Machine Learning Engineer Nanodegree

Capstone Project

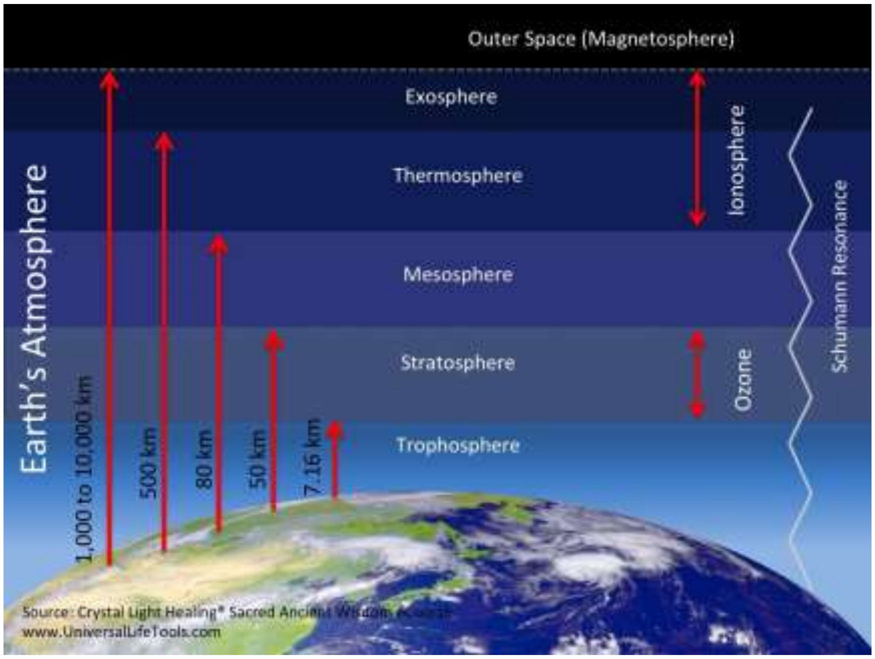
Title: Classification of Radar returns system from Ionosphere layer

Haroon Gharwi
Febrery 4, 2019

I. Definition

Project Overview

The atmosphere is comprised of layers of gases based on temperature. These layers are the troposphere, stratosphere, mesosphere, thermosphere then exosphere. the ionosphere is a series of regions in parts of the mesosphere and thermosphere where it is an Extreme Ultra Violet (EUV) and x-ray solar radiation ionizes the atoms and molecules thus creating a layer of electrons. this ionosphere layer is important because it reflects and modifies radio waves that used for radio communication and satellite navigation.

