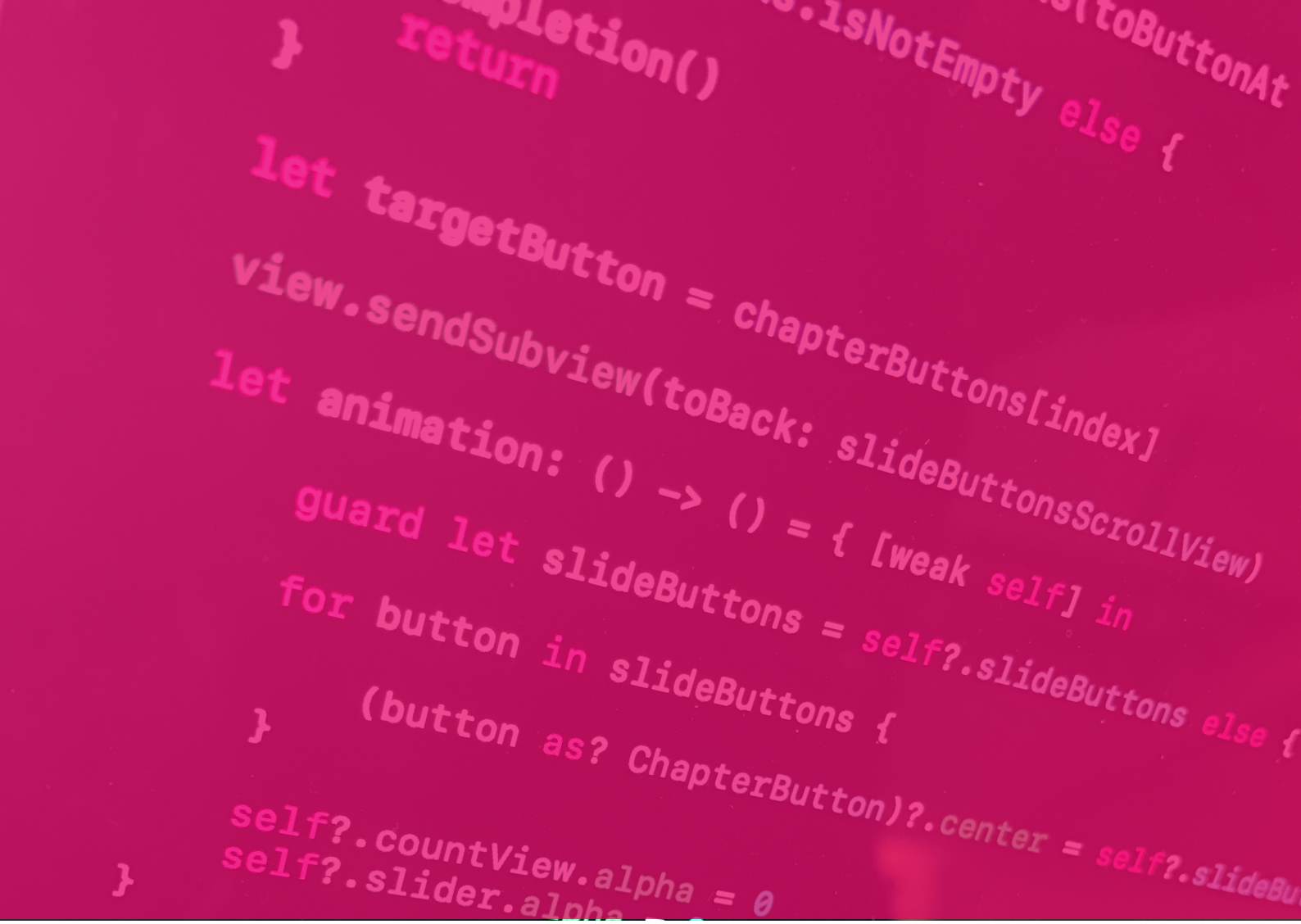


Machine Learning:
Distilling Knowledge from Data

An Experiment in Training Algorithms to Sort Documents

March 2017

OTHER • MEDIA

Contents

Introduction

Artificial intelligence is a field with which many of us are mostly unfamiliar, although we are now hearing about it more and more and it already affects our daily lives more than we may realise. One particular branch of artificial intelligence that is proving an incredibly powerful asset for businesses wanting to analyse data is machine learning.

'Machine learning' is a term used to describe programs that have the ability to learn to recognise specific data by observing data that they are exposed to. It's about building algorithms that learn iteratively from data. The more data you feed in, the better results are produced.

The identification of machine learning as its own branch of technology came about from Arthur Samuel's work at IBM, where, in 1959 he defined it thus:

> 'Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.'

Arthur Samuel utilised machine learning to develop a program that learned how to play checkers (draughts) better than him.

