

Deep Learning for Tabular Data: An Empirical Study

by

Jan André Marais



*Thesis presented in partial fulfilment of the requirements for
the degree of Master of Commerce (Mathematical Statistics)
in the Faculty of Economic and Management Sciences at
Stellenbosch University*

Supervisor: Dr. S. Bierman

December 2018

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date:

Copyright © 2018 Stellenbosch University
All rights reserved.

Abstract

Deep Learning for Tabular Data: An Empirical Study

J. A. Marais

Thesis: MCom (Mathematical Statistics)

December 2018

English abstract.

Uittreksel

Diepleer Tegnieke vir Gestruktrueerde Data: 'n Empiriese Studie

(“Deep Learning for Tabular Data: An Empirical Study”)

J. A. Marais

Tesis: MCom (Wiskundige Statistiek)

Desember 2018

Afrikaans abstract

Acknowledgements

I would like to express my sincere gratitude to the following people and organisations ...

- The UCI Machine Learning Repository (Dheeru and Karra Taniskidou, 2017) for hosting a platform to share datasets.

Contents

Declaration	i
Abstract	ii
Uittreksel	iii
Acknowledgements	iv
Contents	v
List of Figures	viii
List of Tables	ix
List of Abbreviations and/or Acronyms	x
Nomenclature	xi
1 Introduction	1
1.1 Deep Learning	1
1.2 Tabular Data	3
1.3 Challenges for Deep Learning on Tabular Data	3
1.4 Overview of Statistical Learning Theory/Machine learning . . .	4
1.5 Outline	9
2 Neural Networks	10
3 Neural Networks	11
3.1 Introduction	11
3.2 The Structure of a Neural Network	12
3.3 Training a Neural Network	16

3.3.1	Optimisation	16
3.3.2	Optimisation Example	19
3.3.3	Backpropagation	19
3.3.4	Learning Rate	22
3.3.5	Basic Regularisation	22
3.4	Representation Learning	23
3.5	Vanilla MLP on Tabular Data	24
4	Deep Learning	25
4.1	Introduction	25
4.2	Autoencoders	25
4.3	Combating Overfitting	26
4.3.1	Transfer Learning	26
4.3.2	Dropout	26
4.3.3	Data Augmentation	26
4.4	Modern Architectures	27
4.4.1	Normalisation	27
4.4.2	Residual Networks	27
4.4.3	Embeddings	27
4.4.4	Attention	28
4.5	One-cycle Policy	28
4.6	Model Interpretation	30
4.6.1	Model Agnostic	31
4.6.2	Neural Network Specific	31
5	Deep Learning for Tabular Data	32
5.1	Introduction	32
5.2	Related Work	33
5.3	Input Representation	34
5.3.1	Numerical Features	34
5.3.2	Categorical Features	35
5.3.3	Combining Features	36
5.4	Still to categorise	36
5.5	Learning Interactions	42
5.5.1	Fully-Connected Layers	42
5.5.2	Going Deeper	43
5.5.3	Attention	43

5.5.4	Parallel Streams	43
5.6	Dealing with Small Datasets	43
5.6.1	Data Augmentation	43
5.6.2	Unsupervised Pretraining	44
5.7	Other	45
5.8	Recommended Approach	45
6	Experiments	46
6.1	Introudction	46
6.2	Datasets	46
6.3	Evaluation	47
6.3.1	Metrics	48
6.3.2	Cross-validation	49
6.4	General Approach	49
6.5	Architectural Search	49
6.6	Sample Size	50
6.7	Mixup	50
6.8	Pretraining	50
6.9	Attention	50
6.10	Comparisons To Tree-based Methods	51
6.11	Example Interpretation	51
7	Conclusion	52
	Appendices	53
A	Appendix A	54
	Bibliography	55

List of Figures

- 1.1 The exponential growth of published papers containing the term
Deep Learning 2
- 3.1 The structure of an artifical neural network (This is still a placeholder). 12

List of Tables

List of Abbreviations and/or Acronyms

ANN	Artificial Neural Network
CNN	Convolutional Neural Network
CV	Computer Vision
DL	Deep Learning
kNN	k -Nearest Neighbour
mAP	Mean Average Precision
ML	Machine Learning
NLP	Natural Language Processing
NN	Neural Network
SGD	Stochastic Gradient Descent
SotA	State-of-the-Art

Nomenclature

N	number of observations in a dataset
p	input dimension or the number of features for an observation
K	number of labels in a dataset
\mathbf{x}	p -dimensional input vector $(x_1, x_2, \dots, x_p)^\top$
λ	label
\mathcal{L}	complete set of labels in a dataset $\mathcal{L} = \{\lambda_1, \lambda_2, \dots, \lambda_K\}$
Y	labelset associated with \mathbf{x} , $Y \subseteq \mathcal{L}$
\hat{Y}	predicted labelset associated with \mathbf{x} , $\hat{Y} \subseteq \mathcal{L}$, produced by $h(\cdot)$
\mathbf{y}	K -dimensional label indicator vector, $(y_1, y_2, \dots, y_K)^\top$, associated with observation \mathbf{x}
$(\mathbf{x}_i, Y_i)_{i=1}^N$	multi-label dataset with N observations
D	dataset
$h(\cdot)$	multi-label classifier $h : \mathbb{R}^p \rightarrow 2^{\mathcal{L}}$, where $h(\mathbf{x})$ returns the set of labels for \mathbf{x}
θ	set of parameters for $h(\cdot)$
$\hat{\theta}$	set of parameters for $h(\cdot)$ that optimise the loss function
$L(\cdot, \cdot)$	loss function between predicted and true labels
$f(\cdot)$	label prediction module, $f : \mathbb{R}^p \rightarrow \mathbb{R}^K$
$t(\cdot)$	thresholding function, $t : \mathbb{R}^K \rightarrow \{0, 1\}^K$
$\mathcal{N}(\mathbf{x})$	points in the input space neighbourhood of \mathbf{x}

Chapter 1

Introduction

1.1 Deep Learning

This thesis is about using *Deep Learning* (DL) approaches to solve *Machine Learning* (ML) tasks where tabular data are the inputs. The field of DL is an extension of the class of ML algorithms called *Artificial Neural Networks* (NNs). The rapid development in computing power and the growing abundance of data available, awoken the slumbering field of NNs and resulted in optimisation and architecture design advancements, creating the DL field as we know it today (Lecun *et al.*, 2015).

DL is receiving a remarkable amount of attention in academia (see Figure 1.1). DL has already shown tremendous value in application areas such as *Computer Vision* (CV) [cite], audio processing [cite], and *Natural Language Processing* (NLP) [cite], significantly improving upon the then *State of the Art* (SotA). In the aforementioned application areas, DL reached a maturity level sufficient to be able to run these systems in a production/commercial environment, *e.g.* voice assistants [cite], face recognition [cite], language translation [cite] to name a few.

One of the most attractive attributes of DL is its ability to model almost any input-output relationship. DL has been used to generate art [cite] and music [cite], controlling autonomous cars [cite], playing video games [cite], beating the world's best Go player [cite], suggesting which videos to watch [cite], and improving the quality of images [cite].

One thing that these successful DL applications have in common is that the modality of the data on which they operate is homogeneous. In CV the data



Figure 1.1: The exponential growth of published papers containing the term *Deep Learning*

are pixel values, in NLP the data are words and in audio processing the data are sound waves. This is not a criteria for DL to be successful but is certainly a driver for its success in these domains. Modelling of homogeneous data is easier since every input feature can be treated the same. Furthermore, universal patterns exist in each of these domains allowing for knowledge to be transferred between tasks of the same domain, both knowledge acquired by humans and that learned by the DL model. For example in CV, advancements in classifying pictures of pets will most likely also be applicable to identifying tumors in X-rays and the patterns learned by the model to do the one task may also be useful to do the other (see *Transfer Learning*).

A data domain in which DL does not flourish is that of tabular data. Although work is being done on the problem [cite] and SotA results were received on rare occasions [cite], the area is nowhere near as mature or receiving as much attention compared to CV and NLP. ML tasks operating on tabular data are typically more effectively solved by using tree-based methods [cite], which is also evident by looking at the winning solutions of relevant Kaggle

competitions¹. This is possibly largely influenced by the heterogeneity of tabular data (Shavitt and Segal, 2018), which we discuss in the next section.

1.2 Tabular Data

- how is it different to other data sources
- what type of application
- sota approaches (look for that one paper on 121 UCI datasets)
- common workflow

Tabular data is also important. But each column is different and thus in a way more difficult to learn representations. At the moment methods on tabular data are dominated by tree based boosting methods. See kaggle competitions. In some cases where there was enough data deep learning got a slight upperhand. But it is still not clear when a tabular dataset is best suited for dl and neither how then to apply dl to such a dataset.

Nns have been use on tabular data but old techniques and very few of the moden techniques have been tested on tabular data.

1.3 Challenges for Deep Learning on Tabular Data

- summarise the core research of this thesis
- also serves as a motivation for the thesis
- give the aim of this thesis

Many tabular data sets are challenging to represent and model due to its high dimensionality, noise, heterogeneity, sparseness, incompleteness, random errors, and systematic biases (Miotto *et al.*, 2016).

This thesis acts as a tutorial for applying Deep Learning to tabular data and also includes a literature review and empirical study on what works where.

This thesis acts as an tutorial on applying dl to tabular data. We will look at existing work on the matter, see that it is lacking, see what we can borrow from the other domains, do an empirical study to look for clues. Especially

¹<https://www.kaggle.com>

layers, embeddings, pretraining, augmentation, modern training policies, batch size. The use of dl is often restricted by its perceived lack of interpretability and the here we will explore ways that we can interpret them with model agnostic and nn specific methods.

1.4 Overview of Statistical Learning

Theory/Machine learning

Overview of core concepts in Machine Learning used and referred to in the rest of the thesis. - Supervised vs Unsupervised - Regression vs Classification - Overfitting - Bias-variance trade-off - Interpretability

Machine or statistical learning algorithms (used interchangeably) are used to perform certain task that are too difficult or inefficient to solve with fixed rule-based programs. These algorithms are able to learn how to perform a task from data. For an algorithm to learn from data means that it can improve its ability in performing an assigned *task*, with respect to some *performance measure*, by processing *data*. This section gives a brief look at some of the important types of tasks, data and performance measures in the field of statistical learning.

A learning task describes the way an algorithm should process an observation. An observation is a collection of features that have been measured from some object or event that we want the system to process, for example an image. We will represent an observation by a vector $\mathbf{x} \in \mathbb{R}^p$ where each element x_j of the vector is an observed value of the j -th feature, $j = 1, \dots, p$. For example, the features of an image are usually the color intensity values of the pixels in the image.

Many kinds of tasks can be solved with statistical learning. One of the most common learning tasks is that of *classification*, where it is expected of an algorithm to determine which of K categories an input belongs to. To solve the classification task, the learning algorithm is usually asked to produce a function $f : \mathbb{R}^p \rightarrow \{1, \dots, K\}$. When $y = f(\mathbf{x})$, the model assigns an input described by the vector \mathbf{x} to a category identified by the numeric code y , called the *output* or *response*. In other variants of the classification task, f may output a probability distribution over the possible classes.

Regression is the other main learning task and requires the algorithm to predict a continuous value given some input. This task requires a function

$f : \mathbb{R}^p \rightarrow \mathbb{R}$, where the only difference to classification is the format of its output.

Learning algorithms can learn to perform such tasks by observing a relevant set of data points, *i.e.* a dataset. A dataset containing N observations of p features is commonly described as a design matrix $X : N \times p$, where each row of the matrix represents a different observation and each column corresponds to a different feature of the observations, *i.e.*

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{Np} \end{bmatrix}.$$

Often the dataset includes annotations for each observation in the form of a label (classification) or a target value (regression). The N annotations are represented by the vector \mathbf{y} , where element y_i is associated with the i -th row of X . Therefore the response vector may be denoted by

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}.$$

Note that in the case of multiple labels or targets, a matrix representation $Y : N \times K$ is required.

