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

9th April 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 Linear Regression

1.2 Logistic Regression

1.3 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 as neural networks in modern terms was proposed by McCulloch and 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
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 $a(x)$ as a monotonic, once differentiable, function on R^1. The function $a(x)$ 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.

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

In equation 1.1 the subscript denotes that the function is applied element-wise

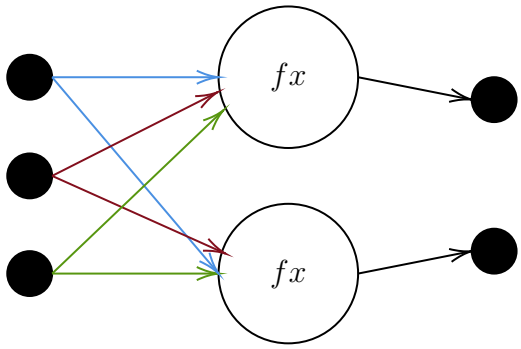

Figure 1.1: An illustration of the graph constructed by two artificial neurons with three input nodes. Colored lines illustrate that each of the input nodes are connected to each of the neurons in a manner we denote as fully-connected.

and we denote the matrix 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:

$$\hat{y} = a(\langle x|W\rangle)_D + b \tag{1.2}$$

As a tie to more traditional methods we note that if we only have one layer and a linear activation $a(x) = x$ 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 by matrix inversion, 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

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

1.3.1 Backpropagation

In the vernacular of the machine learning literature the aim of the optimization procedure is to "train" the model to perform better on the regression, reconstruc-

tion or classification task at hand. Training the model requires the computation of the total derivative in equation 1.3. This is also where the biological metaphor breaks down, as the brain is almost certainly not employing an algorithm so crude as to be formulated by gradient descent. Backpropagation, or automatic differentiation, described most famously in Chapter 8 of **?** is a method of computing the partial derivatives required to go from the gradient of the loss w.r.t the output of the ANN to the gradient w.r.t the individual neuron weights in the layers of the ANN. The algorithm begins with computing the total loss, here exemplified with the squared error function, in equation 1.4

$$E = \mathcal{L}(y, \hat{y}, W) = \frac{1}{2} \sum_n \sum_j (y_{nj} - \hat{y}_{nj})^2 \tag{1.4}$$

The factor one half is included for practical reasons to cancel the exponent under differentiation. As the gradient is multiplied by an arbitrary learning rate η this is ineffectual on the training itself. The sums define an iteration over the number of samples, and number of output dimensions respectively. Taking the derivative of 1.4 w.r.t the output, $\hat{y}$, we get

$$\frac{\partial E}{\partial \hat{y}_j} = \hat{y}_j^M - y_j \tag{1.5}$$

Recall now that for an ANN with M layers the output fed to the activation function is

$$x_j^M = \langle a^{M-1} | W^M \rangle + b_j \tag{1.6}$$

Where the superscript in the inner product denote the output of the second-to-last layer and the weight matrix being the last in the layers. The vector x_j is then fed to the activation to compute the output

$$\hat{y}_j^M = a(x_j^M) \tag{1.7}$$

The activation function was classically the sigmoid (logistic) function but during the last decade the machine learning community has shifted to largely using the rectified linear unit (ReLU) as activation. Especially after the success of Krizhevsky et al. (2012) with AlexNET in image classification. Depending on the output (be it regression or classification) it might be useful to apply the identity transform or a soft max function in the last layer. This does not change the derivation except to change the derivatives in the last layer. We here exemplify the back propagation with the ReLU, which has the form

$$\text{ReLU}(x) = \begin{cases} x, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{1.8}$$

The ReLU is obviously monotonic and its derivative can be approximated with the Heaviside step-function.

$$H(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{1.9}$$

We again make explicit that knowing the form of equations 1.4, 1.8 and 1.9 is not necessary but provided for clarity. We then return to equation 1.5 and manipulate the expression via the chain rule

$$\frac{\partial E}{\partial x_j} = \frac{\partial E}{\partial \hat{y}_j} \frac{\partial \hat{y}_j}{\partial x_j} \tag{1.10}$$

The second derivative of the r.h.s we know from our choice of the activation to be equation 1.9, inserting to evaluate the expression we find

$$\frac{\partial E}{\partial x_j^M} = (\hat{y}_j - y_j) H(x_j^M) \tag{1.11}$$

To complete the derivation we further apply the chain rule to find the derivative in terms of the weight matrix elements.

$$\frac{\partial E}{\partial w_{ij}^M} = \frac{\partial E}{\partial x_j^M} \frac{\partial x_j^M}{\partial w_{ij}^M} \tag{1.12}$$

Recall the definition of x_j as the affine transformation defined in equation 1.1. The derivative of the inner product w.r.t the matrix elements is simply the previous layers output. Inserting this derivative of equation 1.6 we have the expression for our derivatives of interest.

