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

21st January 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 what we think of in modern terms was proposed by McCulloch und Pitts (1943). They propose a computational structure that acts in a fashion like a human neuron. That is it takes input from multiple sources,weights that input and produces an activation if the signal is strong enough. These lose definitions will be made more mathematically clear but for the moment we stick with the simple intuition. Ordinarily a neuron is said to be located in a layer where all the neurons connect to the same input. This layer then produces an output for each neuron in that layer, controlling the output-space of the transformation. A simple illustration of two neurons is provided in figure 1.1

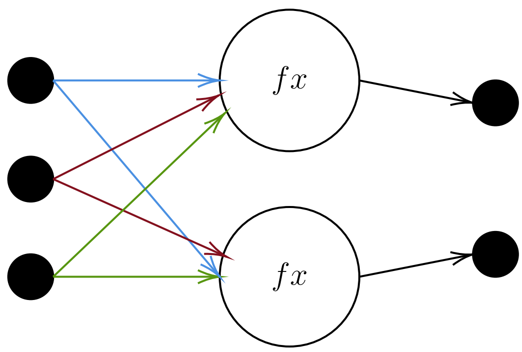

Figure 1.1: An illustration of the graph constructed by two artificial neurons with three input nodes. Colored lines illustrate that each of the input nodes are connected to each of the neurons in a manner we denote as fully-connected.

To formalize this idea let the input be $x \in \mathrm{R}^N$, let then the matrix $W \in \mathrm{R}^{N \times D}$ be the representation of the weight matrix forming the connenctions between the input and the artificial neurons. Lastly we define the activation function $f(x)$ as a univariate, monotonic function on R^1. We will explore this function in some detail later. 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 aditionally associated with a bias node ensuring that even zeroed-cells 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 as presented in equation 1.2

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

As a tie to more traditional methods we note that in the face of the identity transform $f(x) = x$ the equation 1.2 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}. What remains then is to formulate a loss function w.r.t the intended output y that we will describe as $\mathcal{L}(y, \hat{y}, W)$. Based on whether the output is described by a set of probabilities, or real values 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 crossentropy. We will also explore these functions in some detail later. To find the optimal values for W_{ij} we then formulate a gradient optimization like 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

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

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.

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

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

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

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