Statistical learning algorithms can be divided into two main categories, *supervised* and *unsupervised* algorithms, determined by the presence (or absence) of annotations in the dataset to be analysed. Unsupervised learning algorithms learn from data consisting only of features, X , and are used to find useful properties and structure in the dataset (see Hastie *et al.*, 2009, Ch. 14). On the other hand, supervised learning algorithms learn from datasets which consist of both features and annotations, (X, Y) , with the aim to model the relationship between them. Therefore, both classification and regression are considered to be supervised learning tasks.

In order to evaluate the ability of a learning algorithm to perform its assigned task, we have to design a quantitative performance measure. For example, in a classification task we are usually interested in the accuracy of the algorithm, *i.e.* the percentage of times that the algorithm makes the correct

classification. We are mostly interested in how well the learning algorithm performs on data that it has not seen before, since this demonstrates how well it will perform in real-world situations. Thus we evaluate the algorithm on a *test set* of data points, independent of the *training set* of data points used during the learning process.

For a more concrete example of supervised learning, and keeping in mind that the linear model is one of the main building blocks of neural networks, consider the learning task underlying *linear regression*. The objective here is to construct a system which takes a vector $\mathbf{x} \in \mathbb{R}^p$ as input and predicts the value of a scalar $y \in \mathbb{R}$ in response. In the case of linear regression, we assume the output be a linear function of the input. Let \hat{y} be the predicted response. We define the output to be

$$\hat{y} = \hat{\mathbf{w}}^T \mathbf{x},$$

where $\hat{\mathbf{w}} = [w_0, w_1, \dots, w_p]$ is a vector of parameters and $\mathbf{x} = [1, x_1, x_2, \dots, x_p]$. Note that an intercept is included in the model (also known as a *bias* in machine learning). The parameters are values that control the behaviour of the system. We can think of them as a set of *weights* that determine how each feature affects the prediction. Hence the learning task can be defined as predicting y from \mathbf{x} through $\hat{y} = \hat{\mathbf{w}}^T \mathbf{x}$.

We of course need to define a performance measure to evaluate the linear predictions. For a set of observations, an evaluation metric tells us how (dis)similar the predicted output is to the actual response values. A very common measure of performance in regression is the *mean squared error* (MSE), given by

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2.$$

The process of learning from the data (or fitting a model to the data) can be reduced to the following optimisation problem: find the set of weights, $\hat{\mathbf{w}}$, which produces a $\hat{\mathbf{y}}$ that minimises the MSE. Of course this problem has a closed form solution and can quite trivially be found by means of *ordinary least squares* (OLS) (see Hastie *et al.*, 2009, p. 12). However, we have mentioned that we are more interested in the algorithm's performance evaluated on a test set. Unfortunately the least squares solution does not guarantee the solution

to be optimal in terms of the MSE on a test set, rendering statistical learning to be much more than a pure optimisation problem.

The ability of a model to perform well on previously unobserved inputs is referred to as its *generalisation* ability. Generalisation is the key challenge of statistical learning. One way of improving the generalisation ability of a linear regression model is to modify the optimisation criterion J , to include a *weight decay* (or *regularisation*) term. That is, we want to minimise

$$J(\mathbf{w}) = MSE_{\text{train}} + \lambda \mathbf{w}^T \mathbf{w},$$

where $J(\mathbf{w})$ now expresses preference for smaller weights. The parameter λ is non-negative and needs to be specified ahead of time. It controls the strength of the preference by determining how much influence the penalty term, $\mathbf{w}^T \mathbf{w}$, has on the optimisation criterion. If $\lambda = 0$, no preference is imposed, and the solution is equivalent to the OLS solution. Larger values of λ force the weights to decrease, and thus referred to as a so-called *shrinkage* method ((cf. for example Hastie *et al.*, 2009, pp. 61-79) and Goodfellow *et al.* (2016)).

We can further generalise linear regression to the classification scenario. First, note the different types of classification schemes. Consider \mathcal{G} , the discrete set of values which may be assumed by G , where G is used to denote a categorical output variable (instead of Y). Let $|\mathcal{G}| = K$ denote the number of discrete categories in the set \mathcal{G} . The simplest form of classification is known as binary classification and refers to scenarios where the input is associated with only one of two possible classes, *i.e.* $K = 2$. When $K > 2$, the task is known as multiclass classification. In multi-label classification an input may be associated with multiple classes (out of K available classes), where the number of classes that each observation belongs to, is unknown. A thorough discussion of MLC methods is given in ???. Here we start by introducing the two single label classification setups, *viz.* binary and multiclass classification.

In multiclass classification, given the input values \mathbf{X} , we would like to accurately predict the output, G , which we denote by \hat{G} . One approach would be to represent G by an indicator vector $\mathbf{Y}_G : K \times 1$, with elements all zero except in the G -th position, where it is assigned a 1, *i.e.* $Y_k = 1$ for $k = G$ and $Y_k = 0$ for $k \neq G$, $k = 1, 2, \dots, K$. We may then treat each of the elements in \mathbf{Y}_G as quantitative outputs, and predict values for them, denoted by $\hat{\mathbf{Y}} = [\hat{Y}_1, \dots, \hat{Y}_K]$. The class with the highest predicted value will then be the final categorical prediction of the classifier, *i.e.* $\hat{G} = \arg \max_{k \in \{1, \dots, K\}} \hat{Y}_k$.

Within the above framework we therefore seek a function of the inputs which is able to produce accurate predictions of the class scores, *i.e.*

$$\hat{Y}_k = \hat{f}_k(\mathbf{X}),$$

for $k = 1, \dots, K$. Here \hat{f}_k is an estimate of the true function, f_k , which is meant to capture the relationship between the inputs and output of class k . As with the linear regression case described above, we can use a linear model $\hat{f}_k(\mathbf{X}) = \hat{\mathbf{w}}_k^T \mathbf{X}$ to approximate the true function. The linear model for classification divides the input space into a collection of regions labelled according to the classification, where the division is done by linear *decision boundaries* (see ?? for an illustration). The decision boundary between classes k and l is the set of points for which $\hat{f}_k(\mathbf{x}) = \hat{f}_l(\mathbf{x})$. These set of points form an affine set or hyperplane in the input space.

After the weights are estimated from the data, an observation represented by \mathbf{x} (including the unit element) can be classified as follows:

- Compute $\hat{f}_k(\mathbf{x}) = \hat{\mathbf{w}}_k^T \mathbf{x}$ for all $k = 1, \dots, K$.
- Identify the largest component and classify to the corresponding class, *i.e.* $\hat{G} = \arg \max_{k \in \{1, \dots, K\}} \hat{f}_k(\mathbf{x})$.

One may view the predicted class scores as estimates of the conditional class probabilities (or posterior probabilities), *i.e.* $P(G = k | \mathbf{X} = \mathbf{x}) \approx \hat{f}_k(\mathbf{x})$. However, these values are not the best estimates of posterior probabilities. Although the values sum to 1, they do not lie within $[0, 1]$. A way to overcome this problem is to estimate the posterior probabilities using the *logit transform* of $\hat{f}_k(\mathbf{x})$. That is,

$$P(G = k | \mathbf{X} = \mathbf{x}) \approx \frac{e^{\hat{f}_k(\mathbf{x})}}{\sum_{l=1} e^{\hat{f}_l(\mathbf{x})}}.$$

Through this transformation, the estimates of the posterior probabilities both sum to 1 and are squeezed into $[0, 1]$. The above model is the well-known *logistic regression* model (Hastie *et al.*, 2009, p. 119). With this formulation there is no closed form solution for the weights. Instead, the weight estimates may be searched for by maximising the log-likelihood function. One way of doing this is by minimising the negative log-likelihood using gradient descent, which will be discussed in the following section.

Finally in this section, note that any supervised learning problem can also be viewed as a function approximation problem. Suppose we are trying to predict a variable Y given an input vector \mathbf{X} , where we assume the true relationship between them to be given by

$$Y = f(\mathbf{X}) + \epsilon,$$

where ϵ represents the part of Y that is not predictable from \mathbf{X} , because of, for example, incomplete features or noise present in the labels. Then in function approximation we are estimating f with an estimate \hat{f} . In parametric function approximation, for example in linear regression, estimation of $f(\mathbf{X}, \theta)$ is equivalent to estimating the optimal set of weights, $\hat{\theta}$. In the remainder of the thesis, we refer to \hat{f} as the *model*, *classifier* or *learner*.

1.5 Outline

Give an outline of what will be discussed where in the thesis.

Chapter 2

Neural Networks

Chapter 3

Neural Networks

3.1 Introduction

A Neural Network (NN), like any other machine learning model, is a function that maps inputs to outputs, *i.e.*

$$f : \mathbf{x} \rightarrow y.$$

The NN, f , receives input, \mathbf{x} , and produces output, y . What happens inside of f is loosely based on biological neural systems, or the brain. The brain consists of a collection of interconnected neurons, each sending and receiving signals between each other. An artificial NN tries to copy this structure by modelling what happens inside of a single neuron by outputting a weighted combination of its inputs, combined with a simple non-linear transformation. The output of a neuron is referred to as activations. These neurons are grouped in so-called layers. At each layer the input is passed through each of the neurons and their activations, then in turn, gets passed to the next layer. See Figure 3.1 for an illustration of this structure. A more detailed explanation of the structure of a NN is given in section 3.2

The transformation at each neuron is controlled by a set of parameters, also known as weights. These weights can be tuned to obtain a desired output. When training a NN to perform a certain machine learning task, for instance classification, the NN is fed a bunch of data and tweaks its weights so that the resulting output matches the true target as close as possible. This process of tweaking the weights according to the data is done by an optimisation algorithm called Stochastic Gradient Descent (SGD). SGD and NN training is covered in detail in section 3.3.

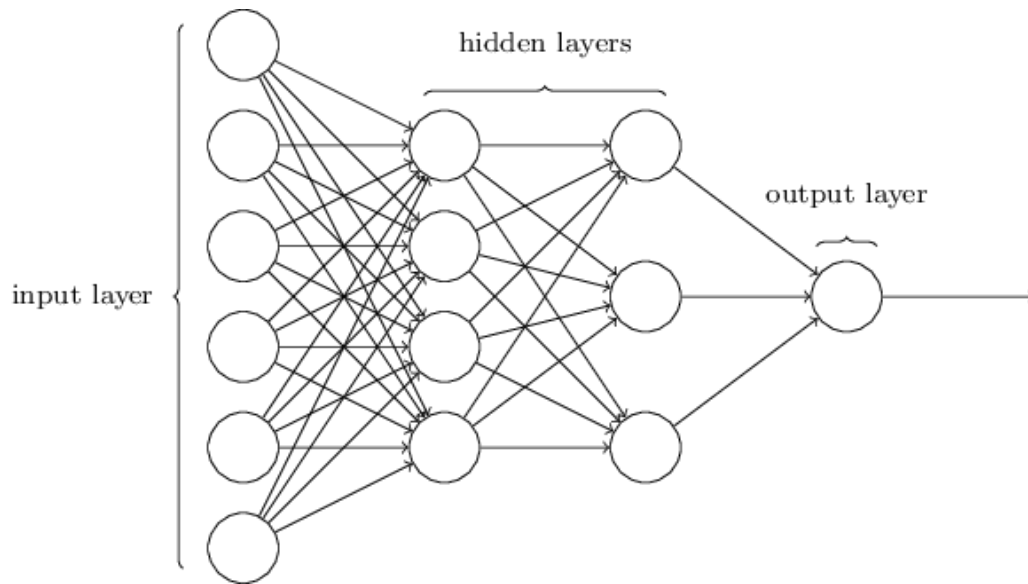


Figure 3.1: The structure of an artificial neural network (This is still a placeholder).

There has been plenty of excitement around NNs recently, but in fact NNs have quite a bit of history. The development of NNs dates at least as far back as Perceptrons in (Rosenblatt, 1962). It is also interesting to compare modern NNs with the Projection Pursuit Regression algorithm (Friedman and Stuetzle, 1981) developed in statistics. Only recently a series of breakthroughs allowed NNs to be more efficient and effective and therefore the revitalisation of the field.