$$\frac{\partial E}{\partial w_{ij}} = (\hat{y}_j - y_j) H(x_j) \text{ReLU}(x_i^{M-1}) \tag{1.13}$$

Separately we compute the derivatives of 1.11 in terms of the bias nodes.

$$\frac{\partial E}{\partial b_j} = \frac{\partial E}{\partial x_j} \frac{\partial x_j}{\partial b_j} = (\hat{y}_j - y_j) H(x_j) \cdot 1 \tag{1.14}$$

This procedure is then repeated for the earlier layers computing the $\partial E / \partial w$ as we go. The backward propagation framework is highly generalizable to variations of activation functions and network architectures. The two major advancements in the theory of ANNs are both predicated on being fully trainable by the backpropagation of errors. Before we consider these improvements made by the introduction of Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) we remark that not only are we free to chose the activation function remarkably freely the backpropagation algorithm also makes no assumptions on the transformation that constructs x_j. As long as it is once differentiable in terms of w_{ij} we are free to pick this transformation also.

there should be a note on the importance of initialization of the weights

1.4 Autoencoders

1.4.1 Introduction to autoencoders

An Autoencoder is an attempt at learning a distribution over some data by reconstruction. The interesting part of the algorithm is in many applications that it is in the family of latent variable models. Which is to say the model encodes the data into a lower dimensional latent space before reconstruction. The goal, of course, is to learn the distribution $P(\mathcal{X})$ over the data with some parametrized model $Q(\mathcal{X}|\theta)$ The model consists of two discrete 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 model 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

$$\mathcal{O} = \arg\min_{\phi,\psi} ||X - \phi \circ \psi||^2 \tag{1.15}$$

Where the $\circ$ operator denotes function composition in the standard manner. As the name implies the encoder creates a lower-dimensional "encoded" representation of the input. This objective function is optimized by a mean-squared-error cost in the event of real valued data, but more commonly through a binary cross-sentropy for data normalized to the range $[0, 1]$. This representation can be useful for identifying whatever information-carrying variations are present in the data. This can be thought of as an analogue to Principal Component Analysis (PCA) (Marsland (2009)). In practice the non-linear maps, ψ and ϕ, are most often parametrized and optimized as ANNs. ANNs are described in detail in section 1.3. The autoencoder was used perhaps most successfully in a de-noising tasks. 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.

Citation needed. Also should I include example of denoising autoencoders ? Maybe a description at least.. Link to notebook maybe?

1.4.2 Variational Autoencoder

Originally presented by Kingma and Welling (2013) the Variational Autoencoder (VAE) is a twist upon the traditional autoencoder. Where the applications of an ordinary autoencoder largely extended to de-noising with some authors using it for dimensionality reduction before training an ANN on the output the VAE seeks to control the latent space of the model. The goal is to be able to generate samples from the unknown distribution over the data. Imagine trying to draw a sample from the distribution of houses, we'd be hard pressed to produce anything remotely useful but this is the goal of the VAE. In this thesis the generative properties of the algorithm is only interesting as a way of describing the latent space. Our efforts largely concentrate on the latent space itself and importantly discerning whether class membership, be it a physical property or something more abstract, [1], is encoded.

The variational autoencoder cost

In section 1.4.1 we presented the structure of the autoencoder rather loosely. For the VAE which is a more integral part of the technology used in the thesis a more rigorous approach is warranted. We will here derive the loss function for the VAE in such a way that makes clear how we aim to impose known structure of the latent space. We begin by considering the family of problems encountered in variational inference, where the VAE takes its theoretical inspiration from. We define the joint probability distribution of some hidden variables z and our data x conditional on some β. In a traditional modeling context we would coin z as including model parameters and β would then denote the hyperparameters. The variational problem is phrased in terms of finding the posterior over z, given β

$$p(z|x,\beta) = \frac{p(z,x|\beta)}{\int_z p(z,x|\beta)} \tag{1.16}$$

citation?

The integral in the denominator is intractable for most interesting problems . This is also the same problem that Markov Chain Monte Carlo (MCMC) methods aim at solving. In physics this family of algorithms has been applied to solve many-body problems in quantum mechanics primarily by gradient descent on variational parameters .

citation? Comph-phys 2 compendium?

Next we introduce the Kullback-Leibler divergence (KL-divergence) (Kullback and Leibler (1951)) which is a measure of how much two distributions are alike, it is important to not that it is however not a metric. We define the KL-divergence in equation 1.17 from a probability measure P, to another Q, by their probability density functions p, q over the set $x \in \mathcal{X}$.

[1] examples include discerning whether a particle is a proton or electron, or capturing the "five-ness" of a number in the MNIST dataset

$$D_{KL}(P||Q) = -\int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx \tag{1.17}$$

$$= \langle \log\left(\frac{p(x)}{q(x)}\right) \rangle_p \tag{1.18}$$

In the context of the VAE the KL-divergence is a measure of dissimilarity of P approximating Q (Burnham et al. (2002)). The derivation then sensibly starts with a KL-divergence.