Despite having been around for decades, it's only now that it is coming into its own as a powerful tool for analysis of data. We have long known the importance of data and analytics, but the processing is so labour-intensive. We have so much data available to us, but often struggle to make meaningful sense of it and it can be out of date by the time we have drilled down into the numbers. Machine learning algorithms' value lies

in their ability to process more data and spot more patterns than their human counterparts. Data analysis need no longer be the domain of the highly skilled few.

There are examples of machine learning in use that you may be familiar with – recommendations for future purchases by Amazon or Netflix, fraud detection amongst insurance companies and the driver-less car could be coming to a road near you sooner than you might think.

This document will enable you to understand more about what machine learning is and how it can help businesses. The experiment that Other Media conducted is showcased as a real-world example of the power of machine learning.

What is machine learning – how does it work?

Machine learning algorithms largely fall into one of 3 categories, based on the feedback (training) that they are able to access: supervised, unsupervised and reinforcement learning.

Supervised learning

Trained algorithms

Used when: you know how to classify your data but you need the algorithm to do it for you

Here, you provide a set of training data where you define what the outputs are and tell the algorithm what the outcome should be for each data item. You then run a set of test data through the system for it to classify. Each time the algorithm is exposed to new data, it adds this to its learning and therefore improves its results with time. Think of it like training a puppy. It will bring you your newspaper instead of a pizza menu from the doormat, but you first have to show it what the newspaper looks like, what its features are. Likewise you teach it the features of a takeaway menu flyer. When the mail arrives, the puppy categorises the mail and each day, with more practice, it can do this more accurately.

Unsupervised learning

Untrained algorithms

Used when: you do not know how to classify your data so you ask the algorithm to create the classifier and identify categories

This system does not rely on the data features being identified manually, but instead the system looks for patterns to classify data. Here, you don't know the outcomes (outputs) you are looking for. So your algorithm will look for patterns and identify classifications for you. Here, your puppy will assess features of the items falling onto the doormat (such as colour, size, print finish, paper) and attempt to classify them for you.

What is machine learning – how does it work?

Reinforcement learning

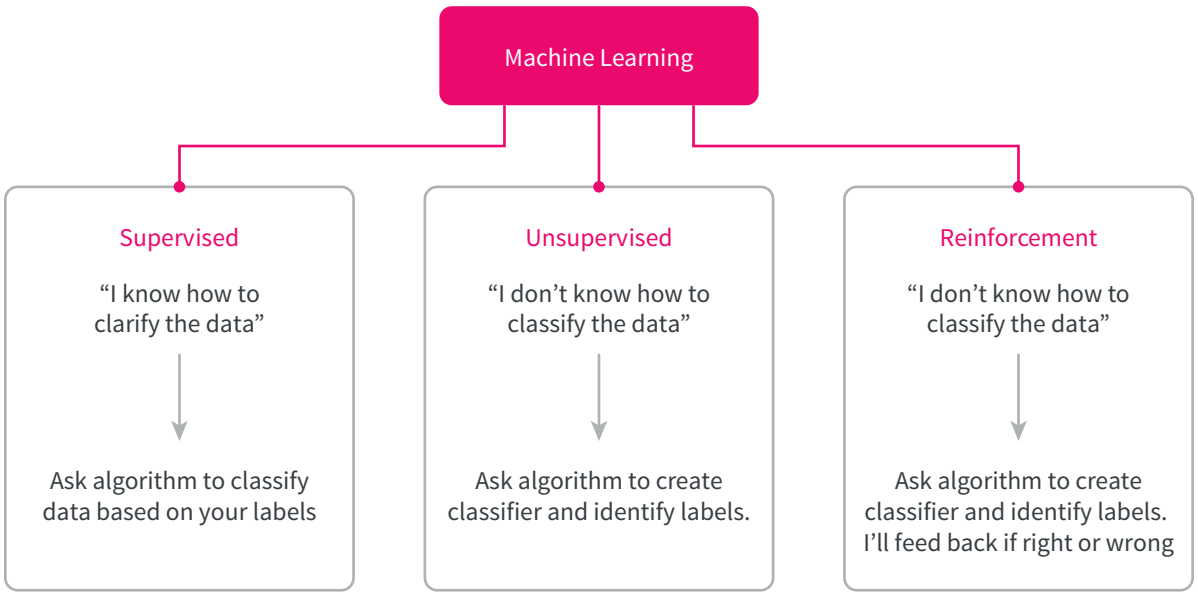
Retro-trained algorithms

Used when: you do not know how to classify the data so the algorithm can do this for you but you have an idea what outputs are right or wrong when presented to you

These algorithms are retro-fed in order to improve their outcomes. They are not trained prior to exposure to test data, they need to decide on an action and then they will be rewarded or punished based on the output they give. The algorithm is not supervised as it is not initially trained with labelled data; yet it is not unsupervised as we provide it with feedback in order for it to learn. In this case, your puppy will be given no information about the post, but when it classifies the post for you, you tell it, for each item, whether it got it right or not. If wrong, you don't tell it the right answer, but it can still learn this way.