The modular nature of a NN allows it to accept inputs and produce outputs of various shapes and sizes. Therefore NNs can be used for just about any machine learning task, from doing simple binary classification on tabular data, to generating full color images from black and white sketches. Modern structures like the Convolutional Neural Network and the Recurrent Neural Network are all based on the vanilla NN structure and training procedure explored in the rest of this chapter.

3.2 The Structure of a Neural Network

Recall, a NN processes an input by sending it through a series of layers, each applying some transformation to its input, to eventually produce an output and each layers consists of smaller computational units, called neurons. To understand and formulate the NN structure, we will start by describing the

operation inside a single neuron and then gradually put the pieces together to form layers and then a complete NN. Suppose we want a function that estimates a taxi fare given the distance travelled, duration of the trip and number of passengers. A single neuron can act as such a function by taking a weighted average of these three inputs to produce an estimate of the taxi fare. ?? is a graphical representation of this function. In equation form, this function can be written as:

$$w_1 \cdot \text{distance} + w_2 \cdot \text{time} + w_3 \cdot \text{passengers} + b = \text{fare},$$

where w_i , $i = \{1, 2, 3\}$, are the weights applied to each of the inputs and b a constant added to the equation, better known as the bias term in machine learning. Clearly, this equation is simply the very common linear model and thus also can be written as:

$$\mathbf{x}^\top \mathbf{w} + b = z,$$

where $\mathbf{x} = [\text{distance} \quad \text{time} \quad \text{passengers}]^\top$ is the input, $\mathbf{w} = [w_1 \quad w_2 \quad w_3]^\top$ the weights and z the output, *i.e.* the taxi fare. For convenience, we sometimes compress the above equation to $\mathbf{x}^\top \mathbf{w} = z$, where \mathbf{x} includes the bias term and the weight vector \mathbf{w} a unit element, *i.e.* $\mathbf{x} = [b \quad \text{distance} \quad \text{time} \quad \text{passengers}]^\top$ is the input, $\mathbf{w} = [1 \quad w_1 \quad w_2 \quad w_3]^\top$.

The weights determine how much each of the inputs contribute to the fare. For example, the distance (in km's) may be the most important driver of the taxi fare but the duration of the trip (in minutes) has little influence and the number passengers has no effect. Then the weights may look something like this:

$$w_1 = 10, \quad w_2 = 0.5 \quad \text{and} \quad w_3 = 0.$$

But we do not know what these weights are before hand and therefore need to estimate them. With the classical linear model, these weights (or coefficients) are estimated using the ordinary least squares (OLS) method. Since a NN consists of many inter-connected neurons, the OLS methods will not suffice. This is the topic of the next section.

Suppose a single neuron (or a linear model if you like) is not flexible enough to model the taxi fare given the distance, time and number of passengers. Now we decide to add another neuron. This neuron also accepts the same inputs as the first, but uses a different set of weights to estimate the fare. Now we have

two neurons, each producing a different output:

$$\mathbf{x}^\top \mathbf{w}_1 = z_1 \quad \text{and} \quad \mathbf{x}^\top \mathbf{w}_2 = z_2.$$

So how do we get a final estimate of the fare from these two initial estimate? We feed it to another neuron of course, *i.e.*

$$\mathbf{z}^\top \mathbf{w}_3 = y$$

See ?? for a graphical representation.

The first two neurons both took in the distance, time and passengers as input and produced a single output. These operations can be expressed as a single equation, *i.e.*

$$\mathbf{x}^\top W = \mathbf{z}^\top,$$

where

$$W = [\mathbf{w}_1 \quad \mathbf{w}_2] = \begin{bmatrix} 1 & 1 \\ w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \end{bmatrix} \quad \text{and} \quad \mathbf{z} = [z_1 \quad z_2]^\top.$$

The collection of these two neurons is what is called a layer. Since our third neuron (which is also a layer but with a single neuron) takes the output of this layer as input, it is possible to express the complete input-output relationship in one equation, *i.e.*

$$\mathbf{z}^\top \mathbf{w}_3 = \mathbf{x}^\top W \mathbf{w}_3 = y.$$

Note here that the weights from the first layer, W , and the third neuron, \mathbf{w}_3 , can collapse into a single vector \mathbf{w} , effectively reducing all of the neuron operations back into a single neuron representation and thus is clearly not a good way to model a network

However, a NN has a way to prevent this collapsing from happening and to allow for non-linear relationships between the inputs and outputs. It does this through the use of an activation function, a simple non-linear transformation. An activation is applied after each linear layer. So now the NN equation can be represented as:

$$a_2(a_1(\mathbf{x}^\top W) \mathbf{w}_3) = y,$$

Where a_1 is the activation function after the first linear layer and a_2 the activation after the final layer.

By introducing the non-linear activations, it greatly enlarges the class of functions that can be approximated by the network.

TBC

The activation function, $a(\cdot)$, was usually chosen to be the sigmoid function, $a(v) = \frac{1}{1+e^{-v}}$

In the previous section, we introduced activation functions, which are simple non-linear functions of its input. These are usually applied after a fully connected layer (linear transformation) and are crucial for the flexibility of a deep neural network. We also mentioned that the sigmoid activation, which was originally the go-to activation, is currently not the most popular choice. Another activation function originally thought to work well was, $a(x) = \tanh(x)$. However, by far the most common activation function used at the time of writing is the Rectified Linear Units (ReLU) non-linearity. Its definition is much simpler than its name and is defined as $a(x) = \max(0, x)$. It was introduced in (Krizhevsky *et al.*, 2012) and they showed that using ReLUs in their CNNs reduced the number of training iterations to reach the same point by a factor of 6 compared to using $\tanh(x)$. The ReLU limits the gradient vanishing problem as its derivative is always one when x is positive. Gradient vanishing problem?

There are a plethora of proposals for activation functions, since any simple non-linear (differentiable?) function can be used. Some of the recent most popular choices are exponential linear units (ELUs) (Clevert *et al.*, 2015) and scaled exponential linear units (SELUs) (Klambauer *et al.*, 2017). The choice of activation function usually influences the convergence time and some might protect the training procedure from overfitting in some cases. The different activation functions can be experimented with, however it would be sufficient in most cases to use ReLUs. The other mentioned proposals have inconsistent gains over ReLUs and therefore it remains the standard choice.

However, very recently (Ramachandran *et al.*, 2017) used automated search techniques to discover novel activation functions. The exhaustive and reinforcement learning based searched identified a few promising novel activation functions on which the authors then did further empirical evaluations. They found that the so-called *Swish* activation function,

$$a(x) = x \cdot \sigma(\beta x),$$

where β is a constant (can also be a trainable parameter), gave the best empirical results. It consistently matched or outperformed ReLU's on deep

networks applied to the domains of image classification and machine translation.

The number of units in the hidden layer, M , is also a value to be decided on. Too few units will not allow the network enough flexibility to model complex relationships and too many takes longer to train and increases the chance of overfitting. M is mostly chosen by experimentation. A good starting point would be to choose a large value and training the network with regularisation (discussed shortly).

The difference between the above discussed neural networks and current state-of-the-art deep learning methods, is the number and type of hidden layers. The following section discusses the popular activation functions used in DNNs.

The units in \mathbf{Z} are called hidden since they are not directly observed. The aim of this transformation is to derive features, \mathbf{Z} , so that the classes become linearly separable in the derived feature space (Lecun *et al.*, 2015). Many more of these hidden layers (combination of linear and non-linear transformations) can be used to derive features to input into the final classifier. This is what we refer to as deep neural networks (DNNs) or deep learning methods.

- comment on number and size of layers
- lead into modern architectures
- lead into parameter optimisation

3.3 Training a Neural Network

3.3.1 Optimisation

As mentioned before, fitting a linear regression model can be reduced to finding the optimal weights to minimise the MSE function (with or without weight decay). In fact, typically model training procedures can be described as the search for its internal parameters that minimises or maximises some *objective function*. Therefore statistical learning and optimisation are closely related. Optimisation refers to the task of either minimising or maximising some function $J(x)$ by altering x . The function we want to optimise is called the objective function. When we are minimising the objective function, we may also refer to the objective function as the *cost* or *loss function*. These terms will be used interchangeably throughout the remainder of the thesis.

As mentioned in the previous section, parameter estimation (or optimisation) of a linear (or logistic regression) model is usually done using OLS or maximum

likelihood estimation (MLE). In this section, however, we discuss an alternative parameter estimation method which is also relevant for the optimisation of neural networks.

Consider the MSE loss function:

$$\begin{aligned} L &= \sum_{i=1}^N L_i \\ &= \sum_{i=1}^N \sum_{k=1}^K (y_{ik} - f_k(\mathbf{x}_i))^2 \\ &= \sum_{i=1}^N \sum_{k=1}^K (y_{ik} - \mathbf{w}_k^T \mathbf{x}_i)^2, \end{aligned}$$

where $f_k(\cdot)$ in this case is the linear model used to predict the k -th class posterior probability. Although the MSE loss is mostly used in a regression setup and not really well suited for classification, we make use of it here for illustration purposes.

To find the weights, \mathbf{w} , that minimise L , we can follow a process of iterative refinement. That is, starting with a random initialisation of \mathbf{w} , one iteratively updates the values such that L decreases. The updating steps are repeated until the loss converges. In order to minimise L with respect to \mathbf{w} , we calculate the gradient of the loss function at the point $L(\mathbf{x}; \mathbf{w})$. The gradient (or slope) of the loss function indicates the direction in which the function has the steepest rate of increase. Therefore, once we have determined this direction, we can update the weights by a step in the opposite direction - thereby reaching a smaller value of L .

The gradient of L_i is computed by obtaining the partial derivative of L_i with respect to \mathbf{w}_k , *i.e.*:

$$\frac{\partial L_i}{\partial \mathbf{w}_k} = -2(y_{ik} - \mathbf{w}_k^T \mathbf{x}_i) \mathbf{x}_i.$$

After obtaining the above N partial derivatives, an update at the $(r + 1)$ -th iteration may be obtained as follows:

$$\mathbf{w}_k^{(r+1)} = \mathbf{w}_k^{(r)} - \gamma \sum_{i=1}^n \frac{\partial L_i}{\partial \mathbf{w}_k^{(r)}},$$

where γ is called the *learning rate* and determines the size of the step taken toward the optimal direction. One typically would like to set the learning rate small enough so that one does not overshoot the minimum, but large

enough to limit the number of iterations before convergence. This value can be determined via a line search but is not always ideal since this may render the training time of DNNs too long. Another option is to reduce the learning rate after every fixed number of iterations. More detail regarding the implication of the learning rate will be given in ??.

The procedure of repeatedly evaluating the gradient of the objective function and then performing a parameter update, is called *gradient descent* [Cauchy, 1847]. Gradient descent forms the basis of the optimisation procedure for neural networks.

Note that a weight update is made by evaluating the gradient over a set of observations, $\{\mathbf{x}_i, i = 1, \dots, n\}$. One of the advantages of gradient descent is that at an iteration, the gradient need not be computed over the complete training dataset, *i.e.* $n \leq N$. When updates are iteratively determined by using subsets of the data, the process is called *mini-batch gradient descent*. This is extremely helpful in large-scale applications, since it obviates computation of the full loss function over the entire dataset. This leads to faster convergence, because of more frequent parameter updates, and allows processing of data sets that are too large to fit into a computer's memory. The choice regarding batch size depends on the available computation power. Typically a batch consists of 64, 128 or 256 data points, since in practice many vectorised operation implementations work faster when their inputs are sized in powers of 2. The gradient obtained using mini-batches is only an approximation of the gradient of the full loss but it seems to be sufficient in practice (Li *et al.*, 2014). Note at this point that the collection of iterations needed to make one sweep through the training data set is called an *epoch*.

