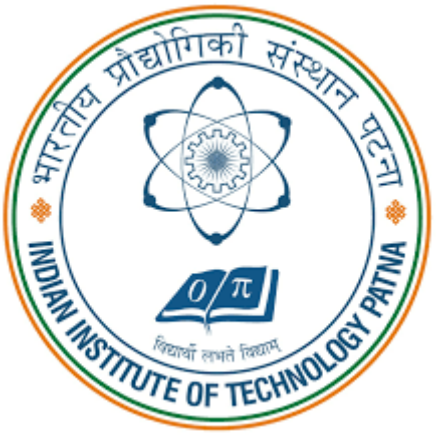


Indian Institute of Technology Patna

CS515

Computer Systems Lab II

Machine Learning on Ionosphere Data-set using Scikit-learn

Dr.Rajiv Mishra
IIT Patna

Ankit Kumar 1711CS03
Manish Kumar Kaushik 1711CS10
Mainak Maulik 1711CS09

Project abstract

The ionosphere is the ionized part of Earth's upper atmosphere, from about 60 km (37 mi) to 1,000 km (620 mi) altitude, a region that includes the thermosphere and parts of the mesosphere and exosphere. The ionosphere is ionized by solar radiation. It plays an important role in atmospheric electricity and forms the inner edge of the magnetosphere. It has practical importance because, among other functions, it influences radio propagation to distant places on the Earth.

We have taken a dataset of Ionosphere and applied some machine learning technique (Gaussian Naive-Bayes, Decision Tree and Support Vector Machine) and predicted the presence/absence (i.e. Good/Bad) of structure in Ionosphere.

Description about data-set

For this study, Ionosphere dataset was taken from UCI Machine learning website: https://archive.ics.uci.edu/ml/datasets/ionosphere This radar data was collected by a system in Goose Bay, Labrador. This system consists of a phased array of 16 high-frequency antennas with a total transmitted power on the order of 6.4 kilowatts. See the paper for more details. The targets were free electrons in the ionosphere. "Good" radar returns are those showing evidence of some type of structure in the ionosphere. "Bad" returns are those that do not; their signals pass through the ionosphere.

Received signals were processed using an autocorrelation function whose arguments are the time of a pulse and the pulse number. There were 17 pulse numbers for the Goose Bay system. Instances in this database are described by 2 attributes per pulse number, corresponding to the complex values returned by the function resulting from the complex electromagnetic signal.

– All 34 are continuous – The 35th attribute is either "good" or "bad" according to the definition summarized above. This is a binary classification task.

Machine learning

Machine learning is a field of computer science that uses statistical techniques to give computer systems the ability to "learn" with data, without being explicitly programmed.

Here in this project we have used three different types of ML algorithms:
1. Gaussian Naive Bayes:

Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values.
It is called naive Bayes or idiot Bayes because the calculation of the probabilities for each hypothesis are simplified to make their calculation tractable. Rather than attempting to calculate the values of each attribute value P(d1, d2, d3—h), they are assumed to be conditionally independent given the target value and calculated as P(d1—h) * P(d2—H) and so on.

Naive Bayes can be extended to real-valued attributes, most commonly by assuming a Gaussian distribution.
This extension of naive Bayes is called Gaussian Naive Bayes. Other functions can be used to estimate the distribution of the data, but the Gaussian (or Normal distribution) is the easiest to work with because you only need to estimate the mean and the standard deviation from your training data.

2. Support Vector Machine:

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

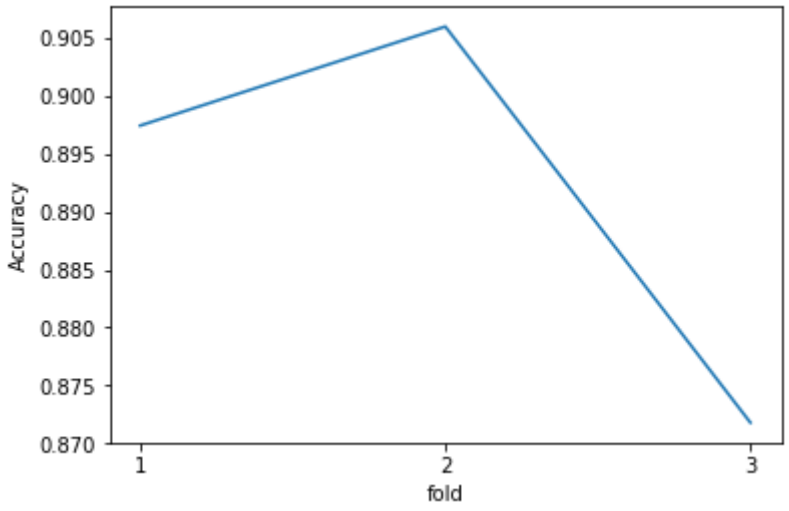
3. Decision tree:

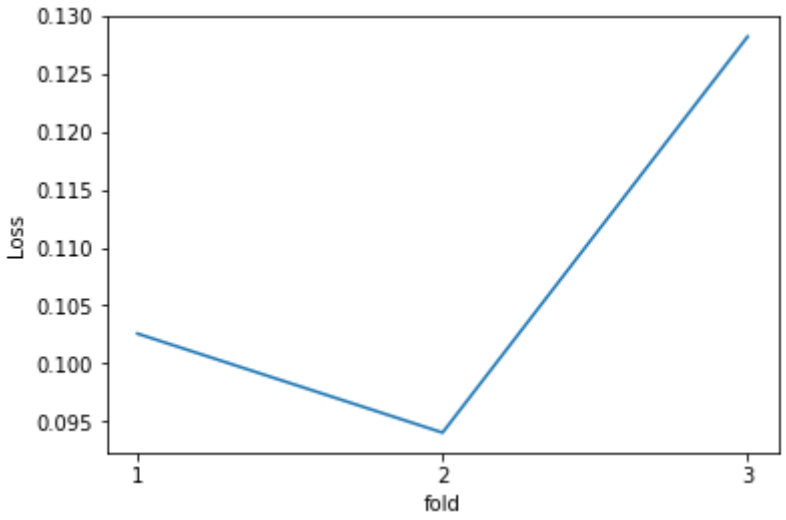
A decision tree is a map of the possible outcomes of a series of related choices. It allows an individual or organization to weigh possible actions against one another based on their costs, probabilities, and benefits. They can can be used either to drive informal discussion or to map out an algorithm that predicts the best choice mathematically.

Experiments and Results

Gaussian Naive-Bayes:

Here we have used 3 folds
confusion matrix for fold 1
$\begin{bmatrix} 31 & 8 \\ 2 & 76 \end{bmatrix}$
confusion matrix for fold 2
$\begin{bmatrix} 29 & 9 \\ 1 & 78 \end{bmatrix}$
confusion matrix for fold 3
$\begin{bmatrix} 38 & 11 \\ 3 & 65 \end{bmatrix}$
Sensitivity or True Positive rate 0.77809235110196
Specificity (SPC) or True Negative rate 0.9653104787584059
Precision Mean 0.9031339031339032
Recall Mean 0.8751871655237199
F1-score Mean 0.8898841918144619
Accuracy Mean 0.9031339031339032
Log Loss Mean 1.053367754101865
Error Mean 0.09686609686609689
Mean Squared Mean 0.09686609686609687

