### **158.736 Assignment 1**

# **Abstract**

Write this up last

# **Summary**

### **Write this up second to last**

# **Background / Method**

## **Datasets**

Describe the nature of each dataset

Six dataset were used to adopt the 2 shallow learning and 2 deep learning method for machine learning.

#### **Forest Data**

This dataset is a cross sectional dataset comprising 581,012 observations across 53 different features.

The class label for each record is a type of tree type that is observed at a particular 30m2 section of the Roosevelt National Park in Colorado.

The features in the dataset relate to various geographical features which might affect the type the forest cover observed such as: elevation, slope, proximity to water/roads, sunlight and soil types.

Of particular note is that there are 40 soil type indicator variables that identify the soil type within the plot whereas the remaining variables are continuous variables.

#### **Street View House Numbers (SVHN)**

This dataset is 3D (colour) 32x32 image set of house numbers observed on letterboxes. The dataset is distinguishable in that the images generally do not have a single number in the image but instead have ‘noise’ relating to other numbers or part-numbers in the image.

#### **Pump Sensor Data**

The data were the raw value came from 52 unit of sensors. The sensors were related to water pump of a small area far from big town. There are 7 system failure in last year. There is a timestamp value showing the recorded time of the sensor data. The class label for each record is the machine status of the water pump with 3 possible values: Normal, Recovering and Broken.

#### **Point Cloud MNIST 2D dataset**

It take the original MNIST dataset and converts each of the non-zero pixels into points in a 2D space. The idea is to classify each collection of point (rather than images) to the same label as in the MNIST.

There are 2 files: train.csv and test.csv. Each file has the columns label,x0,y0,v0,x1,y1,v1,...,x350,y350,v350

where

    label contains the target label in the range [0, 9]

    x{i} contain the x position of the pixel/point as viewed in a Cartesian plane in the range [-1, 27].

    y{i} contain the y position of the pixel/point as viewed in a Cartesian plane in the range [-1, 27].

    v{i} contain the value of the pixel in the range [-1, 255].

The maximum number of point found on a image was 351, images with less points where padded to this length using the following values:

    x{i} = -1

    y{i} = -1

    v{i} = -1

#### **Traffic Signs dataset(German traffic sign classification)**

Detecting Street Signs is one of the most important tasks in Self Driving Cars.This dataset is a benchmark in Street signs and symbols including 43 different classes. Classifying road symbols using Deep Convolutional Neural Network is the aim of this dataset.

#### **Pulsar Star dataset**

Each candidate is described by 8 continuous variables, and a single class variable. The first four are simple statistics obtained from the integrated pulse profile (folded profile). This is an array of continuous variables that describe a longitude-resolved version of the signal that has been averaged in both time and frequency . The remaining four variables are similarly obtained from the DM-SNR curve .

## **Learning Methods**

Outline what the various learning methods are, how they work, strengths and weaknesses etc.

#### **Logistic Regression**

This standard and widely utilised classification technique utilises a linear combination of variables through the logistic/sigmoid function to generate class prediction probabilities which lie between 0 and 1.

#### **Random Forest**

A random forest is an ensemble of numerous underlying simplified random decision trees.

A standard decision tree classifier will consider all features and at each stage select the feature and cut point which creates the biggest information gain. Decision Trees however suffer from high variability in that classification results are highly dependent on the initial feature and cut-off selections.

The Random Forest produces less variable results by utilising an ensemble of decision trees which only consider a random selection of features at each decision point. This approach reduces the variability of results while still allowing the key differentiating features to dominate via the ‘majority vote’ approach.

In this project the standard approach of allowing the square root of the total number of features to be considered at each decision point.

#### **Neural Network - Multi-Layer Perceptron (MLP)**

The Multi-Layer Perceptron model is characterised by multiple modelling layers comprising a variable number of neurons at each layer. Each neuron takes all inputs from the preceding layer and gives the ability to model different aspects of the data. Layering allows modelled features to be refined and re-used by other layers. The downside of MLPs is that all layers are fully connected which can result in an explosion of parameters that need to be trained and the resultant computational complexity.

#### **Neural Network - Convolutional Neural Network (CNN)**

The CNN model is characterised by use of filters and pooling techniques which help to reduce the dimensionality of problems which have large number of features. The use of filters helps to identify common and recurring features within the overall feature space. The use of pooling then allows those identified features to be shrunk into a smaller representative feature space for modelling.

CNN’s are particular strong in classifying image datasets where the feature space is large (each individual pixel) where the filters can identify recurring patterns in the images and shrink the results to reduce computational effort.

## **Expectations**

Given the nature of the datasets and what we know about the learning methods, what would we expect to occur, ie which methods would we expect to perform best on each data set

#### **Street View House Numbers (SVHN)**

Given that the SVHM dataset comprises colour images of 32x32 this equates to 3,072 features/pixels.