The extreme case of mini-batch gradient descent is when the batch size is selected to be 1. This is called *Stochastic Gradient Descent* (SGD). Recently SGD has been used much less, since it is more efficient to calculate the gradient in larger batches compared to only using one example. However, note that it remains common to use the term SGD when actually referring to mini-batch gradient descent. Gradient descent in general has often been regarded as slow or unreliable but it works well for optimising DNNs. SGD will most probably not find even a local minimum of the objective function. It typically however finds a very low value of the cost function quickly enough to be useful.

3.3.2 Optimisation Example

To illustrate the SGD algorithm, consider the linear model in a classification context. Suppose we are given a training data set with two-dimensional inputs and only two possible classes. Let the data be generated in the same way as described in (Hastie *et al.*, 2009, pp. 16-17).

We want to fit a linear regression model to the data such that we can classify an observation to the class with the highest predicted score. In the binary case it is only necessary to model one class probability and then assign an observation to that class if the score exceeds some threshold (usually 0.5), otherwise it is assigned to the other class. Therefore the decision boundary is given by $\{\mathbf{x} : \mathbf{x}^T \hat{\mathbf{w}} = 0.5\}$.

The example is illustrated in ???. The colour shaded regions represent the parts of the input space classified to the respective classes, as determined by the decision boundary based upon OLS parameter estimates. Gradient descent was applied to determine the optimal weights using a learning rate of 0.001. Since the total number of training observations are small, it is not necessary to use SGD. In ??, the dashed lines represent the decision boundary defined by the gradient descent parameter estimates at different iterations. We observe that initially the estimated decision boundary is far from the OLS solution, but as the update iterations proceed, the decision boundary is rotated and translated until finally matching the OLS line. It took 29 iterations for the procedure to reach convergence.

3.3.3 Backpropagation

In Section 3.3.1 we discussed how to fit a linear model using the Stochastic Gradient Descent optimisation procedure. Currently, SGD is the most effective way of training deep networks. To recap, SGD optimises the parameters θ of a networks to minimise the loss,

$$\theta = \arg \min_{\theta} \frac{1}{N} \sum_{i=1}^N l(\mathbf{x}_i, \theta).$$

With SGD the training proceeds in steps and at each step we consider a mini-batch of size $n \leq N$ training samples. The mini-batch is used to approximate the gradient of the loss function with respect to the parameters by computing,

$$\frac{1}{n} \frac{\partial l(\mathbf{x}_i, \theta)}{\partial \theta}.$$

Using a mini-batch of samples instead of one at a time produces a better estimate of the gradient over the full training set and it is computationally much more efficient.

This section discusses the same procedure, but applied to a simple single hidden layer neural network. This is made possible by the *backpropagation* algorithm. Note, this process extends naturally to the training of deeper networks.

The neural network described in the previous section has a set of unknown adjustable weights that defines the input-output function of the network. They are the $\alpha_{0m}, \boldsymbol{\alpha}_m$ parameters of the linear function of the inputs, \mathbf{X} , and the $\beta_{0k}, \boldsymbol{\beta}_k$ parameters of the linear transformation of the derived features, \mathbf{Z} . Denote the complete set of parameters by θ . Then the objective function for regression can be chosen as the sum-of-squared-errors:

$$L(\theta) = \sum_{k=1}^K \sum_{i=1}^N (y_{ik} - f_k(\mathbf{x}_i))^2$$

and for classification, the cross-entropy:

$$L(\theta) = - \sum_{i=1}^N \sum_{k=1}^K y_{ik} \log f_k(\mathbf{x}_i),$$

with corresponding classifier $G(\mathbf{x}) = \arg \max_k f_k(\mathbf{x})$. Since the neural network for classification is a linear logistic regression model in the hidden units, the parameters can be estimated by maximum likelihood. According to Hastie *et al.* (2009, p. 395), the global minimiser of $L(\theta)$ is most likely an overfit solution and we instead require regularisation techniques when minimising $L(\theta)$.

Therefore, one rather uses gradient descent and backpropagation to minimise $L(\theta)$. This is possible because of the modular nature of a neural network, allowing the gradients to be derived by iterative application of the chain rule for differentiation. This is done by a forward and backward sweep over the network, keeping track only of quantities local to each unit.

In detail, the backpropagation algorithm for the sum-of-squared error objective function,

$$\begin{aligned}
L(\theta) &= \sum_{i=1}^N L_i \\
&= \sum_{i=1}^N \sum_{k=1}^K (y_{ik} - f_k(\mathbf{x}_i))^2,
\end{aligned}$$

is as follows. The relevant derivatives for the algorithm are:

$$\begin{aligned}
\frac{\partial L_i}{\partial \beta_{km}} &= -2(y_{ik} - f_k(\mathbf{x}_i))g'_k(\beta_k^T \mathbf{z}_i)z_{mi}, \\
\frac{\partial L_i}{\partial \alpha_{ml}} &= -\sum_{k=1}^K 2(y_{ik} - f_k(\mathbf{x}_i))g'_k(\beta_k^T \mathbf{z}_i)\beta_{km}\sigma'(\alpha_m^T \mathbf{x}_i)x_{il}.
\end{aligned}$$

Given these derivatives, a gradient descent update at the $(r+1)$ -th iteration has the form,

$$\begin{aligned}
\beta_{km}^{(r+1)} &= \beta_{km}^{(r)} - \gamma_r \sum_{i=1}^N \frac{\partial L_i}{\partial \beta_{km}^{(r)}}, \\
\alpha_{ml}^{(r+1)} &= \alpha_{ml}^{(r)} - \gamma_r \sum_{i=1}^N \frac{\partial L_i}{\partial \alpha_{ml}^{(r)}},
\end{aligned}$$

where γ_r is called the learning rate. Now write the gradients as

$$\begin{aligned}
\frac{\partial L_i}{\partial \beta_{km}} &= \delta_{ki}z_{mi}, \\
\frac{\partial L_i}{\partial \alpha_{ml}} &= s_{mi}x_{il}.
\end{aligned}$$

The quantities, δ_{ki} and s_{mi} are errors from the current model at the output and hidden layer units respectively. From their definitions, they satisfy the following,

$$s_{mi} = \sigma'(\alpha_m^T \mathbf{x}_i) \sum_{k=1}^K \beta_{km} \delta_{ki},$$

which is known as the backpropagation equations. Using this, the weight updates can be made with an algorithm consisting of a forward and a backward pass over the network. In the forward pass, the current weights are fixed and the predicted values $\hat{f}_k(\mathbf{x}_i)$ are computed. In the backward pass, the errors δ_{ki} are computed, and then backpropogated via the backpropagation equations to give obtain s_{mi} . These are then used to update the weights.

Backpropagation is simple and its local nature (each hidden unit passes only information to and from its connected units) allows it to be implmented efficiently in parallel. The other advantage is that the computation of the

gradient can be done on a batch (subset of the training set) of observations. This allows the network to be trained on very large datasets. One sweep of the batch learning through the entire training set is known as an epoch. It can take many training epochs for the objective function to converge.

3.3.4 Learning Rate

The convergence times also depends on the learning rate, γ_r . There are no easy ways for determining γ_r . A small learning rate slows down the training time, but is safer against overfitting and overshooting the optimal solution. With a large learning rate, convergence will be reached quicker, but the optimal solution may not have been found. One could do a line search of a range of possible values, but this usually takes too long for bigger networks. One possible strategy for effective training is to decrease the learning rate every time after a certain amount of iterations.

Recently, in (<https://arxiv.org/abs/1711.00489>) (no bibtex entry), the authors found that, instead of learning rate decay, one can alternatively increase the batch size during training. They found that this method reaches equivalent test accuracies compared to learning rate decay after the same amount of epochs. But their method requires fewer parameter updates.

3.3.5 Basic Regularisation

There are many ways to prevent overfitting in deep neural networks. The simplest strategies for single hidden layer networks are by early stopping and weight decay. Stopping the training process early can prevent overfitting. When to stop can be determined by a validation set approach. Weight decay is the addition of a penalty term, $\lambda J(\theta)$, to the objective function, where,

$$J(\theta) = \sum_{km} \beta_{km}^2 + \sum_{ml} \alpha_{ml}^2.$$

This is exactly what is done in ridge regression (Hastie *et al.*, 2009, Ch. 4). $\lambda \geq 0$ and larger values of λ tends to shrink the weights towards zero. This helps with the generalisation ability of a neural network, but recently more effective techniques to combat overfitting in DNNs have been developed. These are dicussed in ??.

It is common to standardise all inputs to have mean zero and standard deviation of one. This ensures that all input features are treated equally. Now we have covered all of the basics for simple (1-layer) neural networks.

- move regularisation to next chapter
- lead into modern learning policies
- lead into what it is learning

3.4 Representation Learning

- What is the Neural Network actually doing?
- See (Bengio *et al.*, 2013)

Each layer of the network is trained to produce a higher-level representation of the observed patterns, based on the data it receives as input from the layer below, by optimizing an objective function. Every level produces a representation of the input pattern that is more abstract than the previous level because it is obtained by composing more non-linear operations.

The most central idea in deep learning is that of representation. Traditionally, input features to a machine learning algorithm must be hand-crafted from raw data, relying on practitioner expertise and domain knowledge to determine explicit patterns of prior interest. The engineering process of creating, analyzing, selecting, and evaluating appropriate features can be laborious and time consuming, and is often thought of as a "black art" requiring creativity, trial-and-error, and oftentimes luck.

In contrast, deep learning techniques learn optimal features directly from the data itself, without any human guidance, allowing for the automatic discovery of latent data relationships that might otherwise be unknown or hidden.

That being said, preprocessing of data in deep learning is not totally free of human engineering. See in the Normalisation section what measures should be taken to ensure a NN can learn effectively from tabular data.

Also manifold learning. Show PCA of different level activations and show how the classes get separated during training. The feature extraction layers transforms the inputs, which are not separable in the input space, to a representation where they are separable. A sufficiently detailed architecture can separate nearly any arrangements of points. PCA vs AE. If non-linear, find

any surface manifold, in contrast to hyperplane without. AEs can only learn to represent data that it has actually seen before, not necessarily out of sample space.

3.5 Vanilla MLP on Tabular Data

- run through an example of a simple MLP
- acts as a baseline
- identify its drawbacks
- how the next chapters will build on it

Chapter 4

Deep Learning

4.1 Introduction

- Modern Neural Network Approaches
- story of combatting overfitting
- Recent advancements in deep learning which could be useful to applying in tabular data

4.2 Autoencoders

An autoencoder takes an input and first transforms it into some (smaller) latent representation using the part of the network called the encoder. From the latent representation the second part of the network, called the decoder, tries to reconstruct the input by doing some transformation. Both the encoder and the decoder networks are NNs in their own right and thus usually consist of either fully connected layers or convolutional layers (or both).

During training a reconstruction loss is minimised. A reconstruction loss measures the distance between the reconstruction of the input based on the latent representation and the actual input.

Autoencoders technically belong to the self (or semi) supervised class of methods, although many still think of it as unsupervised. It is unsupervised in the sense that it does not require labelling, but it is still supervised in the sense that it predicts an output; the input and thus self-supervised.

A denoising autoencoder (DAE) is a variant of the vanilla autoencoder. A DAE also learns to reconstruct the input vector, but in this case from a

corrupted version thereof. So during training, before an input is sent through the encoder, it first get injected with random noise. However, the output of the decoder is still being compared to the original input and thus the DAE is supposed to learn how to remove noise from the input - therefore, denoising.

(Miotto *et al.*, 2016) used a stacked denoising autoencoder to learn patient representations from EHR data. They found that these representations were useful features in predicting future health states of patients. By using these learned representations as input significantly improved the performance of predictive models compared to those only using the raw inputs.

See also (Vincent *et al.*, 2008).

<https://arxiv.org/pdf/1803.09820.pdf>

4.3 Combatting Overfitting

4.3.1 Transfer Learning

4.3.2 Dropout

Interesetingly, (Haldar *et al.*, 2018) found that dropout was not effective in their application. They pinned it down to dropout producing invalid input scenarios that distracted the model. Therefore they opted for hand crafted noise shapes taking into account the distribution of the relevant feature.

