Machine Learning Engineer Nanodegree

Capstone Proposal

Title: Ionosphere Data Set

Haroon gharwi
janurey 30, 2019

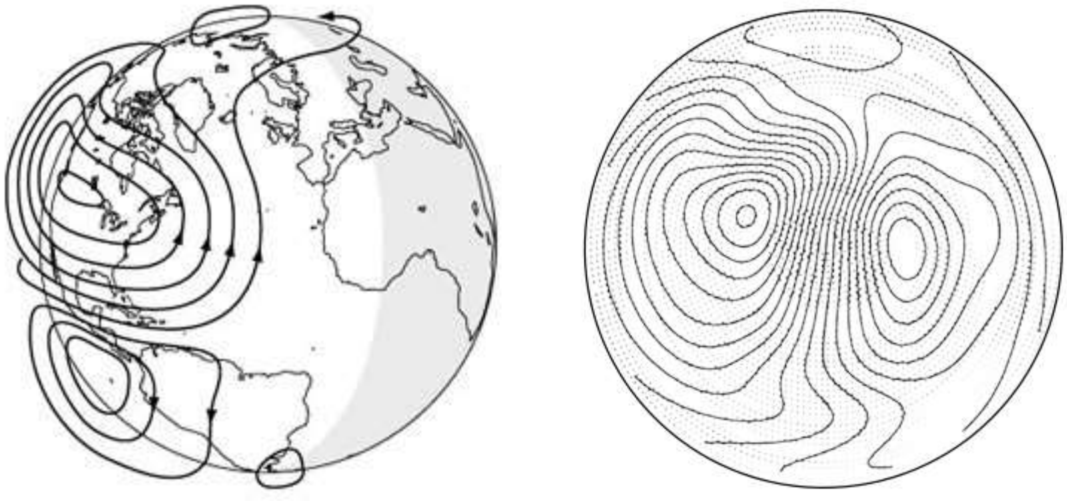
Proposal

Domain Background

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.

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.



Datasets and Inputs

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.

Solution Statement

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.

Benchmark Model

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

Evaluation Metrics

The confusion matrix will be used to evaluate the accuracy the PLM model:

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

Project Design

The dataset is clean 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 remove this feature. Moreover, the value target is a classification type and it has two possible classes (Good/Bad). Sense the target value is known and it is a classification type, I will use the supervisor methods to solve this problem. There are many methods applicable to solve the classification problem, mostly I will go directly to use the multi layers perceptron neural network to solve this problem.

The steps that to solve the problem:

Data preparation and exploration:

-import libraries

-define variables

Training Machine learning algorithm and Evaluation:

-splitting the data into training and testing set

-define the model

-configure the models' settings

-cross validation

-fit and apply the inputs of training set to the models to train the model

-predict the output of training set

-find the accuracy of model

-tuning the model.

-fit and apply the input of test set to the model

-scoring the model

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