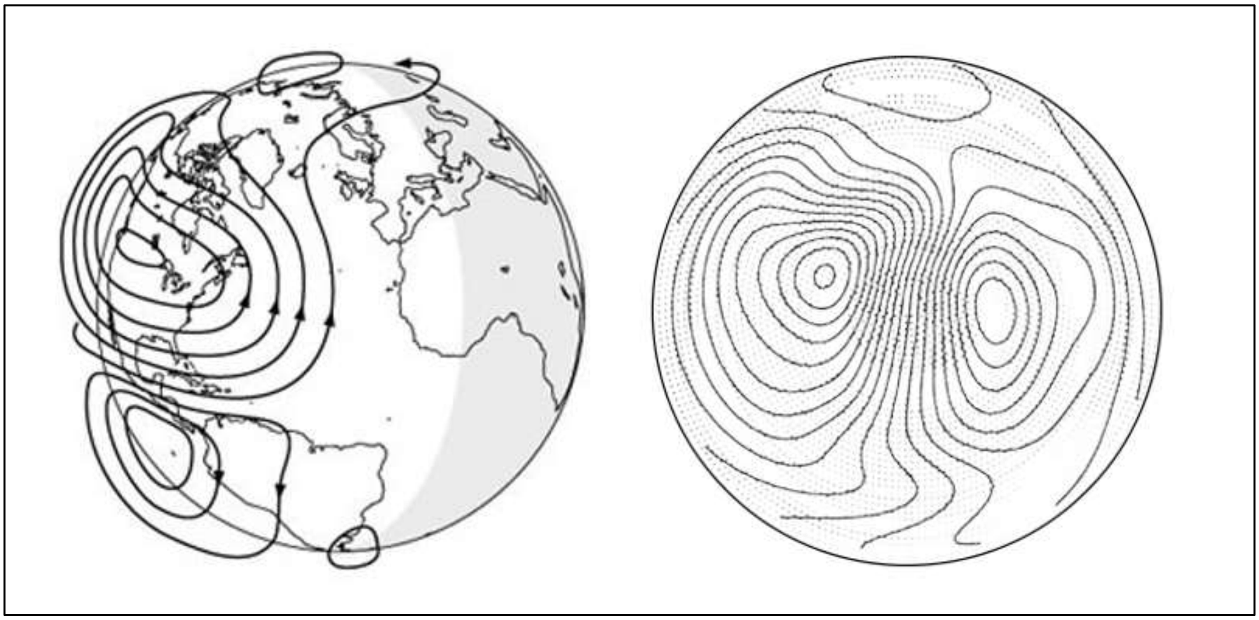


Source: Crystal Light Healing® Sacred Ancient Wisdom Academy
www.UniversalLifeTools.com

Problem Statement

The radar data used in our study were collected by the Space Physics Group of The Johns Hopkins University Applied Physics Laboratory at 1989. The radar system, located in Goose Bay, Labrador, consists of a phased array of 16 high-frequency antennas, with a total transmitted power on the order of 6.4 kW and an antenna gain of about 30 dBm at frequency ranges of 8 to 20 MHz. The signals returns are used to study the physics of the ionosphere. Good signal returns can show evidence of some structure type in the ionosphere while bad radar returns cannot where their signals pass through the ionosphere. The received signals were processed using an autocorrelation function whose arguments are the time of a pulse and the pulse number.

According the Goose Bay radar system, there were 17 pulse numbers for the Goose Bay system. Each pulse in this dataset are described by 2 attributes, corresponding to the complex values returned by the function resulting from the complex electromagnetic signal. The target is showing the free electrons in the ionosphere. “Good” radar returns are those showing evidence of some type of structure in the ionosphere. “Bad” (b) returns are those that do not; their signals pass through the ionosphere. The radar dataset having 351 samples with 35 features. All 34 features are represented as continuous numeric values and the last one is the target value and has two possible result either be Good or Bad.



The radar signal operates by transmitting a multi-pulses pattern to the ionosphere. The receiver is turned on between pulses, and the target velocity is determined by measuring the phase shift of the all returns. If we denote the received signal from the pulse at time t by C(t) = A (t) + iB(t) , then the autocorrelation function (ACF), R, is given by 16 R(t,k) = E C(t + iT)C*[t + (i + k)T] where T is the pulse repetition period, k indicates the pulse number, and the * indicates complex conjugation.

the machine learning has a potential in many wide areas to solve the problems with highly efficient and down the human efforts. One of these machine learning algorithms is a multi layers perceptron neural network method (MLP). where that used to solve the classification problems. In this capstone project will we use the MLP to solve this radar classification problem that normally would require human interference. The MLP network will identify the "good" and "bad" radar returns from the ionosphere.

Metrics

Sense here we involved with a classification problem,the confusion matrix will be used to evaluate the performance of the MLP model:

	Predicted bad	**Predicted good**
Actual bad	True negative	False negative
Actual good	False positive	True positive

According the confusion matrix, there are many methods to measure the performance of the model such as accuracy and F1 score. An Accuracy method will used here to measure the performance of MLP model. The Accuracy is the number of correct predictions made by the MLP model over the total of predictions made:

$$Accuracy = \frac{True\ ngative + True\ positive}{True\ negtive + True\ positive + Flase\ negtive + False\ Positive}$$

The accuracy method is best fit to measure the MLP model performance because we are here carrying about the total number of correct predictions by the model. Unlike F1 score method, where F1 score method is used when it is necessary to prefer some side of confusion matrix to other.

II. Analysis

Data Exploration

The Ionosphere dataset is contained 35 columns. First 34 columns (V0-V34) represent the 17 pulse values of the radar system. These columns are cleaned and there is no any missed values. Also, all the data are continuous values ranging between -1 and 1 this mean the data not need to make pre-processing step. However, the second feature contains only zeros, so it better removes this feature. The last column (class) on the dataset is represent the return result of the radar system either be 0 or 1. Where 0 means the ionosphere layer is bad while 1 means the ionosphere layer is good for a radio communication.