4.3.3 Data Augmentation

As mentioned before, our aim with predictive models is to generalise well to an unseen test set. In an ideal world we would train a model on all possible variations of the data to capture all interactions and relationships. This is not possible in the real world. Such a dataset is not available and would be infinitely large.

In reality we have a finite subset of the full data distribution to train on. Any new samples with unique feature combinations will likely improve the models generalisablity. If the collection of new samples is not available, we can try to artificially create more.

This is a standard approach especially in computer vision applications. For example, from a single image, we can rotate it, flip it horizontally, shift it any direction, crop it, and many other transformations without destroying the

semantic content of the image. But by doing so we are artificially increasing the size of the training set to help with overfitting. Of course this is not as effective as genuine new data samples, but it is a very effective and efficient substitute (Perez and Wang, 2017).

4.4 Modern Architectures

4.4.1 Normalisation

The batch normalisation layer attempts to normalise neuron activations to zero mean and unit variance (Ioffe and Szegedy, 2015). It has become the standard when training deep CNNs. Training with normalisation techniques is perturbed by Stochastic Gradient Descent (SGD), stochastic regularisation (like dropout) and the estimation of the normalisation parameters. Both RNNs and CNNs can stabilise learning via weight sharing, therefore they are less prone to perturbations. Fully-connected NNs do not have this luxury and shows high variance in the training error when trained with normalisation techniques.

- batchnorm
- selu?

4.4.2 Residual Networks

- resnet and densenet

Residual Networks became very popular after it was used to win one of the ImageNet competitions (He *et al.*, 2015). The residual connection layer can simply be formalised as

$$y = F(x) + x,$$

i.e. combining the input to the layer(s) with the output of the layer(s). Here, the combination is by addition, but other ways can also be used, like multiplication or concatenation. These layers are very useful when training deeper neural networks since they encourage gradient flow through internal layers.

4.4.3 Embeddings

Like word embeddings. Learnable mapping of an item to a vector.

4.4.4 Attention

First proposed in neural machine translation (Bahdanau *et al.*, 2014) and now almost used ubiquitously in natural language processing applications.

4.5 One-cycle Policy

- can include experiment here

(Smith, 2018)

- reduce training time and increase performance

Currently the process of setting the hyper-parameters, including designing the network architecture, requires expertise and extensive trial and error and is based more on serendipity than science.

Currently there are no simple and easy ways to set hyper-parameters – specifically, learning rate, batch size, momentum, and weight decay. Grid search or random search is expensive. Optimal parameters make a huge difference in training time and performance.

Look for clues of overfitting and underfitting to determine best parameters.

The experiments discussed herein indicate that the learning rate, momentum, and regularization are tightly coupled and optimal values must be determined together.

by monitoring validation/test loss early in the training, enough information is available to tune the architecture and hyper-parameters and this eliminates the necessity of running complete grid or random searches.

Underfitting is when the machine learning model is unable to reduce the error for either the test or training set. The cause of underfitting is an under capacity of the machine learning model; that is, it is not powerful enough to fit the underlying complexities of the data distributions. Overfitting happens when the machine learning model is so powerful as to fit the training set too well and the generalization error increases.

The takeaway is that achieving the horizontal part of the test loss is the goal of hyper- parameter tuning

The art of setting the network’s hyper-parameters amounts to ending up at the balance point between underfitting and overfitting

If the learning rate (LR) is too small, overfitting can occur. Large learning rates help to regularize the training but if the learning rate is too large, the training will diverge.

To use CLR, one specifies minimum and maximum learning rate boundaries and a stepsize. The stepsize is the number of iterations (or epochs) used for each step and a cycle consists of two such steps – one in which the learning rate linearly increases from the minimum to the maximum and the other in which it linearly decreases.

In the LR range test, training starts with a small learning rate which is slowly increased linearly throughout a pre-training run. This single run provides valuable information on how well the network can be trained over a range of learning rates and what is the maximum learning rate. When starting with a small learning rate, the network begins to converge and, as the learning rate increases, it eventually becomes too large and causes the test/validation loss to increase and the accuracy to decrease. The learning rate at this extrema is the largest value that can be used as the learning rate for the maximum bound with cyclical learning rates but a smaller value will be necessary when choosing a constant learning rate or the network will not begin to converge.

the amount of regularization must be balanced for each dataset and architecture

Contrary to this early work, this Section recommends using a larger batch size when using the 1cycle learning rate schedule, which is described in the above

Weight decay is one form of regularization and it plays an important role in training so its value needs to be set properly. The important point made above applies; that is, practitioners must balance the various forms of regularization to obtain good performance. the interested reader can see kuka et al. (2017) for a review of regularization methods.

1. Learning rate (LR): Perform a learning rate range test to a “large” learning rate. The max LR depends on the architecture (for the shallow 3-layer architecture, large is 0.01 while for resnet, large is 3.0), you might try more than one maximum. Using the 1cycle LR policy with a maximum learning rate determined from an LR range test, a minimum learning rate as a tenth of the maximum appears to work well but other factors are relevant, such as the rate of learning rate increase (too fast and increase

will cause instabilities).

2. Total batch size (TBS): A large batch size works well but the magnitude is typically constrained by the GPU memory. If your server has multiple GPUs, the total batch size is the batch size on a GPU multiplied by the number of GPUs. If the architecture is small or your hardware permits very large batch sizes, then you might compare performance of different batch sizes. In addition, recall that small batch sizes add regularization while large batch sizes add less, so utilize this while balancing the proper amount of regularization. It is often better to use a larger batch size so a larger learning rate can be used.
3. Momentum: Short runs with momentum values of 0.99, 0.97, 0.95, and 0.9 will quickly show the best value for momentum. If using the 1cycle learning rate schedule, it is better to use a cyclical momentum (CM) that starts at this maximum momentum value and decreases with increasing learning rate to a value of 0.8 or 0.85 (performance is almost independent of the minimum momentum value). Using cyclical momentum along with the LR range test stabilizes the convergence when using large learning rate values more than a constant momentum does.
4. Weight decay (WD): This requires a grid search to determine the proper magnitude but usually does not require more than one significant figure accuracy. Use your knowledge of the dataset and architecture to decide which values to test. For example, a more complex dataset requires less regularization so test smaller weight decay values. A shallow architecture requires more regularization so test larger weight decay values.

4.6 Model Interpretation

Although Deep Learning is now the state-of-the-art for many machine learning tasks, it still trails behind other algorithms in terms of model interpretability. But keep in mind this is not an unusual trade-off; between prediction performance and model interpretability. DNNs are occasionally referred to as “black boxes” since it is very difficult to interpret what is going on inside the stacks of linear and non-linear layers. This is one of deep learning’s greatest criticisms and is large reason why it cannot be used in some production environments.

onments. For example, in the clinical domain, model transparency is of utmost importance, given that predictions might be used to affect real-world medical decision-making and patient treatments (Shickel *et al.*, 2017).

Fortunately, some work has been done to gain insights from NNs.

4.6.1 Model Agnostic

4.6.1.1 Permutation Importance

(Haldar *et al.*, 2018) notes that the permutation test only produces sensical results on the assumption that the features are independent. Permuting the feature independently created examples that never occurred in real life, and the importance of features in that invalid space sent us in the wrong direction. The test however is somewhat useful in determining features that were not pulling their weight. If randomly permuting a feature did not affect the model performance at all, it was a good indication that the model is probably not dependent on it.

- Partial Dependence
- SHAP

4.6.2 Neural Network Specific

- Distilling Neural Networks, i.e. training a decision tree on train neural network generated data. <https://arxiv.org/pdf/1711.09784.pdf>
- Mimic learning (Che *et al.*, 2016)
- Plotting embeddings in lower dimensional space with PCA or t-sne
- evaluate which inputs get activated by a certain unit.
-

Chapter 5

Deep Learning for Tabular Data

5.1 Introduction

- core work
- combination of other work with tabular data and other domains
- not received much attention
- automate feature engineering

It is not exactly clear why DNNs are still in many cases inferior to gradient boosted trees when applied to tabular data, even though it outperforms all other algorithms in other application domains like text and speech. We can look for differences between tabular data and unstructured data in their properties to try and understand why this is the case. A difference between the two data types that stand out is the relative importance of each of the important features with respect to the target. In computer vision a large amount of pixels should change before an image is of something else. Whereas in tabular data a very small change in a single feature may have totally different behaviour with respect to the target (Shavitt and Segal, 2018). The same authors mention that this can be addressed by including a separate regularisation term for each of the weights in the network. These regularisation terms are seen as additional model hyperparameters. It is easy to see that this approach is totally intractable since the only way to train these hyperparameters are brute force and repetitive tweaking and validating (derivative free methods). A workaround is to make these regularisation parameters trainable like all of the other points in the network. This is achieved by minimising the counterfactual loss, a novel loss function proposed by (Shavitt and Segal, 2018). They found that training

NNs by optimising the counterfactual loss, outperform other regularisation approaches for NNs and results in NNs that are comparable to gradient boosted trees. The learned regularisation parameters can even help with interpreting feature importance.

- NNs proved to be useful for tabular data at AirBnB (?).

Deep Learning has set new records on various benchmarks and led to various commercial applications. Recurrent Neural Networks achieved new levels in speech and natural language processing and are already deployed on mobile devices. Their counterparts, Convolutional Neural Networks (CNNs), excel in vision tasks. CNNs are on par with human experts on detecting skin cancer. The latest self-driving cars rely on CNNs to understand video imagery. CNNs were also used in AlphaGo to evaluate board positions. (Klambauer *et al.*, 2017)

But on structured data, Deep Neural Networks have still a long way to go. Random Forests, Gradient Boosted Trees and Support Vector Machines often outperforms DNNs on structured data, both in terms of accuracy and efficiency.

DNNs on tabular data also struggle to go deeper than 3 or 4 layers.

Adopting NNs for tabular data receives far less attention and remains challenging. Fully connected model structure leads to very complex optimization hyper-planes with a high risk of falling into local optimums.

Explanation can start with the most naive DNN and then systematically add components to improve it, from architectural changes to learning changes.

5.2 Related Work

- Recommender Systems: AirBnB (Haldar *et al.*, 2018)
- Healthcare/EHR: (Rajkomar *et al.*, 2018) showed how effective NNs are for EHR data. State of the art on various predictive tasks.
- Taxi Trajectory: One of the first successful implementations of modern NNs for tabular data was in predicting the destination of a taxi ride based on its initial trajectory (de Brébisson *et al.*, 2015). It was hosted as a Kaggle competition and this solution outperformed all other entries by a significant margin. **Click-through rate prediction (CTR)** (Song

et al., 2018). To predict the probability of a user clicking an item, critical to online applications and recommender systems.

5.3 Input Representation

It is widely held that 80% of the effort in an analytic model is preprocessing, merging, customizing, and cleaning datasets, not analysing them for insights (Rajkomar *et al.*, 2018).

The success of predictive algorithms largely depends on feature selection and data representation. The feature selection process and finding the best data representation is largely a manual and painful process.

In most machine learning tasks the greatest performance gains can be achieved by feature engineering whereas better algorithms only result in incremental boosts. In feature engineering one strives to create new features from the original features based on some domain knowledge of the data or otherwise, that makes it easier for the model to estimate the target. Although a crucial step to make the most out of the data, this can be a very laborious process. There is no formal path to follow in this stage and thus usually consists of many a trial and error, benefitted by domain knowledge of the data, only accessible in some cases. A huge advantage of using NNs on tabular data (and other data structures) is that the feature engineering process gets automated to some extent. A NN learns these optimal feature transformations implicitly during the training process. The hidden layers of a NN can be viewed as a feature extractor that was optimised to map the inputs into the best possible features space for a model (the final layer of the network) to operate in.

5.3.1 Numerical Features

One of the things that make tree-based methods so attractive is that the numeric values of the features hardly matter, as long as their relative ordering is meaningful. On the other hand, NNs are very sensitive to the numeric value of the input. This is related to the optimisation procedure. If an abnormal feature value is fed to the network during training, large gradients can backpropagate through the network and/or result in vanishing gradients (Clevert *et al.*, 2015).