With any system, the more data you feed it, the better the result will be.

So, every morning, your puppy gets another chance to recognise and categorise your mail and with time, it becomes more reliable at bringing you the newspaper…until you fancy a pizza, then you have to teach it the new rules…



How can machine learning help businesses?

Traditional analytics produces reports that are often out of date by the time they are published. Once you have the information in a meaningful format, it could be too late to act on it effectively.

ML can process millions of data points in the time a human can process only a few, massively increasing the efficiency of the analysis and what's more, it gets better the more data it processes, so improves over time.

What the experts say

IBM announced in February 2017 that they have extracted the core machine learning technology from IBM Watson to make it available to their customers storing data on their mainframes (those enormous machines that still house data for many of the world's biggest banks, insurance firms, and large retailers) and eventually to any company holding data in a private cloud. They recognise that "Even using the most advanced techniques, data scientists – in shortest supply among today's IT skills – might spend days or weeks developing, testing and retooling even a single analytic model one step at a time."

As this company, whose technology and personnel helped to put the first man on the moon, is recognising, machine learning is a significant focus for the future of businesses.

How is this of interest to my business?

Websites are increasingly dependent on data from many external sources, much of it unstructured, textual data, such as tweets and posts on other social media sites, comments from end users, news stories and reviews, amongst many others. Being able to classify this data is the beginning of being able to utilise it. For example: What is the subject matter? What language is it in? Does it convey positive or negative sentiment? Does it relate to other topics?

How can machine learning help businesses?

Text classification: for answering these questions

Many of the above questions can be answered by employing text classification algorithms. This allows you to break down data and classify documents into pre-defined classes. A machine learning classifier learns to assign a category to text, also known as tagging. Each category is called a label and you can classify using multiple labels or assign only one label to each data item.

Text Classification: Basic Steps

1. Assign labels to a set of training data
2. Train the algorithm by submitting the training data
3. Test the algorithm with a different set of data
4. Deploy the algorithm to classify future data that you feed it
5. Use the insight distilled from the data

Current examples of how this technology is being used include identifying shopping patterns and allowing offers to be targeted to customers and suggesting product recommendations. Identifying patterns can enable financial institutions to see deviations from patterns that could signify fraudulent activity.

Real-time application

You now have meaningful data, rather than raw data. Not only that but it has taken you barely any time at all to compile it. Perhaps the icing on the cake is now to present this in a real-time, digestible format? Our experiment with classifying competitor emails does just that…

The experiment: classifying email newsletters to gain competitor insights

With a view to adding value to the collection and display of competitor newsletters, we asked ourselves this question:

> "Can we build a picture of what is going on in a competitive market by using artificial intelligence and machine learning techniques to automate the reading and classification of news?"

The aim

We focussed our experiment on market surveillance by classifying email newsletters to look at competing fashion brands' activity. This, now useful, knowledge that we have distilled from the raw data can then be visually displayed in real time.

