex algorithms that contain rules to decide is a message is junk. The program would look for certain words to determine if the subject or context mirrored that which could be recognized as junk or whether the sender was known to you. To establish the efficacy of the program, one would have to feed in test emails and then see how if they were segmented appropriately. This process is laborious

because it is iterative in its nature and is prone to numerous false positives. Add to this the fact that those with nefarious intentions are known to change their tactics often, we are left with a system that requires constant, diligent maintenance.

Better would be a solution where the computer independently establishes and refines logic for filtering of emails; the fundamental premise for machine learning. The programmer would gather a significant number of emails and segment them into two groups; those that are known spam and those that are valid and then apply them to any of a number of readily available, off the shelf and often without cost machine learning tools that do not require any custom code. The ML algorithm would look at the two groups and create its own rules to tell the respective groups apart; a process called 'training'. Put simply, we are giving the system original data and the expected output only and it creates as series of truths that can be used to recreate the output for other data sources. If we provide more data the accuracy increases, albeit to a point constrained by diminishing returns.

With machine learning we don't have to do the laborious and intellectually challenges aspects, in that we don't have to establish any email filtering rules.

This same algorithm can by utilized to resolve lots of other challenges solely through the provision of alternative data, that is that, there is not a requirement to modify a single line of code. We could just as easily feed in pictures of hand written numbers. The algorithm could decide which number each picture represents; it learns how to do new things without you having to explicitly program it. Instead, you show the computer data and the computer learns from the data how to approximate functions that you would have had to program in by hand. Machine learning is a great solution for many complex real world problems that are hard to solve with traditional programming.

Practical applications of Machine learning

The practical applications are numerous and expanding at a colossal rate. The following list is a fragment of the most prevalent uses –

Financial institutions make use of machine learning in two major ways; determine important characteristics of financial data and to prevent fraud. Insights provide early identification of investment opportunities, market leverage and the optimal time to action trades. Data mining can identify clients that are high-risk and cyber surveillance enables the pinpointing of frailties in financial systems.

The **energy** industry utilizes predictive analytics as a measure to reduce the time and cost associated with finding new energy sources. The early detection of systems failures ensures that downtime is avoided and distribution is streamlined.

Another sizable arena for machine learning is the **healthcare** industry, accelerated by the prevalence of wearable sensors and devices that offer real-time access to a patient's health. Machine learning can also empower medical experts to analyze data and see trends that lend themselves to increased efficacy when it comes to diagnosis and treatment.

Websites recommending things that you are highly likely to procure is a machine learning technique that is frequently used in **Commerce,** in that one's purchasing and browsing history tell a story which can be used to generate revenue. This ability to collate information, analyze it and then deploy it as a very personal shopping experience or marketing campaign directive is the future of retail.

Government agencies employ such techniques to resolve a wide spectrum of issues. The challenges faced by 'public safety' centered departments are the most visible to the public, if only because when they fail they are front page news. The ability to trawl through massive amounts of data in real time ensures that those with nefarious objectives are less successful that would otherwise be the case.

Functional applications of machine learning.

As discussed previously, machine learning is a field of activity where computer systems ‘learn’ without the requirement for explicit programming; the output being predictive models. These predictive models can be grouped to a certain extent, (albeit with some cross-over), in the following manner.

Classification is the process where data is divided into classes and the system is required to produce a solution that assigns new inputs to one of these classes. Spam filtering is the standard examples for this, in that, the model will assign new emails a classification of either “spam” or “not spam”. Classification answers questions like, ‘is that a picture fo a cow or an horse?’.

Another supervised learning approach is **Regression analysis,** where the outputs are a prediction of a continuous numerical value e.g. prediction the cost of a product.

With **Clustering** the data is also divided into groups based on similarity. Put simply, the goal of clustering is to segregate data into groups with similar trends. This process is commonly used in retail marketing where it is appropriate to understand the purchasing preferences of certain demographics. It is also applied often to anomaly detection in that, data that cannot be ‘clustered’ would be a strong candidate for being identified as an anomaly datum.