(Haldar *et al.*, 2018) suggest to restrict the values in the range of $\{-1, 1\}$ and so that the median is mapped to zero. They achieved this by inspecting each

of the features and if a feature looks gaussian, do the normalisation $(x - \mu)/\sigma$ and if the feature looks more like a power law distribution, transform is by $\log((1 + x)/(1 + \text{median}))$.

Another step the same authors suggest is to ensure the continuous variables follow a smooth distribution. This helps for generalisation, checking for bugs and general training efficiency. It also helps the analyst to determine whether a feature is generated by some other underlying process.

- how to normalize continuous variables
- mean subtract and error divide
- rankGauss
- scale to 0-1

5.3.2 Categorical Features

Entity embedding not only reduces memory usage and speeds up neural networks compared with one-hot encoding, but more importantly by mapping similar values close to each other in the embedding space it reveals the intrinsic properties of the categorical variables, which you cannot obtain with one-hot encoding.

Companies like Instacart and Pinterest have reported the effective use of entity embeddings on their internal datasets. These embeddings can be reused on different machine learning tasks and do have to be relearned for each dataset.

First published work in modern times on entity embeddings was in the taxi destination prediction challenge (de Brébisson *et al.*, 2015). Another Kaggle success story is for predicting the total sales of a store (Guo and Berkhahn, 2016). This embedding of discrete data was inspired by work done word embeddings in the Natural Language Processing community. There a word is mapped into a vector space of fixed size. The vector representing a word is known as its embedding. The table of embeddings for the words in the dataset is included in the model as a parameterised mapping that can be learned in the same way as the rest of the NN layers. The parameters of the embedding function (or layer) are first randomly initialised and then gets tuned along with the rest of the NN during training.

The embedding for discrete variables act in the exact same way. The embedding for each categorical variable gets concatenated to the continuous variables and then gets passed to the rest of the layers in the network.

In (de Brébisson *et al.*, 2015) they found that embeddings helped a lot. The embeddings can also be visualised to investigate whether make sense or to gain further insight into the data and model decision making. The weights associated with each categories projection onto the embedding space can be plotted with any dimension reduction technique like t-sne or PCA. Then we can compare the categories based on their relative distances and positions in this reduced space.

Entity embeddings are not too different to one-hot encoding a categorical input and sending it through a standard fully connected layer. An embedding is essentially the same operation but a separate one for each of the categorical features. Doing it this way reduces memory usage and speeds up training of a NN. This makes is incredibly useful for datasets with high cardinality features and many of them. It will also not be possible to interpret categories based on its embedding of the one-hot encoded path is followed.

We further demonstrate in this paper that entity embedding helps the neural network to generalize better when the data is sparse and statistics is unknown (Guo and Berkhahn, 2016).

As proof that these entity embeddings actually learns something useful, besides plotting the embedding matrix, one can also feed them along with the continuous features to other learning algorithms and see how it affects performance. (Guo and Berkhahn, 2016) found that the embeddings obtained from the trained neural network boosted the performance of all tested machine learning methods considerably when used as the input features.

5.3.3 Combining Features

5.4 Still to categorise

Feature engineering in general is hard and time consuming with no clear recipe to follow. But it is also very crucial to an effective learning system. The main aim is to find a low-dimensional representation of sparse and high-dimensional raw features and their meaningful combinations. Some of the challenges of feature engineering is listed below.

Can used the Criteo¹ dataset as an example of these challenging datasets.

¹<http://labs.criteo.com/2014/09/kaggle-contest-dataset-now-available-academic-use/>

(Zhou *et al.*, 2017) is very similar to the rest of these citations. (Covington *et al.*, 2016) also embedding + MLP

Sparse and high-dimensional inputs. (Song *et al.*, 2018), (Wang *et al.*, 2017), (Qu *et al.*, 2016), (Cheng *et al.*, 2016), (?), (Covington *et al.*, 2016) Not always the case, but common. High cardinality categorical features are sparse when one-hot encoded. DNNs can easily overfit these sparse and high-dimensional datasets. One-hot encoding + fully connected layer deals with the sparse inputs and the weight matrix is bigger. Separate embeddings are also easier to interpret. Each feature is fed separately into the network so that the layers above can learn specialised representations per feature. This improves generalisation (less parameters) speeds up training and reduces memory footprint (Covington *et al.*, 2016).

Extracting high-order combinations of features. (Song *et al.*, 2018), (Wang *et al.*, 2017), (Qu *et al.*, 2016), (Guo *et al.*, 2017) They key question here is to determine which feautres to combine and how to form meaningful high-order features. Effective prediction usually relies on modelling high-order interactions between features. Majority of the time needs domain experts to help massage the data. Can follow a brute force approach but enumerating all the possible high-order features will exponentially increase the model search space which will just further increase the risk of overfitting. Can use multiple fully connected layers with non-linear activations of a NN. Fully-connected layers model all feature interactions implicitly, but is not good enough to learn all types of interactions. These layers are inefficient in learning multiplicative feature interactions. Hard to explain which features and combinations were important. (Song *et al.*, 2018) uses a multi-head self-attention mechanism which they call the interacting layer. (Zhou *et al.*, 2017) also uses some form of attention but without the softmax layer to reserve intensity of activations. The idea comes from (Vaswani *et al.*, 2017) which itself stems from work done in (Bahdanau *et al.*, 2014). Within in the interacting layer each feature is allowed to interact with every other feature and automatically determine which of those interactions are relevant to the output. They also combine a residual connection between layers so that different orders of feature interactions can be combined. To explain the attention mechanism, consider feature j and the step to determine which high-order features involving feature j are meaningful. We

first define the correlation between features j and k under attention head h as:

$$\alpha_{j,k}^{(h)} = \frac{\exp(\phi^{(h)}(\mathbf{e}_j, \mathbf{e}_k))}{\sum_{l=1}^L \exp(\phi^{(h)}(\mathbf{e}_j, \mathbf{e}_l))}$$

where $\phi^{(h)}(.,.)$ is an attention function which defines the similarity between two features. It can be defined by a neural network or a simple inner product like in (Song *et al.*, 2018):

$$\phi^{(h)}(\mathbf{e}_j, \mathbf{e}_k) = \langle W_{\text{query}}^{(h)} \mathbf{e}_j, W_{\text{key}}^{(h)} \mathbf{e}_k \rangle$$

where $W_{\text{query}}^{(h)}$ and $W_{\text{key}}^{(h)}$ are transformation matrices which map the original embedding space into a new space. The representation of feature j in subspace h is then updated by combining all relevant features guided by coefficients $\alpha_{j,k}^{(h)}$:

$$\tilde{\mathbf{e}}_j^{(h)} = \sum_{k=1}^K \alpha_{j,k}^{(h)} W_{\text{value}}^{(h)} \mathbf{e}_k$$

$\tilde{\mathbf{e}}_j^{(h)}$ is a combination of feature j and its relevant features under attention head h . Therefore it is a learned combinatorial feature. Since a feature can be involved in various different combinations, we use multiple heads to extract combinations, *i.e.* $\{\tilde{\mathbf{e}}_j^{(h)}\}_{h=1}^H$. (Song *et al.*, 2018) used $H=2$. All of these combinatorial features are concatenated into a single vector, $\tilde{\mathbf{e}}_j$. Then finally the output is combined with its raw input (residual connection) and sent through a ReLU:

$$\mathbf{e}_j^{\text{res}} = \text{ReLU}(\tilde{\mathbf{e}}_j + W_{\text{res}} \mathbf{e}_j)$$

This mapping from \mathbf{e}_j to $\mathbf{e}_j^{\text{res}}$ is done for each features to form the interacting layer. The interacting layer is thus a representation of high-order features. These interacting layers can be stacked on-top of each other to form arbitrary order combinatorial features. (Song *et al.*, 2018) shows that residual connection gives better results. (Wang *et al.*, 2017) uses the cross-network which is an automated way of building cross-features. Each layer produces higher-order interactions based on existing ones, and keeps the interactions from previous layers. The cross-network is trained jointly with a DNN. (Wang *et al.*, 2017) also used a residual connection. (Wang *et al.*, 2017) makes a case for finding a bounded-degree feature interactions, saying that all the Kaggle competitions are won with feature engineering of low-degree interactions, whereas DNNs learn highly non-linear interactions implicitly. (Wang *et al.*, 2017) cross-network

consists of cross-layers that can be formalised as:

$$\mathbf{x}_{l+1} = \mathbf{x}_0 \mathbf{x}_l^\top \mathbf{w}_l + \mathbf{b}_l + \mathbf{x}_l$$

where \mathbf{x}_l is the output of the l -th cross layer; \mathbf{x}_0 is the input vector; \mathbf{w}_l and \mathbf{b}_l are its associate weight and bias parameters respectively. Each cross layer adds back its input after feature crossing in a residual connection fashion. (Wang *et al.*, 2017) experimented with 1-6 cross layers. The degree of cross features grows with cross-network depth. The DNN trained in parallel is just a simple network with fully-connected layers and ReLUs. The output of the two streams are concatenated, send through a fully connected layer and a sigmoid layer. (Qu *et al.*, 2016) used something called a product layer, which takes pairwise inner or outer products of all feature combinations and concatenates it to all linear combinations. The output is then fed to 2 fully-connected layers. According to (Guo *et al.*, 2017) it is necessary to capture both low and high-order interactions (and wide&deep paper). They also have to parallel streams of networks, one the FM capturing the low order interactions and one the DNN capturing the high-order interactions. (Cheng *et al.*, 2016) believes it is both important to learn to memorise and generalise. Where memorise refers to recalling from known observations and generalise to predict accurately on unseen samples. They attempt to achieve this again with two streams, one linear layer (wide) and one deep network (deep). The wide stream learns to memorise and the deep network learns to generalise. Combined by a weighted sum. (?) Fully connected model structure leads to very complex optimization hyper-planes with a high risk of falling into local optimums. Therefore it is necessary to explicitly leverage expressive feature combinations. Furthermore it help to limit the model size to make learning more efficient. To achieve this they use automatic feature grouping, feature group reduction and recursive endocder with share embeddings. These ideas seems a little ad-hoc and not end-to-end.

Dealing with mixed input types. (Song *et al.*, 2018), (Wang *et al.*, 2017), (Qu *et al.*, 2016), (Cheng *et al.*, 2016) Processing numerical and categorical features and their combinations. (Song *et al.*, 2018) embeds both the numerical and categorical features separately into a lower-dimensional representation. By mapping both types in the same feature space facilitates more effective learning of interactions between the mixed features. The embedding for the j -th categorical feature is obtained by (Song *et al.*, 2018), (Wang *et al.*,

2017):

$$\mathbf{e}_j = V_j \mathbf{x}_j$$

where \mathbf{x}_j is the one-hot encoded vector representation of the j -th categorical variable and V_j is the associated embedding/weight matrix. The weights in V_j are learned along with all of the other parameters in the network. The embedding for the j -th numerical features is obtained by:

$$\mathbf{e}_j = \mathbf{v}_j x_j$$

where x_j is a scalar and \mathbf{v}_j is the associated weight vector. Why does it help to increase the dimension of the numerical features? Is it only for balanced representation when combined with categorical features? The numerical and categorical embeddings are concatenated to form a single vector representation (Song *et al.*, 2018). (Wang *et al.*, 2017) does not embed numerical but just stack the normalised numerical features along with the categorical embeddings. (Song *et al.*, 2018) used $\log^2(z)$ if $Z > 2$ to transform numeric features to minimise its variance. (Wang *et al.*, 2017) used normal log transform. (Wang *et al.*, 2017) uses normalisation but dont know what yet. (Song *et al.*, 2018) experimented to find the optimal embedding size [8,16,24,32], found it depended on the dataset. (Wang *et al.*, 2017) used embedding sizes of $6 \times (\text{cardinality})^{\frac{1}{4}}$. Not a lot of work on numerical features (?). (Zhou *et al.*, 2017) has an intersting take on multi-hot categorical features; where a feature can have more than category associated with it. The embedding layer for that instance then outputs a list of embeddings with lenght the same as the number of categories associated with that instance and feature. The list of embeddings then gets projected back into a fixed-length representation by doing a pooling operation. (Covington *et al.*, 2016) addresses numeric feature normalisation. (Ioffe and Szegedy, 2015) shows how sensitive DNNs are to scaling and distribution. Proper normalisation of numeric features was critical for convergence (Covington *et al.*, 2016). (Covington *et al.*, 2016) does a transformation to equally distribute a numeric feature in [0,1) using the cumalitive distribution. The integral is approximated with linear interpolation on the quantiles of the feature values computed in a single pass over the data before training begins. In addition they add \hat{x}^2 and $\sqrt{\hat{x}}$ to give the network more expressive power.

