Predicting Agricultural Output by applying Machine Learning.

Siddhant Ghosh[1] Prof. Sonali Patil[2] Utkarsha Matere[3]
[1,2,3]Suman Ramesh Tulsiani Technical Campus, India

Abstract— Data sets of different related to the agriculture is vastly available. The main motto of our system will be to create a machine learning algorithm and use the available datasets for the unsupervised machine learning process and thereby predict the future crop production, crops which would mostly likely to give a greater profit margin, etc. The prediction can also help various State as well as Central Government for taking appropriate steps as described by the prediction. The methodology will be as follows: By taking various datasets from the government and by the help of clustering for the first step and in the next step linear regression for correctly predicting. The reason why we are using unsupervised learning for the first step is so that we won't lose any data. The whole system will be available as a website with a GUI that can accept data for the learning process. The output will also depend upon the present state of the environment. The features for the learning process are:

1) Weather data (Rainfall, Winds, frequency of droughts, so on.)
2) Soil composition.
3) Types of fertilizers used.
4) The source for seeds/saplings
5) Seasonal data.
6) Whether mechanized farms or manual labor used. (If manual labor is not used what is the extent of mechanization)

Any more help whatsoever received from the government.

Key words: Machine Learning, KNN Filter, Gaussian Process Regression, Perceptron-Based, Single-Layered Perceptron-Based, Multi-Layered Perceptron-Based, Naïve Bayes Classification, Ordinary Least Squares Classification, Logistic Regression

I. INTRODUCTION

Machine learning studies computer algorithms for learning to do stuff. We might, for instance, be interested in learning to complete a task, or to make accurate predictions, or to behave intelligently. The learning that is being done is always based on some sort of observations or data, such as examples (the most common case in this course), direct experience, or instruction. So in general, machine learning is about learning to do better in the future based on what was experienced in the past.

The emphasis of machine learning is on automatic methods. In other words, the goal is to devise learning algorithms that do the learning automatically without human intervention or assistance. The machine learning paradigm can be viewed as "programming by example."

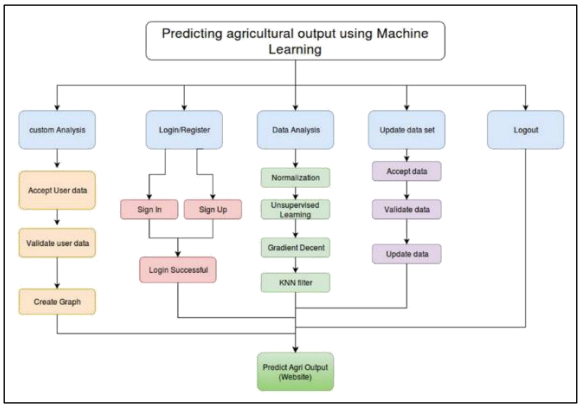
Often we have a specific task in mind, such as spam filtering. But rather than program the computer to solve the task directly, in machine learning , we seek methods by which the computer will come up with its own program based on examples that we provide.

Machine learning is a core subarea of artificial intelligence. It is very unlikely that we will be able to build any kind of intelligent system capable of any of the facilities that we associate with intelligence, such as language or vision, without using learning to get there. These tasks are otherwise simply too difficult to solve. Further, we would not consider a system to be truly intelligent if it were incapable of learning since learning is at the core of intelligence. Although a sub-area of AI, machine learning also intersects broadly with other fields, especially statistics, but also mathematics, physics, theoretical computer science and more.

II. LITERATURE SURVEY

The primary goal of machine learning research is to develop general purpose algorithms of practical value. Such algorithms should be efficient. As usual, as computer scientists, we care about time and space efficiency. But in the context of learning, we also care a great deal about another precious resource, namely, the amount of data that is required by the learning algorithm. Learning algorithms should also be as general purpose as possible. We are looking for algorithms that can be easily applied to a broad class of learning problems, such as those listed above. Of primary importance, we want the result of learning to be a prediction rule that is as accurate as possible in the predictions that it makes. Occasionally, we may also be interested in the interpret ability of the prediction rules produced by learning. In other words, in some contexts (such as medical diagnosis), we want the computer to find prediction rules that are easily understandable by human experts. As mentioned above, machine learning can be thought of as "programming by example." What is the advantage of machine learning over direct programming? First, the results of using machine learning are often more accurate than what can be created through direct programming. The reason is that machine learning algorithms are data driven, and are able to examine large amounts of data.