Training the classifier

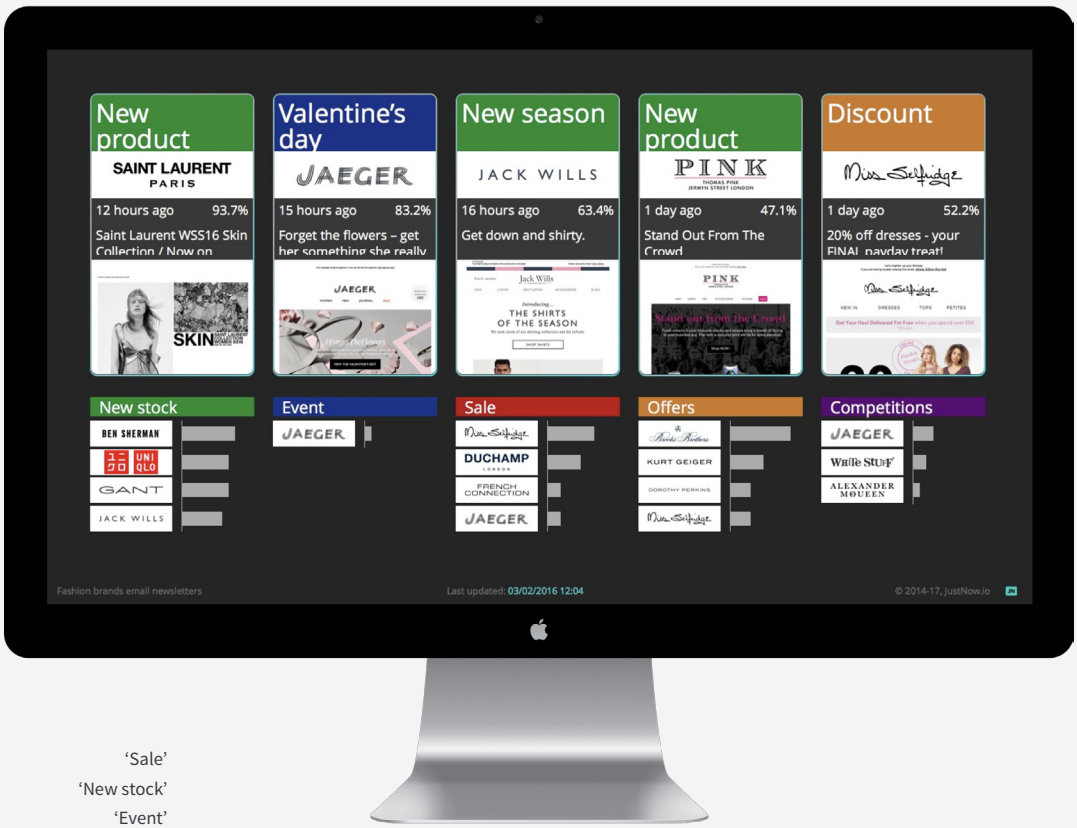
We started by subscribing to a large number of newsletters and then extracted the data automatically using the Google Gmail API (application programming interface). Using a sample of this data we then taught a machine learning algorithm how to classify a subset of these emails: identifying, for example, sale periods, new product announcements and store openings.

The experiment: classifying email newsletters to gain competitor insights

Displaying the insights

Once our classifier was trained we then ran unclassified data through the same system and visualised this flow through our JustNow dashboard system. Using colour codes and a snapshot of the email itself we are able to see, at a glance, each email and a summary of the activities described in recent emails.

The display in this example experiment sections the data into two parts. At the top the emails are displayed as they come in with the computed classification at the top, colour-coded according to the five top-level categories we pre-defined:



Red:	'Sale'
Green:	'New stock'
Blue:	'Event'
Orange:	'Offer'
Purple:	'Competitions'

Note: although there are five items in both the upper and lower parts of the screen there is no vertical relationship between them

The experiment: classifying email newsletters to gain competitor insights

The text in each colour-coded section at the top is the sub-category assigned from within each of these 5 top-level categories, e.g. ‘New season’ and ‘New product’ are both sub-categories within the green ‘New stock’ category.

Below the colour-coded categorisation is the brand logo, allowing the item to be visually recognisable quickly. This is followed by the age of the email and the percentage probability (confidence) value that it belongs to that category (as deduced by the classifier) and then by the subject line of the email. Finally in this top section of the display there is a screen shot image of the HTML version of the email.

This display changes, in real-time, as new emails come in.

Below the latest 5 emails is a summary section of the previous 30 days. Under each of the major category headings are displayed the brands that have been most active in that 30-day period and an associated bar graph to the right of the logos representing the actual number of emails received in those categories.

This summary is refreshed/reordered each day at midnight.

What we learned from the experiment

Classification of data is a complex process and there are a number of factors that will influence the accuracy and ultimate value of any data automatically generated by machine learning tools.

There is a wide range of available tools/services for classifying data to suit varying needs of the user.

Any tool of this type can only be as good as the model data that it uses to learn. Therefore significant thought needs to be given to the structure of the classification being used and much care taken when manually applying this to the model data.

In this experiment, we placed each email into a single category but could have equally as accurately described many emails as being in multiple categories. For this demonstration some classifications have relatively low probabilities (for example where there is little text to process) and these could be considered for omitting or flagging up in some way.

We are excited to see if this approach can be applied to other data sources we are monitoring, including spotting events in Google Analytics and classifying competitor social media actions. We are continuing to focus on this exciting technology to identify further applications to meet the specific needs of our diverse clients.

Conclusion

We have an abundance of data at our disposal.

We have machine learning algorithms widely available to us.

Therefore we have the tools we need to understand data and make it work for us.

Machine learning allows us to do what only a select group of specialists have previously been able to do, in a fraction of the time, giving us up-to-date knowledge. Not only do these machine-based solutions do the job efficiently, they improve continuously.

This knowledge is empowering companies to understand their markets, their competitors and their customers like never before and to better serve those customers.

Artificial intelligence is already in our lives. Machine learning is in our lives. We can choose to misunderstand, be scared of, and ignore these technologies, or embrace them and exploit the efficiencies and insight they have to offer.

Find Out More

Other Media

The Emerson Building
London SE1 9DU
United Kingdom

+44 (0) 20 7089 5959

hello@othermedia.com

visit: other.media