Modern Machine Learning Algorithms: Applications in Nuclear Physics

by

Robert Solli

Thesis

for the degree of

Master of Science

Faculty of Mathematics and Natural Sciences
University of Oslo

26th February 2019

Abstract

In this thesis a novel filtering thechinque of AT-TPC noise events is presented using clustering techniques on the latent space produced by a Variational Autoencoder(VAE)

Chapter 1

Theory

1.1 Neural networks

While the basis for the modern neural network was laid more than a hundred years ago in the late 1800's but what we think of in modern terms was proposed by McCulloch und Pitts (1943). They described a computational structure analogous to a human neuron. Dubbed an Artificial Neural Network (ANN) it takes input from multiple sources, weights that input and produces an output if the signal from the weighted input is strong enough. A proper derivation will follow but for the moment we explore this simple intuition. These artificial neurons are ordered in layers, each successively passing information forward to a final output. The output can be categorical or real-valued in nature. A simple illustration of two neurons in one layer is provided in figure 1.1

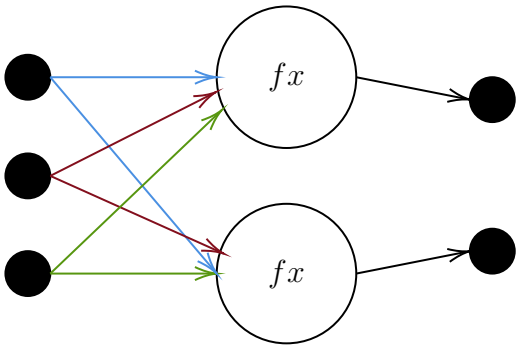

Figure 1.1: An illustration of the graph constructed by two artificial neurons with three input nodes. Colored lines illustrate that each of the input nodes are connected to each of the neurons in a manner we denote as fully-connected.

The ANN produces an output by a "forward pass". If we let the input to an ANN be $x \in \mathrm{R}^N$, and letting the matrix $W \in \mathrm{R}^{N \times D}$ be a representation of the weight matrix forming the connections between the input and the artificial

neurons. Lastly we define the activation function $f(x)$ as a monotonic function on R^1. The function $f(x)$ determines the complexity of the neural network together with the number of neurons per layer and number of layers. For any complex task the activation takes a non-linear form which allows for the representation of more complex problems. A layer in a network implements what we will call a forward pass as defined in function 1.1.

$$\hat{y} = f(\langle x|W\rangle)_D \tag{1.1}$$

In equation 1.1 the subscript denotes that the function is applied element-wise and we denote the inner product in bra-ket notation with $\langle\cdot|\cdot\rangle$. Each node is additionally associated with a bias node ensuring that even zeroed-neurons can encode information. Let the bias for the layer be given as $b \in \mathrm{R}^D$ in keeping with the notation above. Equation 1.1 then becomes:

$$\hat{y} = f(\langle x|W\rangle_D) + b \tag{1.2}$$

As a tie to more traditional methods we note that if we only have one layer and a linear activation $f(x) = x$ the ANN becomes the formulation for a linear regression model. In our model the variables that need to be fit are the elements of W that we denote W_{ij}. While one ordinarily solves optimization problem for the linear regression model with inverting a matrix, we re-frame the problem in more general terms here to prime the discussion of the optimization of multiple layers and a non linear activation function. The objective of the ANN is formulated in a "loss function", which encodes the difference between the intended and achieved output. The loss will be denoted as $\mathcal{L}(y, \hat{y}, W)$. Based on whether the output is described by real values, or a set of probabilities this function, $\mathcal{L}$, takes on the familiar form of the Mean Squared Error or in the event that we want to estimate the likelihood of the output under the data; the binary cross-entropy. We will also explore these functions in some detail later. The ansatz for our optimization procedure is given in the well known form of a gradient descent procedure in equation 1.3

$$W_{ij} \leftarrow -\eta \frac{\partial \mathcal{L}}{\partial W_{ij}} + W_{ij} \tag{1.3}$$

1.2 Autoencoder

An Autoencoder is an attempt at learning a directed reconstruction model of some input. The simplest possible such model is a neural network composed of two parts; an encoder and a decoder. Where the encoder is in general a non linear map ψ

$$\psi : \mathcal{X} \to \mathcal{Z}$$

Where $\mathcal{X}$ and $\mathcal{Z}$ are arbitrary vector spaces with $\dim(\mathcal{X}) > \dim(\mathcal{Z})$. The second part of the network is the decoder that maps back to the original space.

$$\phi : \mathcal{Z} \to \mathcal{X}$$

The objective is then to find the configuration of the two maps ϕ and ψ that gives the best possible reconstruction, i.e the objective $\mathcal{O}$ is given as

$$\mathcal{O} = \arg\min_{\phi,\psi} ||X - \phi \circ \psi(X)||^2 \tag{1.4}$$

As the name implies the encoder creates a lower-dimensional "encoded" representation of the input. This representation can be useful for identifying the information-carrying variations in the data. This can be thought of as an analogue to Principal Component Analysis (PCA)Marsland (2009). More recently the Machine Learning community discovered that the decoder part of the network could be used for generating new samples form the sample distribution, dubbed "Variational Autoencoders" they are among the most useful generative algorithms in modern machine learning.

1.2.1 Variational Autoencoder

Originally presented by Kingma und Welling (2013) the variational autoencoder is a twist upon the traditional

Bibliography

[Kingma und Welling 2013] Kingma, Diederik P. ; Welling, Max: Auto-Encoding Variational Bayes. (2013), dec. – URL http://arxiv.org/abs/1312.6114

[Marsland 2009] Marsland, Stephen: *Machine Learning: An Algorithmic Perspective*. 2009

[McCulloch und Pitts 1943] McCulloch, Warren S. ; Pitts, Walter: A logical calculus of the ideas immanent in nervous activity. In: *The Bulletin of Mathematical Biophysics* 5 (1943), dec, Nr. 4, S. 115–133. – URL http://link.springer.com/10.1007/BF02478259. – ISSN 0007-4985