We begin by defining $q(z|x)$ to be the true posterior distribution over the latent variable $z \in \mathcal{Z}$, conditional on our data $x \in \mathcal{X}$ with a true posterior distribution $p(x)$ and $q(z)$, with an associated probability measure Q as per our notation above. Let then the distribution over the latent space parametrized by the autoencoder be given as $\psi(z|x)$, where the autoencoder parametrizes a distribution $\eta(x)$, and an associated probability measure Ψ. And recalling Bayes rule for conditional probability distributions $p(z|x) = (p(x|z)p(z)/p(x)$

$$D_{KL}(\Psi||Q) = \langle \log\left(\frac{\psi(z|x)}{q(z|x)}\right) \rangle_\psi \tag{1.19}$$

$$= \langle \log\left(\psi(z|x)\right) \rangle_\psi - \langle \log\left(p(x|z)q(z)\right) \rangle_\psi + \log\left(p(x)\right) \tag{1.20}$$

$$= \langle \log\left(\frac{\psi(z|x)}{q(z)}\right) \rangle_\psi - \langle \log\left(p(x|z)\right) \rangle_\psi + \log\left(p(x)\right) \tag{1.21}$$

Rearranging the terms we arrive at the variational autoencoder cost

$$\log(p(x)) - D_{KL}(\Psi||Q) = \langle \log\left(p(x|z)\right) \rangle_\psi - \langle \log\left(\frac{\psi(z|x)}{q(z)}\right) \rangle_\psi \tag{1.22}$$

We are still bound by the intractable integral defining the evidence $p(x) = \int_z p(x,z)$ which is the same integral as in the denominator in equation 1.16. The solution appears by approximating the KL-divergence up to an additive constant by estimating the evidence lower bound (ELBO). This function is defined as

$$ELBO(q) = \langle \log(p(z,x)) \rangle - \langle \log(q(z)) \rangle \tag{1.23}$$

To fit the VAE cost we rewrite the ELBO in terms of the conditional distribution of x given z

$$ELBO = \langle \log(p(z)) \rangle + \langle \log(p(x|z)) \rangle - \langle \log(q(z|x)) \rangle \tag{1.24}$$

Finally the ELBO can be related to the VAE loss by applying Jensen's inequality (J) to the log evidence

$$\begin{aligned}
\log(p(x)) &= \log \int_z p(x|z)p(z) && (1.25)\\
&= log \int_z p(x|z)p(z)\frac{q(z|x)}{q(z|x)} && (1.26)\\
&= log\langle p(x|z)p(z)/q(z|x)\rangle && (1.27)\\
&\overset{(J)}{\geq} \langle \log(p(x|z)p(z)/q(z|x))\rangle && (1.28)\\
&\geq \langle \log(p(x|z))\rangle + \langle \log(p(z))\rangle - \langle \log(q(z|x))\rangle && (1.29)
\end{aligned}$$

Now we have a fully computationally tractable system. We note that in the above notation we would parametrize the distribution $p(x|z)$ as a neural network, in machine learning parlance it is called the generator network. Kingma and Welling (2013) showed that this variational lower bound on the marginal likelihood of our data is feasibly implemented with a neural network when trained with variations of gradient descent.

1.5 Notes

1. L1 regularization on the LSTM cells in the draw network seem to encourage the network to capture "many events". Looks like many spirals in one. While L2 (or sparese) regularization represents the images well. Can we represent the inner workings of the LSTM in some way?

2. Benchmark reconstruction loss for DRAW is at 255 - 1200 nodes, 60 filters, 10 timesteps, L2 regularization, Adam optimizer

3. Nesterov momentum yields suboptimal results. Reconstruction loss of about 1.4 times the loss when using Adam

4. Adadelta yields pure noise reconstructions (short simulation)

5. Adagrad yields localized "clouds" in the output

6. for simulated data it seems we can compress to about $350 \sim 300$ nodes in the encoder lstm. And to 3 dimensions in the latent space

7. In what seems like the minimal compressed state for the simulated data the training seems unstable and will frequently get stuck in local minima or have the gradient explode

8. DRAW without attention seems unable to learn even the simulated distribution at 128 by 128 pixels

9. In the DRAW algorithm the glimpse is specified by an affine weight transformation - but to be comparable it should be constant as a hyperparameter.

10. Implementing the glimpse as a hyperparameter was hugely successful, perhaps surprisingly in decreasing the reconstruction loss. Now remains the task of using the latent representations for classification

Bibliography

Burnham, K. P., Anderson, D. R., and Burnham, K. P. (2002). *Model selection and multimodel inference : a practical information-theoretic approach*. Springer.

Kingma, D. P. and Welling, M. (2013). Auto-Encoding Variational Bayes.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks.

Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Marsland, S. (2009). Machine Learning: An Algorithmic Perspective.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133.