DUNCAN HODGES

A-Z OF DATA

ARTIFICIAL NEURAL NETWORKS

A framework for building machine learning algorithms that is inspired by the brain.

BIG DATA

Data that cannot be efficiently analysed using conventional means, typically because of its volume, veracity, velocity and/or variety.

CONVOLUTIONAL NEURAL NETWORKS

A particular type of Deep Learning that is adept at dealing with images, speech and text.

DEEP LEARNING

An approach using very complex, multi-layered Artificial Neural Networks that requires large amounts of training data but can perform very complex tasks such as image labelling, like identifying cars in a photograph.

EXPLORATORY DATA ANALYSIS

The preliminary investigations of a data set in order to better understand its characteristics.

FEATURE SELECTION

Or feature engineering, the process of selecting inputs to an algorithm and how these inputs should be represented. For example, if we are trying to create an algorithm to predict how many free seats there are on a train journey – what is the best set of information about the journey, is it start time, end time, date, starting station, destination station, what colour the train is, or the weather?

GENETIC ALGORITHM

Process that mimics natural selection, where a solution evolves through the mixing, or 'breeding', and 'mutation' of a set of potential solutions. Most often used in robotic problems or problems where there are a large number of good solutions, and we are trying to find the best from these.

HASHING

A mathematical process that takes data of any size and maps it to data of a fixed size. The process is generally difficult to reverse and is most commonly used in the storage of sensitive data such as passwords or in index structures. Normally seen in action turning passwords like 'Hunter2' into '*****'.

INDEX

A structure that allows the efficient location of a piece of information in a data store.

JUPYTER NOTEBOOK

A document that contains live code, analysis and descriptive text, allows sharing and collaboration around a data analysis task.

KOLMOGOROV-SMIRNOV TEST

A mathematical approach to analysing two datasets to determine if they have equal distributions. Helps with understanding whether two groups in an experiment show different responses to a stimulus.

LOGISTIC REGRESSION

A model that typically predicts a binary outcome (e.g., true / false) from one or more continuous inputs, such as predicting whether someone will repay a loan based on their income.

MACHINE LEARNING

The field of study dealing with algorithms and models that improve their performance as they are provided with more data. This improvement continues until overfitting occurs and maximum performance has been reached.

NATURAL LANGUAGE PROCESSING

A field of study which attempts to train machines to understand and analyse human languages, contributing to applications such as automated customer services assistants on websites.

OVERFITTING

The scourge of modern data science, generally occurs when a system has been 'over trained' on a training data set and cannot generalise to data to which it has not been previously exposed.

PRIVACY

The expectation that personally identifiable information or other sensitive data will be treated securely and sensitively. Getting value from data whilst respecting the privacy of data subjects is the cornerstone of modern data protection laws.

QUALITATIVE DATA

Data that are non-numerical in form.

'R'

A leading free software environment widely used for data analysis tasks.

SUPERVISED LEARNING

The process of learning from a set of labelled data. Typically used in classification techniques where we want to sort inputs into a number of different classes (e.g., email spam / not-spam).

TRUST

Within data science the perception of the credibility of a piece of data, a data source, a data processing system or a prediction.

UNSUPERVISED LEARNING

The process of learning where no previous data is used. Typically used in clustering techniques where we wish to group input data to a number of groups that exhibit similar characteristics, such as grouping movies into genres on a streaming platform.

VISUALISATION

The process for communicating complex information typically through imagery.

WORD EMBEDDINGS

A set of statistical Natural Language Processing techniques where words are allocated a vector of numbers. This vector of numbers effectively encodes the 'meaning' of the word. Machines can then use these vectors to better 'understand' a corpus of text.

X-AXIS

The horizontal axis on a graph, also called the abscissa – a term used at least since the 13th century, by Leonardo of Pisa.

YOTTABYTE

One septillion bytes or 1 trillion Terabytes – about 200,000 trillion photos of Kim Kardashian (see *CSR*, Issue 5)!

ZIPF'S LAW

A feature of all natural languages where the most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc.

Dr Duncan Hodges is a Senior Lecturer in Cyberspace Operations at Cranfield University and is based at the Defence Academy of the United Kingdom. He holds an ESRC National Centre for Research Methods fellowship investigating Digital Identity and is a visitor at the Alan Turing Institute, the UK national institute for data science and artificial intelligence. His research focuses on identity in online and offline spaces, operations in cyberspace and how they can be supported by the innovative and ethical use of data.