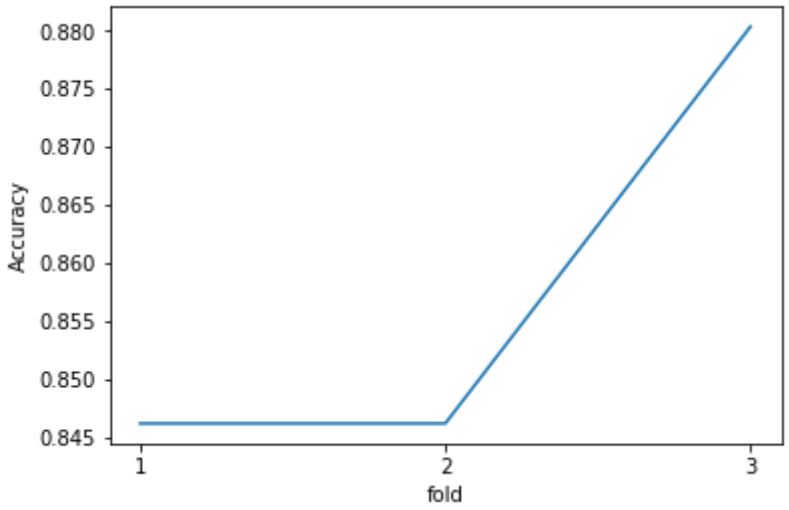


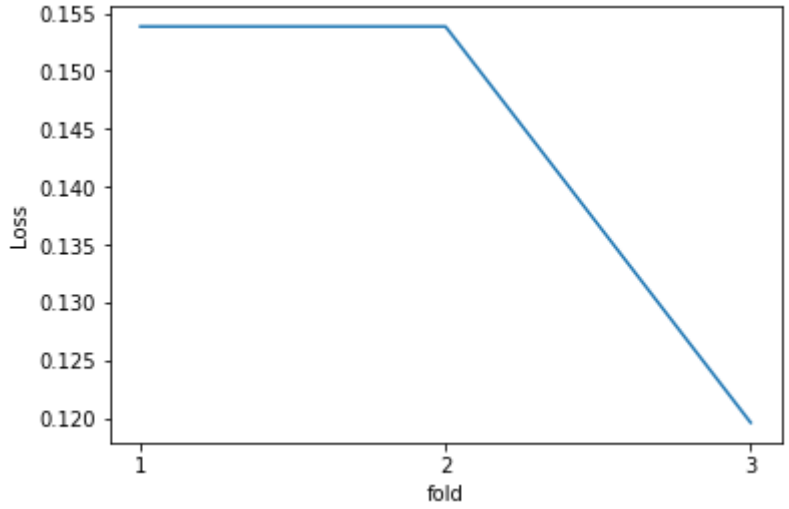


Decision Tree:

Here we have used 3 folds
confusion matrix for fold 1
$\begin{bmatrix} 32 & 8 \\ 5 & 72 \end{bmatrix}$
confusion matrix for fold 2
$\begin{bmatrix} 34 & 6 \\ 1 & 76 \end{bmatrix}$
confusion matrix for fold 3
$\begin{bmatrix} 40 & 6] \\ 10 & 61 \end{bmatrix}$

Sensitivity or True Positive rate 0.8166440831074978
Specificity (SPC) or True Negative rate 0.9159040277461331
Precision Mean 0.8974358974358975
Recall Mean 0.8834663448411152
F1-score Mean 0.8880665161966532
Accuracy Mean 0.8974358974358975
Log Loss Mean 3.542438604606224
Error Mean 0.10256410256410257
Mean Squared Mean 0.10256410256410257

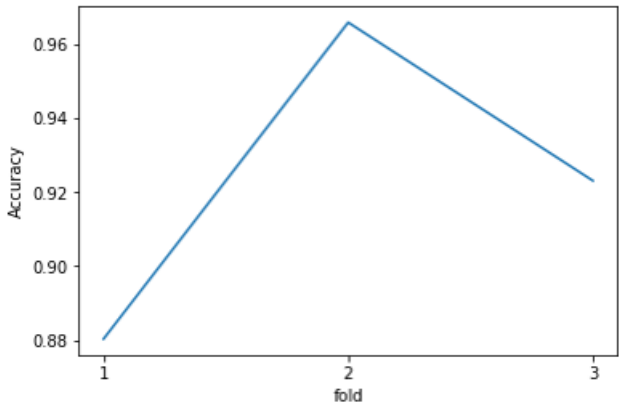


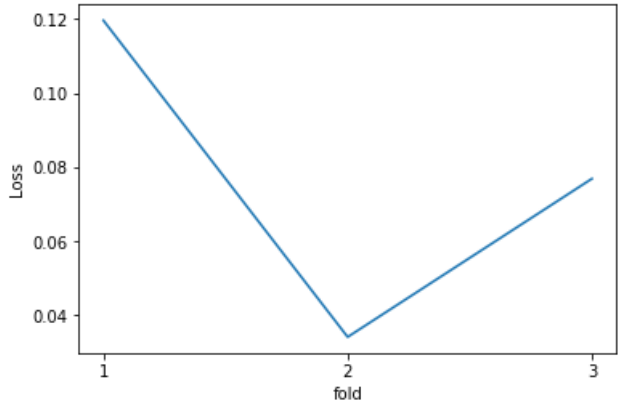


Support Vector Machine:

Here we have used 3 folds
confusion matrix for fold 1

$$\begin{bmatrix} 29 & 11 \\ 0 & 77 \end{bmatrix}$$
confusion matrix for fold 2
$$\begin{bmatrix} 34 & 10 \\ 1 & 72 \end{bmatrix}$$
confusion matrix for fold 3
$$\begin{bmatrix} 35 & 7 \\ 0 & 75 \end{bmatrix}$$
Sensitivity or True Positive rate 0.8131690660760428
Specificity (SPC) or True Negative rate 0.991430454845089
Precision Mean 0.9173789173789174
Recall Mean 0.8862269959872698
F1-score Mean 0.9047016195072377
Accuracy Mean 0.9173789173789174
Error Mean 0.08262108262108263
Mean Squared Mean 0.08262108262108263





Conclusion

In this project, we successfully classified Ionosphere data in two categories Good and Bad with accuracy of 94% with Support Vector Machine.