Interpreting DNNs. (Song *et al.*, 2018) Model explainability is important for various reasons. Helps to know how to improve your model or where it goes wrong. Like a sanity check. If you cannot explain how a prediction is

made, you cannot know how good it is. (Song *et al.*, 2018) uses the multi-head self-attention mechanism to evaluate correlations between features - globally and locally. (Zhou *et al.*, 2017) plots attention and categorical embeddings.

Layer Normalisation. (Klambauer *et al.*, 2017) Fully-connected DNNs with normalisation techniques are very sensitive to perturbations. DNNs exhibit a high variance in training error when trained using BatchNorm. This hinders the learning process. Combined with dropout just magnifies the effect. (Klambauer *et al.*, 2017) suggested the use of SeLUs which is an activation function which helps the network to maintain zero mean and unit variance activations. By using this activation, there is no need for a BatchNorm layer. SNNs do not suffer from exploding or vanishing gradients. They paper tested on 123 tabular datasets to show that on average SNNs are the best. But it is quite finicky to get the implementation right. It requires a very specific weight initialisation, one that does not really make sense for embedding matrices. And even when you get it right, the improvement is not necessarily significant. That said, it has not been tested by others on tabular data, so it is worth a try.

$$\text{selu}(x) = \lambda \begin{cases} x & \text{if } x > 0 \\ \alpha e^x - \alpha & \text{if } x \leq 0 \end{cases}$$

Give more detail if it is proved to be useful. Also needs a specific type of dropout.

Small Datasets. It is well know that DNNs require a large amount of data to generalise well. Typically, tabular datasets are not as large as unstructured datasets like images and texts. There is also no large tabular dataset from which knowledge can be transferred from like ImageNet for computer vision and wikipedia for NLP. We suggest two techniques for overcoming this problem: data augmentation and unsupervised pre-training. (Zhang *et al.*, 2016) also did pretraining with DAEs. DAEs enforce robustness to partially destroyed inputs. Can also be view from a manifold learning perspective (Vincent *et al.*, 2008). Should also consider VAE and GANS (?) uses output from GBDT to train an initial model and then to use it a initialisation of the actual model. They call it the transfer of structured knowledge. Data augmentation for tabular datasets is rarely studied. Can use corruption like DAEs or swap noise but then creates inputs that does not exist in the real data distribution.

Choosing DNN structural hyperparameters. Since there are no shared patterns among the diverse tabular datasets, it is hard to design a

universal architecture that will fit all. Most of these parameters are very dependent on the dataset and other modeling choices and therefore the need to tune them. Structural hyperparameters are usually found using some brute search. (Song *et al.*, 2018) used a sigmoid layer for binary classification. Embedding dimension: 16 (Song *et al.*, 2018), 32 (Cheng *et al.*, 2016) Hidden layer size: 32 (Song *et al.*, 2018) Number of hidden layers: 32 Dropout: 0.5 (Song *et al.*, 2018), tuned (Zhang *et al.*, 2016) (Qu *et al.*, 2016) (Guo *et al.*, 2017) found dependent on dataset and model. Residual connections: (Song *et al.*, 2018), (Wang *et al.*, 2017) (Song *et al.*, 2018) experimented to see how many layers they should choose [1,2,3,4]. (Song *et al.*, 2018) experimented to find the optimal embedding size [8,16,24,32] (Wang *et al.*, 2017) tested number of layers [2-5] (Wang *et al.*, 2017) tested hidden layer sizes at [32-1024] (Zhang *et al.*, 2016), (Qu *et al.*, 2016), (Guo *et al.*, 2017), (Covington *et al.*, 2016) experimented for number and size of layers. Batchnorm: (Wang *et al.*, 2017) Activation functions: tanh (Zhang *et al.*, 2016), tanh vs sigmoid vs relu (tanh and relu depending on dataset) (Qu *et al.*, 2016), relu vs tanh (relu better) (Guo *et al.*, 2017) Shapes: Diamond, constant, increasing, decreasing (Zhang *et al.*, 2016) found that Diamond shape works best. found constant to work the best (Guo *et al.*, 2017)

Choosing DNN learning hyperparameters. Loss function: logloss with L_2 penalty (Song *et al.*, 2018), (Wang *et al.*, 2017), (Zhang *et al.*, 2016), (Qu *et al.*, 2016) mini-batch aware l2 for large inputs (Zhou *et al.*, 2017), Batch size: 1024 (Song *et al.*, 2018) 512 (Wang *et al.*, 2017) Optimiser: Adam (Song *et al.*, 2018), (Wang *et al.*, 2017) Gradient clipping: (Wang *et al.*, 2017) Learning rate: 0.001-0.0001 (Wang *et al.*, 2017), [1, 0.1, ..., 0.0001] (Zhang *et al.*, 2016) Early stopping: (Song *et al.*, 2018), (Wang *et al.*, 2017), (Zhang *et al.*, 2016) (Zhang *et al.*, 2016) compared dropout with L2 and found dropout to be better.

5.5 Learning Interactions

Automated Feature Engineering

5.5.1 Fully-Connected Layers

- how deep and how big

- architecture shape

5.5.2 Going Deeper

- Residual Connections
- Self-normalising Units

5.5.3 Attention

5.5.4 Parallel Streams

5.6 Dealing with Small Datasets

5.6.1 Data Augmentation

- Blankout
- Swap Noise
- Mixup

Tabular data is very different to image data and the standard augmentations used in computer vision does not make sense with tabular data. You cannot rotate or scale an observation from a tabular data without losing its meaning. One transformation that does make sense for tabular input is the injection of random noise.

When working with images, we can randomly perturb the pixel intensities by a small amount so that it is still possible to make sense of its content. By adding 1 for example to all pixels and all colors in an image, will only make it slightly brighter and we will still be able to make sense of it. Bu with tabular data we can just randomly add a small amount to any feature. The input features will probably not all be on the same scale and the addition of noise might result in a feature value that is out of the true distribution. In addition, it does not make sense to add anything to a discrete variables. Thus in order to inject random noise to a tabular data sample, the noise should be scaled relative to each input feature range and the results should be a valid value for that feature. This also helps the model to be more robust to small variations in the data.

(Van Der Maaten *et al.*, 2013) suggests an augmentation approach that does this called Marginalised Corrupted Features (MCF). The MCF approach adds noise to input from some known distribution.

In the original Denoising Autoencoding paper (Vincent *et al.*, 2008), they used a blank-out corruption procedure. Which is randomly selecting a subset of the input features and mask their values with a zero. The only conceptual problem with this approach is that for some features a zero value actually carries some meaning, so a suggestion is to blank-out features with a unique value not already belonging to that feature distribution.

Another input corruption approach shown to work empirically here is what is called Swap Noise (Kosar and Scott, 2018). The swap noise procedure corrupts inputs by randomly swapping input values with those of other samples in the datasets. In this way you ensure that the corrupted input at least have valid feature values. But it still might produce combinations of features that are not actually possible.

All of these methods have hyperparameters that needs to be set. I haven't gone into detail as I still need to decide what is relevant to this thesis.

- Mixup: (Zhang *et al.*, 2017) - Taking linear combinations of pairs of samples.

5.6.2 Unsupervised Pretraining

- DAEs (Miotto *et al.*, 2016) presented a novel unsupervised deep feature learning method to derive a general-purpose patient representation for EHR data that facilitates clinical predictive modelling. A stacked denoising autoencoder was used. Unsupervised feature learning attempts to overcome limitations of supervised feature space definition by automatically identifying patterns and dependencies in the data to learn a compact and general representation that make it easier to automatically extract useful information when building classifiers or other predictors (Miotto *et al.*, 2016). These techniques are very familiar and effective in text, audio and image processing, but not with tabular data.
- Knowledge distillation/pseudo labelling

5.7 Other

- 1cycle not used with tabular data before
- How to interpret the model/decisions?
- other processing
- Regularisation Learning - <https://arxiv.org/pdf/1805.06440.pdf>

5.8 Recommended Approach

Chapter 6

Experiments

"For us, the most important part of rigor is better empiricism, not more mathematical theories."

— Ali Rahimi and Ben Recht, *NIPS 2017*

6.1 Introudction

- empirical comparisons/ablation studies of approaches discussed in core chapter

6.2 Datasets

- which datasets will experiments performed on

Our experiments are done on multiple datasets. Thus we can distinguish between findings that are only true for certain datasets and tasks and findings that hold more universallt.

The criteria for selecting the datasets were: - Strong model performance baselines exist; so that we can determine how far we are from the SoTA and that is actually a relevant problem. - Entirely open source; so that anyone can access it, reproduce it and build on it. - More than 20,000 observations; since NNs are data hungry. - Does not require too much preprocessing; so that most of the energy goes into the modelling phase. - Contain a mix of continuous and categorical features.

We chose two datasets for regression, one for binary classification and one for multi(class/label) classification from the UCI machine learning repository [Dua2017]. The chosen datasets are:

The Adult dataset¹. This dataset was collected during a census. The task here is to predict whether or not a certain person's income exceeds \$50,000 per year. The features available are things like *age*, *education*, *sex* and *race*. In total there are 14 features and 48,842 observations.

Forest Cover Type²: Predicting forest cover type from cartographic variables. This is a multiclass classification task. There are 581,012 observations.

Taxi Fare Prediction: Regression task (possibly) <https://www.kaggle.com/c/new-york-city-taxi-fare-prediction>

- Maybe <https://www.kaggle.com/c/costa-rican-household-poverty-prediction/data>
- Maybe <https://www.kaggle.com/c/home-credit-default-risk>

Look at the datasets used by (?), also (Zhang *et al.*, 2017)

- Criteo (Song *et al.*, 2018), (Wang *et al.*, 2017), (Qu *et al.*, 2016), (Guo *et al.*, 2017)
- Avazu (Song *et al.*, 2018)
- KDD12 (Song *et al.*, 2018)
- MovieLens-1M (Song *et al.*, 2018)
- iPinYou (Qu *et al.*, 2016)
- Forest Cover type (Wang *et al.*, 2017)
- Higgs (Wang *et al.*, 2017)
- Tox21 (Klambauer *et al.*, 2017)
- Yahoo (?)
- Letor (?)
- Protein (?)
- A9A (?)
- Flight (?)

6.3 Evaluation

For most of the experiments we will do a 5-fold cross validation (Hastie *et al.*, 2009, p. 241) to estimate the performance of a model. That is, randomly dividing the dataset in five equal parts and then in turn, hold out one of those

¹<http://archive.ics.uci.edu/ml/datasets/Adult>

²<https://archive.ics.uci.edu/ml/datasets/coverture>

parts for validation purposes and train the model using the remaining four parts. The performance of the model can then be evaluated on the held-out part. This process is repeated for every one of the five segmentations of the dataset and thus five measurements of the performance of model is obtained. We can then compute the average over these five measurements to obtain a less biased estimate of the model performance. Another advantage of this approach is that we can obtain standard error for the model performance.

Cross-validation is rarely done in Deep Learning, since the models typically take very long to train and any repetition is thus more costly. However, Deep Learning is also mostly applied to large datasets and if a large test set is available, the gains from cross-validation diminishes. Fortunately, the NNs applied to tabular data are much smaller than ones used for unstructured data and for this work we have access to sufficient computing power. And therefore cross-validation makes sense.