	Column name	discerption
Features	34 columns (V0 to V34)	Continuous values between [-1,1]
Target	Las column (Class)	Has two values [0,1] 0= bad class ,1= good class

Exploratory Visualization

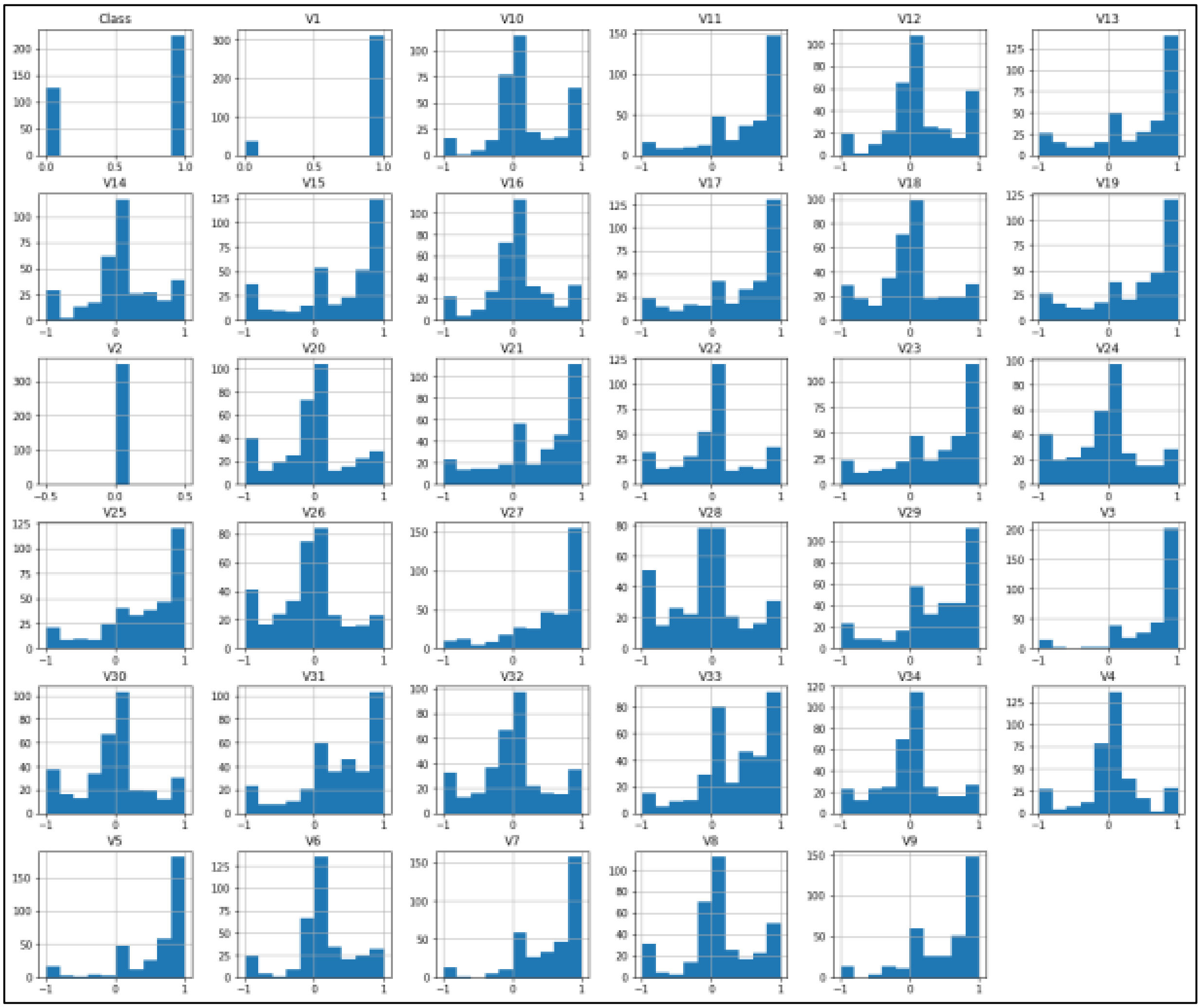
In order to understand the entire of ionosphere dataset, the table below describes the statistical of the dataset briefly:

	V1	V2	V3	V4	V5	.	V33	V34	class
count	351	351	351	351	351	.	351	351	351
mean	0.891738	0.0	0.641342	0.044372	0.601068	.	0.349364	0.01448	0.641026
min	0.000000	0.0	-1.00000	-1.00000	-1.00000	.	-1.00000	-1.00000	0.000000
max	1.000000	0.0	1.000000	1.000000	1.000000	.	1.000000	1.00000	1.000000

Throw looking to the count row, the data has no any missing values. In addition, the dataset is clear and ranging between -1 and 1 according to min and max rows. Unless V2 column, it got zeros. So, it has to remove it. As we can see, we can conclude the data is prepared and there is no preprossing step indeed. The table below shows sample of the dataset:

	V1	V2	V3	V4	V5	.	V33	V34	class
0	1	0	0.99539	-0.05889	0.85243	.	0.18641	-0.45300	1
1	1	0	1.00000	-0.18829	0.93035	.	-0.13738	-0.02447	0
2	1	0	1.00000	-0.03365	1.00000	.	0.56045	-0.38238	1
3	1	0	1.00000	-0.45161	1.00000	.	-0.32382	1.00000	0

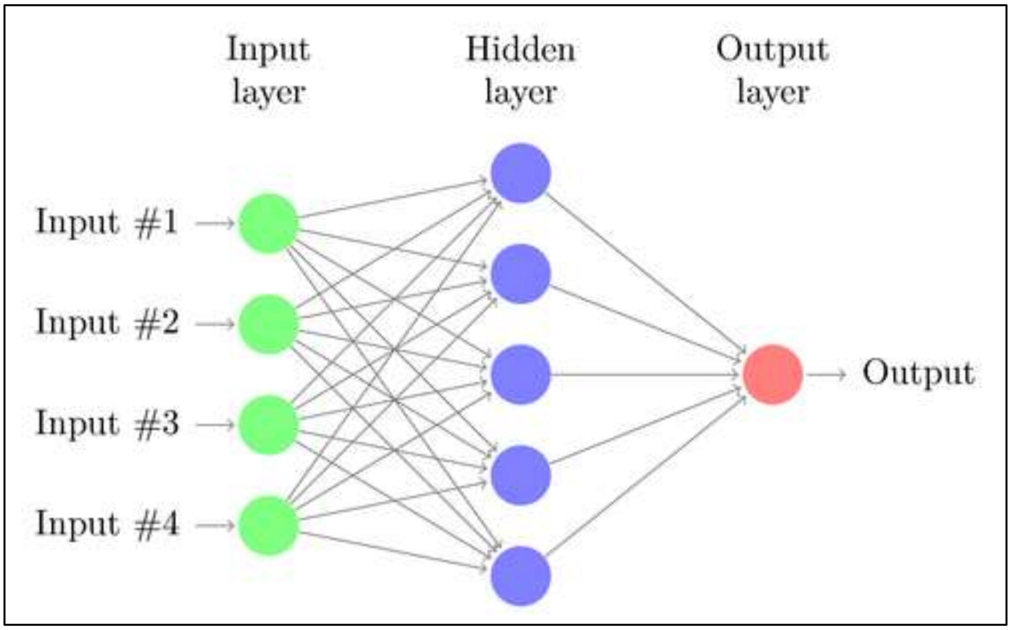
The histogram below shows how the data are distrusted for each column. As we can see, the data ranging between -1 and 1. Moreover, V2 column has only zeros:



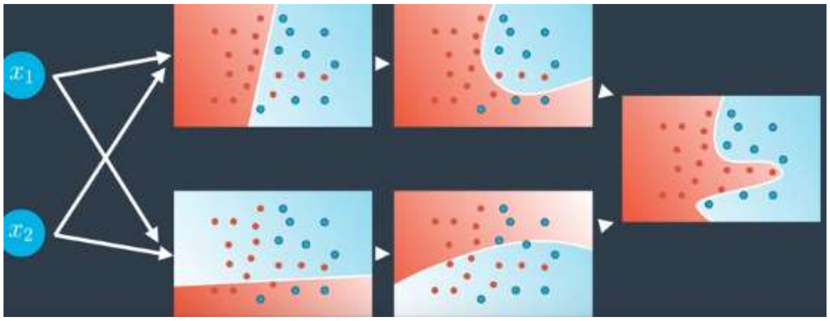
Algorithms and Techniques

The radar system problem is a classification problem based. The multilayer perceptron (MLP) neural network will be used to implement this classification problem.