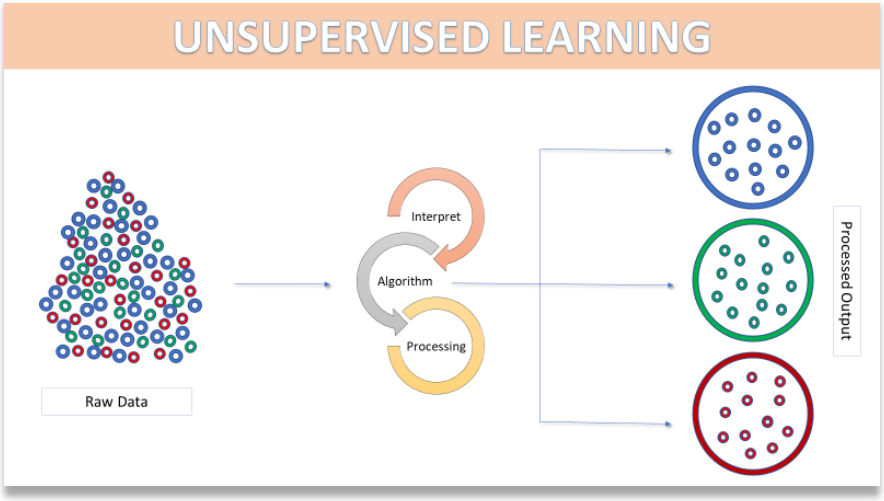
Dimensionality reduction is a process that provides simplification of data through stratification, as often there are too many variables. This technique is often a precursor to classification which can be stymied by a higher the number of features which in turn make visualization of the turning

set harder. Also, it is often seen that variables are correlated, and hence redundant. Dimensionality reduction reduces the number of random variables to those that are of greatest importance, but has with it certain disadvantages, in that, data is discarded and it is often a challenge to establish how many of the principal variables to retain.

Forms of Machine Learning

For the purpose of this area of study I am going to explain how 'supervised machine learning' can be employed to predict value. There are other branches of machine learning, such as unsupervised learning and reinforcement learning and these are detailed below -

Unsupervised learning is analysis carried out against data that has no labels or formal stratification. Most importantly, there is no reference to a 'right' answer and as such the algorithm must establish an answer. The aim of unsupervised learning is to interrogate the data in rode to find some logic or structure within. A practical application would be visual recognition where the system analyzes a vast number of piece of video footage in concert with a text based descriptions of what is being seen. The algorithm using statistical analysis would establish visual patterns and then be correlate them with text to develop theories visual trend and nuance. For example, such a system might be able to establish a high confidence model that can identify a specific item in all videos. The system is never given the 'right' answer but confidence is established through repetition a large number of samples.

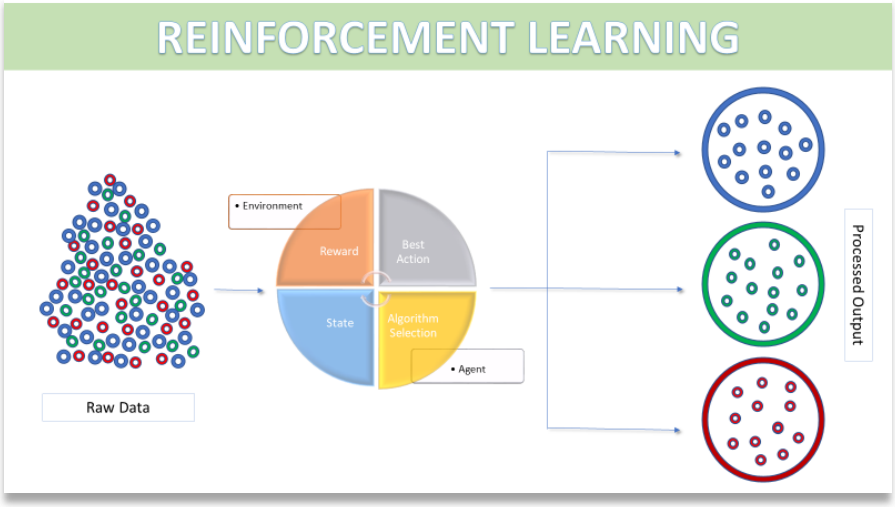


Semi-Supervised learning uses both labelled and unlabeled data to establish a solution, where the latter, (which is more cost effective to procure), is typically an order of magnitude greater in

quantity. This form of learning lends itself well to functions such as regression, classification and forecasting. Humans learn, (to a great extent), in this way; taking a small amount of instruction, e.g. from parents collaborated by large amounts of unlabeled data, ie. experience.

Reinforcement learning is a common technique in the field of robotics. The algorithm determines through iterative trialing which actions yield the 'best' rewards.

Reinforcement learning is concerned with how an agent ought to take actions in an environment so as to maximize some notion of long-term reward. Put simply, the reinforcement learning system will attempt to find a 'policy' that aligns to actions an 'agent' should to take within a given environment. For robotics, the robot uses reinforcement learning to pick a component from one area and the place it elsewhere. Both successes and failures are collated and act as the 'knowledge' to do have increased performance.



Supervised learning is the field of machine learning where the computer learns how to perform a function by looking at labeled training data and the basis for establishing a value prediction. Training a supervised learning model by providing data and directing it what the correct output value should be for that data and our machine learning algorithm uses that data to work out the rules to reproduce those same results. The system is able to generalize what it learned from the training data and use it to predict values for new data.