For the regression tasks we will compare the various models using the mean squared error and for classification we use cross-entropy. These are the metrics directly being optimised during the training process. When comparing the results to previous work, we base it on the metrics that are common for the specific dataset.

6.3.1 Metrics

- loss function
- task specific
- dataset specific
- time and memory

AUC, Logloss for binary classification (Song *et al.*, 2018), (Wang *et al.*, 2017), (Zhang *et al.*, 2016), (Qu *et al.*, 2016) No cross-validation (Song *et al.*, 2018), (Zhang *et al.*, 2016) Run time per epoch (Song *et al.*, 2018), (Guo *et al.*, 2017) Model size (Song *et al.*, 2018) Think (Wang *et al.*, 2017) used CV since the report se's for hyperparameter tunings.

6.3.2 Cross-validation

6.4 General Approach

- which numeric normalisation
- should we embed numerics

Preprocessing. - Remove infrequent features (Song *et al.*, 2018) - numeric transform by $\log^2(z)$ if $Z > 2$ (Song *et al.*, 2018). - numeric features log transform (Wang *et al.*, 2017)

6.5 Architectural Search

- Number of layers (vs with SeLU) (vs other activations) [1-10]
- Layer size [32-2048]
- Architecture shape [Constant, increasing, decreasing, diamond]
- Embedding sizes [proportional, fixed]
- dropout [0-1]

Recently found that the below experiments were already done by (Guo *et al.*, 2017, Qu *et al.* (2016), Zhang *et al.* (2016)). This was however only explored for Click-through rate prediction data. Thus the below experiments should be done in the light of these findings and can be compared to their findings.

Here we investigate the effect of the size of the network on the different datasets. We compare the performance of the models at different numbers and sizes of layers. Larger networks are more flexible and therefore we expect it to act similarly to any learning model flexibility parameter. Increasing the network size will be beneficial up until a certain point until it becomes too big and be more prone to overfitting. We hope to find a rule of thumb that might act as a good starting point and guideline to choose the network size. We also want to get a feel for how important these hyperparameters are.

- Constant size Layer sizes: 32, 64, 128, 256, 512, 1024, 2048 Number of layers: 1,2,3,4,5,6 At a constant dropout.

Suppose we choose three layers, compare the following shape at approximately equal number of parameters. Shapes: Constant, decreasing, increasing, diamond, hourglass

The aim of these experiments are to evaluate performance at different embedding sizes. We explore embedding sizes at different ratios of the cardinality of categorical variables. The ratios we look at are: 10%, 20% 30%, 40%, 50%, 60%, 70%, 80% and 90% of the cardinality of each categorical feature. Possibly explore certain max embedding sizes.

As we increase the size we will also look at the effect it has by visually inspecting the embedding layers in a 2-dimensional space.

Again, we expect there to be optimal embedding size for each variable depending on the cardinality of the variable and how complex its relationship is with the other variables and the target. We expect the ideal embedding size to be as small as possible but still being able to capture all of the information of the variable.

Look at wide and deep models from (Cheng *et al.*, 2016). They restricts all embeddings to be of size 32.

6.6 Sample Size

- accuracy vs size of dataset

6.7 Mixup

- does it help the validation loss

6.8 Pretraining

- does it help the validation loss
- Are these features useful for tree based methods.

6.9 Attention

- with residual

6.10 Comparisons To Tree-based Methods

- Compare Neural Networks to Gradient Boosted Machines and Random Forests.

6.11 Example Interpretation

- plot embeddings
- plot attention matrices
- SHAP and permutation

Chapter 7

Conclusion

- What was done in the thesis?
- Is Deep Learning useful for tabular data?
- If it is, when?
- Where should future work on the subject focus on?

Appendices

Appendix A

Appendix A

Description of each of the datasets used in Experiments.

Bibliography

- Bahdanau, D., Cho, K. and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *CoRR*, vol. abs/1409.0473. 1409.0473.
Available at: <http://arxiv.org/abs/1409.0473>
- Bengio, Y., Courville, A. and Vincent, P. (2013 Aug). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828. ISSN 0162-8828.
- Che, Z., Purushotham, S., Khemani, R.G. and Liu, Y. (2016). Interpretable deep models for icu outcome prediction. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, vol. 2016, pp. 371–380.
- Cheng, H., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., Anderson, G., Corrado, G., Chai, W., Ispir, M., Anil, R., Haque, Z., Hong, L., Jain, V., Liu, X. and Shah, H. (2016). Wide & deep learning for recommender systems. *CoRR*, vol. abs/1606.07792. 1606.07792.
Available at: <http://arxiv.org/abs/1606.07792>
- Clevert, D.-A., Unterthiner, T. and Hochreiter, S. (2015 November). Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *ArXiv e-prints*. 1511.07289.
- Covington, P., Adams, J. and Sargin, E. (2016). Deep neural networks for youtube recommendations. In: *Proceedings of the 10th ACM Conference on Recommender Systems*. New York, NY, USA.
- de Brébisson, A., Simon, É., Auvolat, A., Vincent, P. and Bengio, Y. (2015). Artificial neural networks applied to taxi destination prediction. *CoRR*, vol. abs/1508.00021. 1508.00021.
Available at: <http://arxiv.org/abs/1508.00021>
- Friedman, J.H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American statistical Association*, vol. 76, no. 376, pp. 817–823.

- Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep Learning*. MIT Press.
<http://www.deeplearningbook.org>.
- Guo, C. and Berkahn, F. (2016). Entity embeddings of categorical variables. *CoRR*, vol. abs/1604.06737. 1604.06737.
Available at: <http://arxiv.org/abs/1604.06737>
- Guo, H., Tang, R., Ye, Y., Li, Z. and He, X. (2017). Deepfm: A factorization-machine based neural network for CTR prediction. *CoRR*, vol. abs/1703.04247. 1703.04247.
Available at: <http://arxiv.org/abs/1703.04247>
- Haldar, M., Abdool, M., Ramanathan, P., Xu, T., Yang, S., Duan, H., Zhang, Q., Barrow-Williams, N., Turnbull, B.C., Collins, B.M. and Legrand, T. (2018 October). Applying Deep Learning To Airbnb Search. *ArXiv e-prints*. 1810.09591.
- Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. 2nd edn. Springer.
Available at: <http://www-stat.stanford.edu/~tibs/ElemStatLearn/>
- He, K., Zhang, X., Ren, S. and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, vol. abs/1512.03385. 1512.03385.
Available at: <http://arxiv.org/abs/1512.03385>
- Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, vol. abs/1502.03167.
Available at: <http://arxiv.org/abs/1502.03167>
- Klambauer, G., Unterthiner, T., Mayr, A. and Hochreiter, S. (2017). Self-normalizing neural networks. *CoRR*, vol. abs/1706.02515. 1706.02515.
Available at: <http://arxiv.org/abs/1706.02515>
- Kosar, R. and Scott, D.W. (2018 January). The Hybrid Bootstrap: A Drop-in Replacement for Dropout. *ArXiv e-prints*. 1801.07316.
- Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems, NIPS'12*, pp. 1097–1105. Curran Associates Inc., USA.
Available at: <http://dl.acm.org/citation.cfm?id=2999134.2999257>
- Lecun, Y., Bengio, Y. and Hinton, G. (2015 5). Deep learning. *Nature*, vol. 521, no. 7553, pp. 436–444. ISSN 0028-0836.

- Li, M., Zhang, T., Chen, Y. and Smola, A.J. (2014). Efficient mini-batch training for stochastic optimization. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pp. 661–670. ACM, New York, NY, USA. ISBN 978-1-4503-2956-9.
Available at: <http://doi.acm.org/10.1145/2623330.2623612>
- Miotto, R., Li, L., Kidd, B.A. and Dudley, J.T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. In: *Scientific reports*.
- Perez, L. and Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *CoRR*, vol. abs/1712.04621. 1712.04621.
Available at: <http://arxiv.org/abs/1712.04621>
- Qu, Y., Cai, H., Ren, K., Zhang, W., Yu, Y., Wen, Y. and Wang, J. (2016). Product-based neural networks for user response prediction. *CoRR*, vol. abs/1611.00144. 1611.00144.
Available at: <http://arxiv.org/abs/1611.00144>
- Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Liu, P.J., Liu, X., Sun, M., Sundberg, P., Yee, H., Zhang, K., Duggan, G.E., Flores, G., Hardt, M., Irvine, J., Le, Q.V., Litsch, K., Marcus, J., Mossin, A., Tansuwan, J., Wang, D., Wexler, J., Wilson, J., Ludwig, D., Volchenboum, S.L., Chou, K., Pearson, M., Madabushi, S., Shah, N.H., Butte, A.J., Howell, M., Cui, C., Corrado, G. and Dean, J. (2018). Scalable and accurate deep learning for electronic health records. *CoRR*, vol. abs/1801.07860. 1801.07860.
Available at: <http://arxiv.org/abs/1801.07860>
- Ramachandran, P., Zoph, B. and Le, Q.V. (2017). Searching for activation functions. *CoRR*, vol. abs/1710.05941. 1710.05941.
Available at: <http://arxiv.org/abs/1710.05941>
- Rosenblatt, F. (1962). *Principles of neurodynamics: perceptrons and the theory of brain mechanisms*. Report (Cornell Aeronautical Laboratory). Spartan Books.
Available at: <https://books.google.ca/books?id=7FhRAAAAMAAJ>
- Shavitt, I. and Segal, E. (2018 May). Regularization Learning Networks: Deep Learning for Tabular Datasets. *ArXiv e-prints*. 1805.06440.
- Shickel, B., Tighe, P., Bihorac, A. and Rashidi, P. (2017). Deep EHR: A survey of recent advances on deep learning techniques for electronic health record (EHR)

- analysis. *CoRR*, vol. abs/1706.03446. 1706.03446.
Available at: <http://arxiv.org/abs/1706.03446>
- Smith, L.N. (2018). A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay. *CoRR*, vol. abs/1803.09820. 1803.09820.
Available at: <http://arxiv.org/abs/1803.09820>
- Song, W., Shi, C., Xiao, Z., Duan, Z., Xu, Y., Zhang, M. and Tang, J. (2018). AutoInt: Automatic feature interaction learning via self-attentive neural networks. *CoRR*, vol. abs/1810.11921. 1810.11921.
Available at: <http://arxiv.org/abs/1810.11921>
- Van Der Maaten, L., Chen, M., Tyree, S. and Weinberger, K.Q. (2013). Learning with marginalized corrupted features. In: *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pp. I-410–I-418. JMLR.org.
Available at: <http://dl.acm.org/citation.cfm?id=3042817.3042865>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. (2017). Attention is all you need. *CoRR*, vol. abs/1706.03762. 1706.03762.
Available at: <http://arxiv.org/abs/1706.03762>
- Vincent, P., Larochelle, H., Bengio, Y. and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In: *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pp. 1096–1103. ACM, New York, NY, USA. ISBN 978-1-60558-205-4.
Available at: <http://doi.acm.org/10.1145/1390156.1390294>
- Wang, R., Fu, B., Fu, G. and Wang, M. (2017). Deep & cross network for ad click predictions. *CoRR*, vol. abs/1708.05123. 1708.05123.
Available at: <http://arxiv.org/abs/1708.05123>
- Zhang, H., Cissé, M., Dauphin, Y.N. and Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *CoRR*, vol. abs/1710.09412. 1710.09412.
Available at: <http://arxiv.org/abs/1710.09412>
- Zhang, W., Du, T. and Wang, J. (2016). Deep learning over multi-field categorical data: A case study on user response prediction. *CoRR*, vol. abs/1601.02376. 1601.02376.
Available at: <http://arxiv.org/abs/1601.02376>

Zhou, G., Song, C., Zhu, X., Fan, Y., Zhu, H., Ma, X., Yan, Y., Jin, J., Li, H. and Gai, K. (2017 June). Deep Interest Network for Click-Through Rate Prediction. *ArXiv e-prints*. 1706.06978.