A multilayer perceptron (MLP) is a class of feedforward artificial neural network. An MLP consists of, at least, three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. The most common activation functions are sigmoid and Rectified linear unit (ReLU) MLP utilizes a supervised learning technique called backpropagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron.



The MLP is used to produce a nonlinear classification boundary that utilized for solving a complex classification problem.



There are many libraries available to approach MLP model such as Keras, Sklearn and TensorFlow. The keras library will be used in this project to build MPL model.

Benchmark

The table below shows the results of other researches that applied the machine learning algorithms to solve this problem. This table shows the highest ranks for methods that used in the problem. So, this table will be as the reference to compare it with my algorithm accuracy result at the end.

Method	Accuracy %	Reference
3-NN + simplex	98.7	Our own weighted kNN
VSS 2 epochs	96.7	MLP with numerical gradient
3-NN	96.7	KG, GM with or without weights
IB3	96.7	Aha, 5 errors on test
1-NN, Manhattan	96.0	GM kNN (our)
MLP+BP	96.0	Sigillito
SVM Gaussian	94.9±2.6	GM (our), defaults, similar for C=1-100
C4.5	94.9	Hamilton
3-NN Canberra	94.7	GM kNN (our)
RIAC	94.6	Hamilton
C4 (no windowing)	94.0	Aha
C4.5	93.7	Bennet and Blue
SVM	93.2	Bennet and Blue
Non-lin perceptron	92.0	Sigillito
FSM + rotation	92.8	our
1-NN, Euclidean	92.1	Aha, GM kNN (our)

DB-CART	91.3	Shang, Breiman
Linear perceptron	90.7	Sigillito
OC1 DT	89.5	Bennet and Blue
CART	88.9	Shang, Breiman
SVM linear	87.1±3.9	GM (our), defaults
GTO DT	86.0	Bennet and Blue

III. Methodology

Data Preprocessing

As we mentioned in before, the data is prepared and there is no any preprossing step indeed. the table below shows that the data is ranging between -1 and 1

	V1	**V2**	**V3**	**V4**	**V5**	**.**	**V33**	**V34**	**class**
count	351	351	351	351	351	.	351	351	351
mean	0.891738	0.0	0.641342	0.044372	0.601068	.	0.349364	0.01448	0.641026
min	0.000000	0.0	-1.00000	-1.00000	-1.00000	.	-1.00000	-1.00000	0.000000
max	1.000000	0.0	1.000000	1.000000	1.000000	.	1.000000	1.00000	1.000000

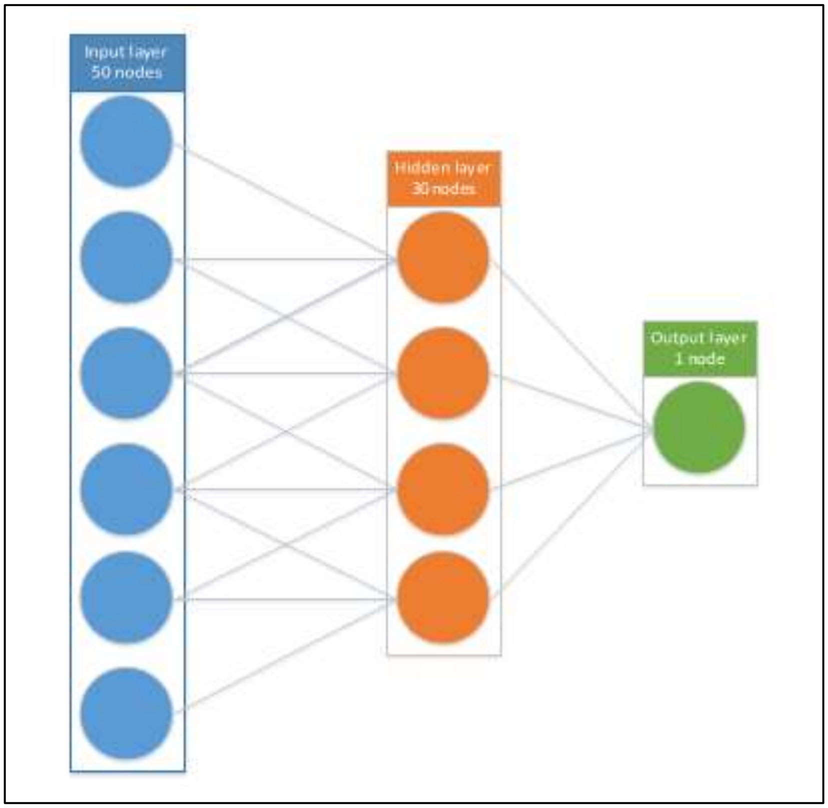
Implementation:

There are many steps are followed to implement the MLP model:

First, define the variable: the dataset is spliced into two variables, one variable that contains pulse columns (V1, V3, V4, V5.... V34) named as the features. While the second variable is a radar_class as the target of classification. Then the dataset is converted to vectors sense the MLP model requires the input data as vector essentially.

second, split the data. the data is divided into train and test set (with percentage 20%) in order to train the model well. The data divided throw Sklearn library. This led to 280 samples out of 351 samples for training and 71 samples for testing

Third, model architecture. the model of MLP is built and designed by using the Keras library with the following architecture:



Fourth,tune the model. The MLP model is tuned, where here the parameters of MLP model is defined:

- Matrix: the accuracy matrix will be used to check the accuracy of the model
- Optimizer: the Adam optimazer will be used to control the training model
- Cross validation: 0.2. split the train set into train and cross validation set with 20% of the data to check how the model is doing during the training.

- Loss function: the binary crossentropy will be used to calculate the losses during training the model.
- Epochs: 10. It is the total number of repeating time that the all training set will be trained.
- Batch size: 30. mean 30% of the training set will be trained in each cycle during training the model.

Fifth, train the model. Where here the model will be started to train the training set and produce the final MLP model

Finally, score the model. after the final MLP model is produced, the test set will be applied to the MLP model and check the accuracy of our model.

Refinement

The initial model produced with accuracy 88.73%. There are many criteria can be tuned to improve the accuracy of the model. these criteria consist: number of nodes of each MLP layer, percentage of nodes to drop out in each cycle, optimizer type, percentage of cross validation set, number of epochs and the batch size. To tune the MLP model and find the best parameters values we will use the grid search methods. The grid search method allows us to train the model with different parameters values and pick up the best setting for the model. Moreover, the grid search method uses K-fold cross validation method. In the k-fold cross validation method, all the entries of the original training dataset are used for both training as well as validation. Where The training dataset will be divided into k number groups. Each group will be used for validation only once while the remain groups will be used for the training. Then do the same process for each group. The final score will be the sum and average all these results. This k fold method takes advantage to tune the model and keep all the training data set for the training at the same time sense the size of our dataset is a little small.

The grid search will be assigned with k =10 and with these possible parameters values:

number of nodes of each MLP layer (50,30,1), (70,30,1) and (80,30,1)

percentage drop out (0.1, 0,1) and (0.2,0.2)

optimizer type: Adm and rmsprop

number of epochs:10 and100

batch size: 20% and 30 %

IV. Results

Model Evaluation and Validation

After a time taking for tuning the MLP model with grid search method, the model has got a high predicting performance with accuracy 95.77% followed by these parameters values:

number of nodes of each MLP layer (80,30,1)

percentage drop out (0.2, 0,2)