This application has far reaching promise for the cost analysis and forecasting community; something that I hope to demonstrate with a simple more generic example.

This method is being increasing utilized in the real estate industry. There are of course a plethora of experienced real estate agents who could provide a robust estimate for the value of a property just by looking at it, but given the regulatory mandate to move away from such subjective assessments, the use of more sophisticated mechanism is preferred. That being said, the human process

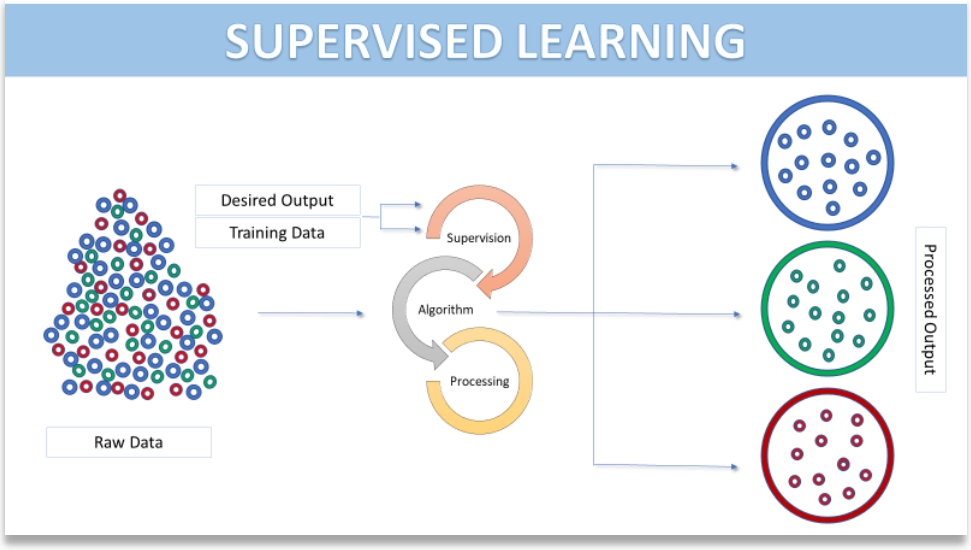
for determining value is worthy of consideration. An expert will know that, a spacious house in a quiet neighborhood, with excellent curb appeal will worth more than a smaller less attractive proposition. Expert knowledge directs us to what variables may be important.

The relatively simple process starts with the collation of data. Sophisticated IS systems and data repositories (MLS), ensure that all pertinent data is collected regarding retail purchases. This data is generally available in XML / JSON formats and so is readily manipulated by the machine learning models. The drivers for value, (neighborhood, square footage, number of bedrooms, etc), are known and have remained constant for decades and a such can be used to direct the training data, along with a final sales price for each property. If the drivers where not know then the dimensionality reduction techniques could be employed to provide indicative insight.

Understanding the overarching function and the within of each of the variables will provide a consistent solution. Luckily the machine learning algorithm carries out these tasks. That being said, there is merit in understanding the steps. An equation is developed to model the problem; in that case to estimate the value of a home. Fundamentally, each of the drivers is assigned a weighting.. Secondly, a cost function is developed to quantify the error in the model, i.e. the variance from what is calculated and what the recorded actuals are and then finally, using a optimization algorithm (typically gradient descent), each of the weighting are ‘tweaked’ until such a point that the cost function variances are statistically insignificant.

We now have a general function, developed by the machine learning algorithm which is applicable in the vast majority of cases.

Value prediction is an incredibly useful technique because it can be used to solve so many of a wide spectrum of problems. Using the exact same technique, it is possible to establish the cost and value of anything based on its attributes, as long as you have suitable training data. Of equal importance is that, estimation in this way ensures that bias is negated and that all attributable factors are accounted for.



The general rule that having more data is better than less holds true for machine learning. The more data you have to training from, the more likely that the user can develop a system that can make an accurate prediction across a wide range of scenarios. At a bare minimum, the 'rule of thumb' is that there should be ten data sources for each of the respective 'features'. Using some of the previously mentioned techniques can help reduce the need for large amount of data, in that, reducing the features, (through dimensionality reduction), will mean that less data is required to get to a robust position. When developing machine learning systems, the programmer should avoid adding in more features. Instead the goal is to include the features that have the most predictive information and exclude the features that aren't predictive

There is of course a cost benefit consideration, in that, whilst more data is better, the cost may be prohibitive and the benefits comparatively minimal.

After training a machine learning model, the next part of the process is to establish accuracy; the error therein due to overfitting and underfitting. This is often done through the application of a measure called the mean absolute error, which looks at the error for each prediction and then offers an average across all of the prediction made by the model.