It would be expected that shallow learning methods would struggle to accurately classify these images due to the high level of dimensionality.

Of the deep learning methods it would be expected that a CNN would be the most effective modelling technique as the filtering and pooling capabilities identify common features in the images and compress the results.

#### **Forest Data**

As the Forest Data only comprised 53 dimensions it would be expected that shallow methods such as logistic regression and random forests would be able to perform a reasonable job of classifying the forest type.

It would be expected that a deep learning method could outperform on this dataset and given the pure cross-sectional nature of this dataset it would be expected that a MLP could well perform best

#### **Pump Sensor Data**

Similar that the Pump Sensor Data comprised 52 dimensions, it would be expected that shallow methods such as logistic regression and random forests would be able to perform a reasonable job of classifying the machine status. But it was noticed that the sample of “Broken” machine status only have 7 records among 220,320 records. The trained model may not able to predict the “Broken” labels.

It would be expected that a deep learning method could outperform on this dataset and given the pure cross-sectional nature of this dataset it would be expected that a MLP could well perform best

#### **Point Cloud MNIST 2D dataset**

After restructure the dataset, it was represented as image of 28x28 and this equates to 784 features/pixels.

It would be expected that shallow learning methods would have some difficulty to accurately classify these images due to the high level of dimensionality.

Within the deep learning methods it would be expected that a CNN would be the most effective modelling technique as the filtering and pooling capabilities identify common features in the images and compress the results.

Pulsar Star Data

The Pulsar Star Data has a simple structure with 9 dimensions and extensive data records. It would be expected that both of shallow learning methods and deep learning methods would be to achieve an acceptable machine learning model. It can be easily predicted that deep learning will take more time than shallow learning.

Traffic Signs Data

The Traffic Signs Data comprises color images of 32x32 similar with SVHN. It would be expected that shallow learning methods would not get an accurately classify model for these images due to the high level of dimensionality. CNN is the most commonly used to analyze visual imagery and it would train a suitable model for the Traffic Signs Data.

## **General Modelling Approach**

Outline approach to data preparation, cleaning etc. How was training vs validation performed. What level of parameter tuning was performed etc.

**Training vs Validation**

For all models the hold-out validation approach was applied by training the model on a random selection of 80% of the data while testing was performed on the remaining 20% of data.

**Scaling**

Min/Max scaling was applied to the underlying Forest/SVHN/…./.../….. Datasets.

Standard Normal scaling was applied to Pump Sensor,Pulsar Star and Point Cloud MNIST 2D Dataset

**Number of Epochs for Deep Learning Training**

Setting the number of epochs for deep learning affects overall results as training results steadily improve as more epochs are run however with diminishing improvements.

Training a model with too many epochs thought can result in test results declining due to overfitting.

The number of epochs that models were trained for was set to the number where for that particular model the test results plateaued.

**Regularisation**

Regularisation is the process of removing insignificant variables from a model to avoid overfitting.

For the deep learning methods the dropout approach was used which random drops out parameters after have layer which forces the model to to find a more general way of fitting rather that relying on certain parameters.

For the logistic regressions, the L2 regularisation parameter (which penalises the use of a variable/feature) was determined via a grid search to optimise the test results.

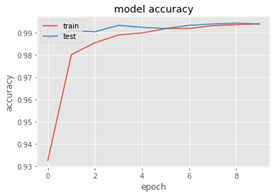
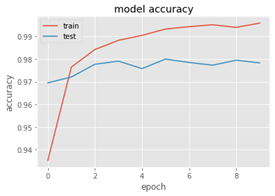
# **Results**

For each dataset suggesting putting in the table of results and then discussing the observed results.

**Comparison between MLP and CNN**

From the result comparison between MLP and CNN, it’s found that the CNN is better especially in controlling the overfitting than the MLP. Compare to MLP, CNN has one or more layers of convolution units and max pooling units. These has beneficial such as:

* Reduce the number of units in the network. That means there are fewer parameters to learn which reduces the chances of overfitting as the model have less complexity than a fully connected network.
* It considered the context/shared information in the small neighborhoods. This is very important in many applications such as image, video, text, and speech processing/mining as the neighboring inputs (eg pixels, frames, words, etc) usually carry related information.



(MLP)         (CNN)

# **Discussion**

Comment on the findings. Which results were as expected? Which results were not as expected. If results are not as expected, rationalise why that has occured.

What further work could be explored? How could modelling approaches be improved.

# **Summary**

# **Appendix 1: Model Configurations**

Maybe put the model diagrams here?

# **Appendix 2: Model Fit Diagnostics**

Put training vs test charts and things like confusion matricies here?