optimizer type: rmsprop

number of epochs:100

batch size: 30%

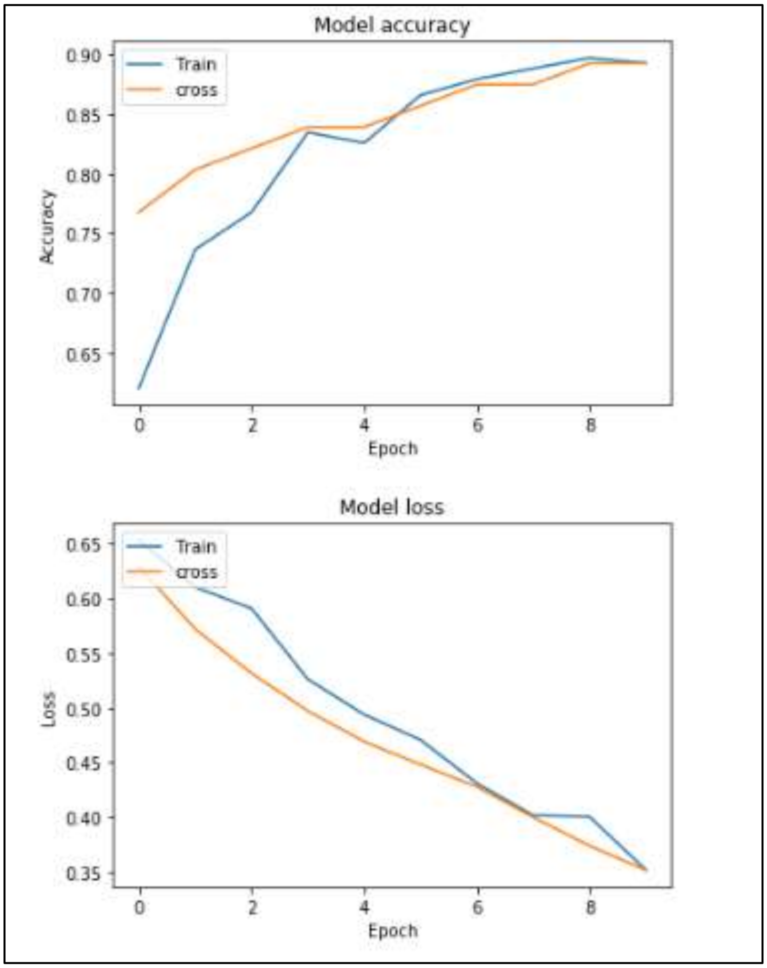
Justification

According the benchmark model, the highest accuracy record is 98.7%,96.7% then 96.0% On the other hand, the MLP model here is got predicting with 95.77 % accuracy. The model accuracy is acceptable according the performance of benchmark model.

V. Conclusion

Free-Form Visualization

The graphs below observe the performance of the initial parameters MLP model that training over the time. The history for the training set is labelled as train while the cross validation set is labelled as cross. From the plot of accuracy, we can see that the model accuracy tends to be rising over the time as well as cross validation. From loss plot, we can see the model is start as underfitting, where both set with high loss. Then the MLP model start learning very well over the time. However, both graphs show that the model is not complete learning enough! This give a great notice to value of epoch, where that the epoch= 10 is not enough to train the model. For this we got to put the epoch=100 in the next training rate.



Reflection

Nowadays, the machine learning has a great benefit to apply in a lot of modern applications that involved to solve the problems with a highly efficient and down the human efforts. One of these machine learning algorithms is a multi layers perceptron neural network method (MLP). where that is used to solve the classification problems. In this capstone project, the MLP model is used to solve this radar classification problem that has the ability to identify how the radar return from the ionosphere layer is good or not with high efficient and effortless that normally would require the human interference.

Actually, the difficulty of this project was during the searching and understanding the domain background of the ionosphere layer to identify the problem and its dataset. But the most interesting on this project is to apply the whole skills that acquired during this nanodegree program.

Improvement

As we can see this MLP model is got a high performance for this project. However, it may be can improve this model throw tuning the model parameters or changing the structure of the neural network model. Moreover, there are many other methods can be involved with this project area such as support vector machine (SVM) and decision tree.

References

https://en.wikipedia.org/wiki/Atmosphere

https://scied.ucar.edu/atmosphere-layers

https://en.wikipedia.org/wiki/Ionosphere

http://www.aeronomie.be/en/topics/earthsystem/ionosphere-gps.htm

https://archive.ics.uci.edu/ml/datasets/ionosphere

http://fizyka.umk.pl/kis-old/projects/datasets.html

http://superdarn.thayer.dartmouth.edu/downloads/96JA01584.pdf

https://pdfs.semanticscholar.org/e0d2/de05caacdfa8073b2b4f77c5e72cb2449b81.pdf