III. PROPOSED TECHNOLOGY



All rights reserved by www.ijsrd.com

IV. ALGORITHMS

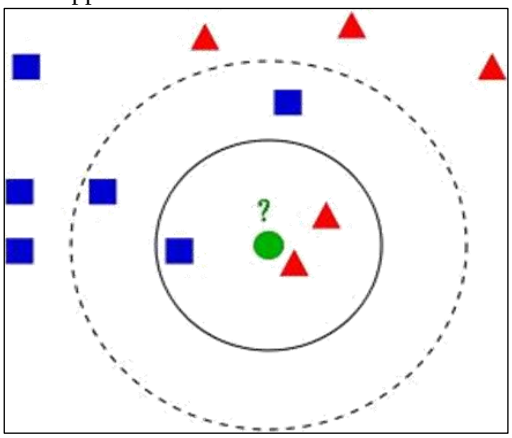
A. KNN - Filter

The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples.

In the classification phase, k is a user-defined constant, and an unlabeled vector (a query or test point) is classified by assigning the label which is most frequent among the k training samples nearest to that query point.

A commonly used distance metric for continuous variables is Euclidean distance. For discrete variables, such as for text classification, another metric can be used, such as the overlap metric (or Hamming distance). In the context of gene expression micro-array data, for example, k-NN has also been employed with correlation coefficients such as Pearson and Spearman. Often, the classification accuracy of KNN can be improved significantly if the distance metric is learned with specialized algorithms such as Large Margin Nearest Neighbor or Neighborhood components analysis.

A drawback of the basic "majority voting" classification occurs when the class distribution is skewed. That is, examples of a more frequent class tend to dominate the prediction of the new example, because they tend to be common among the k nearest neighbors due to their large number. One way to overcome this problem is to weight the classification, taking into account the distance from the test point to each of its k nearest neighbors. The class (or value, in regression problems) of each of the k nearest points is multiplied by a weight proportional to the inverse of the distance from that point to the test point. Another way to overcome skew is by abstraction in data representation. For example, in a self-organizing map (SOM), each node is a representative (a center) of a cluster of similar points, regardless of their density in the original training data. K-NN can then be applied to the SOM.



B. Gaussian process

In probability theory and statistics, a Gaussian process is a particular kind of statistical model where observations occur in a continuous domain, e.g. time or space. In a Gaussian process, every point in some continuous input space is associated with a normally distributed random variable. Moreover, every finite collection of those random variables has a multivariate normal distribution, i.e. every finite linear combination of them is normally distributed. The distribution of a Gaussian process is the joint distribution of all those (infinitely many) random variables, and as such, it is a distribution over functions with a continuous domain, e.g. time or space.

Viewed as a machine-learning algorithm, a Gaussian process uses lazy learning and a measure of the similarity between points (the kernel function) to predict the value for an unseen point from training data. The prediction is not just an estimate for that point, but also has uncertainty information—it is a one-dimensional Gaussian distribution (which is the marginal distribution at that point).

For some kernel functions, matrix algebra can be used to calculate the predictions using the technique of kriging. When a parameterized kernel is used, optimization software is typically used to fit a Gaussian process model.

The concept of Gaussian processes is named after Carl Friedrich Gauss because it is based on the notion of the Gaussian distribution (normal distribution). Gaussian processes can be seen as an infinite-dimensional generalization of multivariate normal distributions.

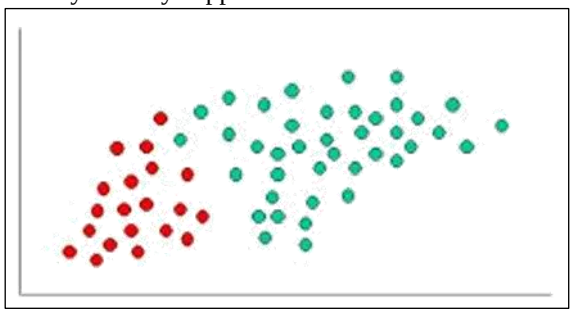
Gaussian processes are useful in statistical modelling, benefiting from properties inherited from the normal. For example, if a random process is modelled as a Gaussian process, the distributions of various derived quantities can be obtained explicitly. Such quantities include the average value of the process over a range of times and the error in estimating the average using sample values at a small set of times.

C. Naïve Bayes Classification

The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods.

To demonstrate the concept of Naïve Bayes Classification, consider the example displayed in the illustration above. As indicated, the objects can be classified as either GREEN or RED. Our task is to classify new cases as they arrive, i.e., decide to which class label they belong, based on the currently exiting objects.

Since there are twice as many GREEN objects as RED, it is reasonable to believe that a new case (which hasn't been observed yet) is twice as likely to have membership GREEN rather than RED. In the Bayesian analysis, this belief is known as the prior probability. Prior probabilities are based on previous experience, in this case the percentage of GREEN and RED objects, and often used to predict outcomes before they actually happen.



Since there are twice as many GREEN objects as RED, it is reasonable to believe that a new case (which hasn't

All rights reserved by *www.ijsrd.com*

been observed yet) is twice as likely to have membership GREEN rather than RED. In the Bayesian analysis, this belief is known as the prior probability. Prior probabilities are based on previous experience, in this case the percentage of GREEN and RED objects, and often used to predict outcomes before they actually happen.

