The Computational biology and deep learning

John Rim

ABSTRACT

The explosion of molecular and cellular profile data from vast numbers of samples has resulted from technological developments in genomics and imaging. Traditional analytic methodologies are being tested by the fast rise in biological data dimension and collection rate. Modern machine learning approaches, such as deep learning, promise to make accurate predictions by using very large data sets to identify hidden structure. We explore regulatory genomics and cellular imaging applications of this new breed of analytic techniques in this study. We give an overview of what deep learning is and how it may be used to obtain biological insights in various scenarios. We emphasise potential dangers and restrictions to educate computational biologists when and how to make the most of this new tool, in addition to showing particular applications and offering recommendations for practical use.

INTRODUCTION

The machine learning methods are generic methodologies for discovering functional correlations from data without having to describe them beforehand. The capacity to construct predictive models without making significant assumptions about underlying processes, which are typically unknown or inadequately characterised, is appealing in computational biology. For example, the most accurate prediction of gene expression levels is presently performed using sparse linear models or random forests; how the selected features affect the transcript levels is still a study issue. Machine learning algorithms are used in genomes, proteomics, metabolomics, and sensitivity to chemicals. The classic machine learning workflow, which includes four steps: data cleaning and pre-processing, feature extraction, model fitting, and assessment, may be used to characterise the majority of these applications. When available, one data sample is labelled with its response variable or output value y (typically a single number) and includes all covariates and characteristics as input x (commonly a vector of numbers). Data pre-processing, feature extraction, model learning, and model assessment are the four processes in the traditional machine learning workflow. Supervised machine learning methods link input characteristics to an output label y, whereas unsupervised methods learn aspects without using labels that have been seen. Many traditional machine learning algorithms struggle with high-dimensional input data that is complicatedly connected to the associated label. Higher-level characteristics retrieved with a deep model, on the other hand, maybe able to better differentiate across classes. Deep networks develop progressively abstract feature representations from raw data using a hierarchical framework.

An artificial neural network is made up of layers of interconnected compute units that are inspired by neural networks in the brain (neurons). The number of hidden layers in a neural network correlates to its depth, and the maximum number of neurons in one of its layers corresponds to its breadth. Artificial neural networks were renamed "deep networks" when it became feasible to train networks with higher numbers of hidden layers. The network takes data in an input layer, which is then modified nonlinearly through numerous hidden layers until final outputs are generated in the output layer in the canonical configuration (panel A). All neurons in the preceding layer are linked to neurons in the hidden or output layer. Each neuron calculates its output f(x) using a weighted sum of its inputs and a nonlinear activation function (panel B). The rectified linear unit (ReLU; panel B) is the most common activation function, which filters negative impulses to 0 and passes through positive signals. The weights w(i) between neurons are free parameters that are learnt from input/output samples and encapsulate the model's interpretation of the data. Learning minimises the loss function L(w), which gauges the model output's fit to a sample's true label (panel A, bottom). The loss function is high-dimensional and non-convex, analogous to a landscape with numerous hills and valleys, making reduction difficult. It took several decades for the backward propagation approach to be used to generate a loss function gradient through the chain rule for derivatives, allowing for fast stochastic gradient descent training of neural networks. The predicted label is compared to the genuine label during learning to compute a loss for the current task.