Overfitting occurs when the model incorporates all, (or too much), of the data but does not establish a consensus pattern and as such will develop bad predictions for any house that was not represented in the training data. Underfitting is the exact opposite, It's when your model is too simple, and doesn't fully learn the pattern in the data. There are a variety of techniques, not discussed here that can be used to redress these issues; many of them are incorporated into the tools developed for machine learning.

The near term future of machine learning

In conclusion the machine learning has dramatically encroached our daily lives, almost to the point of being ubiquitous and whilst there are fears regarding personal liberties, political manipulations and the like, I believe that equal consideration should be afforded to the betterment of society in general and our cost conscious collective in particular. The following are a limited list of examples where Machine learning will be used to meet positive aims -

The wide-scale adoption of **self-driving , (autonomous), vehicles** represents a a far more efficient and effective outlook for transportation. Early reports suggest that such technology could reduce collisions, (a major contributor to congestion), considerably and as such limit the burden on the emergency response health systems and reduce the shared insurance liability. Far more importantly, there is evidence to suggest that fatalities could be reduced by as much as 90%. Vehicles are learning how to drive, in the context of best practice, the law and risk. Fleets of vehicles, equipped with sensors, (LIDAR, RADAR), and cameras have to date recorded hundreds of thousands of data points, all which contribute to a increased intelligence and inherently safer roads.

The US **Healthcare** system is both a significant part of the economy and highly inefficient and as such there is great interest in finding improvements that deliver cost reductions, more profit and ultimately better, more affordable care. Machine learning can provide support in key areas such as diagnosis, pathology and the personalized treatment. Trends can be established that otherwise would not be visible due to prohibitive amounts of data and new medicines can developed through autonomous AI based development regimen.
Additionally, fitness tracking wearables offer the potential for feedback that can act as early or urgent indicators to take action.

Improved management of the **global retail** world is of interest because retail commerce constitutes expenditure in the region of 20 billion US dollars per year. The collation and application of machine learning on consumer data will manifest itself in ways that provide a more personalized purchasing experience. Marketing will be directed to what is of interest to you personally, not just from the perspective of a broad demographic and incentives to purchase will be based on personal preferences. Warehouses will work more efficiency and determining the range ands stocking off an item will be attribute to data rather than subjectivity. Its should be recognized that having and understanding large sources of data ensures that not only are the correct pandects available for purchase but also the correct products are developed that are most likely to be purchased.

It is estimated that in 2015 the costs of damage attributable to **Cybercrime** was in the region of 3 trillion US dollars and that this number likely to double in by 2021, (Accenture citation). Preventative cybersecurity measures are expensive, predominantly due to the fact that mechanisms of attack change with great frequency and ingenuity. Machine learning can be implemented to detect the primary offensive routes early and identify rapid defensive measures; a form of cyber self defense. These systems, unlike their human counterparts are ever vigilant; working all hours, every day.

The **moderation of digital content** is a current and significant concern in general and for social media platforms in particular, as they strive to retain credibility when it comes to the delivery of accurate information to their subscribers; advertising revenue, (an ultimately share price), depends on it. In response to the 'fake news' groundswell both Facebook and Twitter have extensively increased the sixes of their respective content monitoring teams. Emergent machine learning based Artificial Intelligence systems are being employed to meet this challenge head with solutions that adapt to interactions with humans such that content can be moderated on a massive scale.

International **communications** will become hugely more effective once real-time speech translation capabilities are more common place. Machine learning can be used to improve the recognition of speech and dialect nuance and as such ensure that all parties are clearly understood in real time.

There are of course applications for the cost analysis and forecasting community. The following are a few of the less obvious applications –

Actuarial experts are making fervent use of this field of analysis and as such it would be a reasonable assumption that there follows a congruent path for those who estimate **cost uncertainty and contingency.** Machine learning, (generally unsupervised), is used to establish trends that often would be quite elusive or laborious to ascertain. Risk analyst will be able to transition from subjective measures of frequency and impact, to a realm where data drives a risk response.

Integrated logistics support encompasses planning and performing the post production phases of an equipment life in a manner that is concerted, affordable and with assurance that what is needed by the war fighter, is available for the war fighter. Machine learning provides insight and a way through the massive amounts of data available such that logisticians can develop a robust understanding as to what is possible operationally and how to manage deployments efficiently and effectively.

Machine learning is an ubiquitous aspect of our modern data driven lives and despite a myriad of reporting to the contrary; can have positive impacts I content that it has many applications across the cost communities

References

Sebastian Raschka and Vahid Mirjalili Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow, 2nd Edition – Sept 2016

Ryan Roberts, Machine Learning: The Ultimate Beginners Guide For Neural Networks, Algorithms, Random Forests and Decision Trees Made Simple – Jul 29th 2017

Karen Mourikas, Joe King, Denise Nelson, Machine Learning Approach to Cost Analysis, ICEAA SoCal Workshop – Sept 2017