Thus, we can write,

$$\text{Prior probability for GREEN} \propto \frac{\text{Number of GREEN objects}}{\text{Total number of objects}}$$

$$\text{Prior probability for RED} \propto \frac{\text{Number of RED objects}}{\text{Total number of objects}}$$

Although the assumption that the predictor (independent) variables are independent is not always accurate, it does simplify the classification task dramatically, since it allows the class conditional densities p(xk | Cj) to be calculated separately for each variable, i.e., it reduces a multidimensional task to a number of one-dimensional ones. In effect, Naive Bayes reduces a high-dimensional density estimation task to a one-dimensional kernel density estimation.

Furthermore, the assumption does not seem to greatly affect the posterior probabilities, especially in regions near decision boundaries, thus, leaving the classification task unaffected.

Naive Bayes can be modeled in several different ways including normal, lognormal, gamma and

$$p(x_k \mid C_j) = \begin{cases} \frac{1}{\sigma_{kj}\sqrt{2\pi}} \exp\left(\frac{-(x-\mu_{kj})^2}{2\sigma_{kj}}\right), & -\infty < x < \infty, -\infty < \mu_{kj} <, \sigma_{kj} > 0 \quad \text{Normal} \\ \mu_{kj}: \text{mean}, \sigma_{kj}: \text{standard deviation} & \\ \frac{1}{x\sigma_{kj}(2\pi)^{1/2}} \exp\left\{\frac{-[\log(x/m_{kj})]^2}{2\sigma_{kj}^2}\right\}, & 0 < x < \infty, m_{kj} > 0, \sigma_{kj} > 0 \quad \text{Lognormal} \\ m_{kj}: \text{scale parameter}, \sigma_{kj}: \text{shape parameter} & \\ \frac{\left(\frac{x}{b_{kj}}\right)^{c_{kj}-1}}{b_{kj}\Gamma(c_{kj})} \exp\left(\frac{-x}{b_{kj}}\right), & 0 \le x < \infty, b_{kj} > 0, c_{kj} > 0 \quad \text{Gamma} \\ b_{kj}: \text{scale parameter}, c_{kj}: \text{shape parameter} & \\ \frac{\lambda_{kj}\exp(-\lambda_{kj})}{x!}, & 0 \le x < \infty, \lambda_{kj} > 0, x = 0,1,2,\ldots \quad \text{Poisson} \\ \lambda_{kj}: \text{mean} & \end{cases}$$

Poisson density functions:

D. Performance Parameters

1) Goal:

Applying Machine Learning to predict which crop would be the most profitable by making use of various data available in the government offices which are related to agriculture and to help farmers to maximize their profit.

2) Approach:

To make utmost use of the available data and through supervised learning create a system which would be efficient and give an output which is within the admissible range.

E. Design Fundamentals

1) Inputs for learning:

- Weather data.
- Soil data.
- Types of fertilizers used.
- Source of seeds.
- Seasons.
- Funds received from Government
- Mechanized farms or Manual Farms.

2) Framework:

The system will be divided into Three parts:

- Machine learning
- GUI
- Database

The GUI will be web based which can be accessed from anywhere for easily accessibility. It will have a user database which users can access and enter data into the database for the machine learning process.

The Database will be used for the supervised Learning procedure.

The Machine Learning part will have the core logic and the algorithms required for the whole system.

F. Testing

To test our system, first we would give the system sufficient data to 'learn' and create its own database to refer and give the desired output. The methodology will be as follows: Suppose that we have data of 10 years. We will enter the data for supervised learning of 9 years and if the output is in a nearby range of the 10th year, we will consider our work to be a success.

V. Applications

Farmers, Government bodies and Entrepreneurs can use this system to maximize profit and minimize losses. The customers who will use this system will profit from using this system because they can take informed decisions

One of the most important issues in machine learning is whether one can improve the performance of a supervised learning algorithm by including unlabeled data. Methods that use both labeled and unlabeled data are generally referred to as semi-supervised learning. Although a number of such methods are proposed, at the current stage, we still don't have a complete understanding of their effectiveness. This paper investigates a closely related problem, which leads to a novel approach to semi-supervised learning. Specifically we consider learning predictive structures on hypothesis spaces (that is, what kind of classifiers have good predictive power) from multiple learning tasks. We present a general framework in which the structural learning problem can be formulated and analyzed theoretically, and relate it to learning with unlabeled data. Under this framework, algorithms for structural learning will be proposed, and computational issues will be investigated. Experiments will be given to demonstrate the effectiveness of the proposed algorithms in the semi-supervised learning setting.

VI. Conclusion

Thus by using different Algorithms we will analysis agricultural data which is gathered from the government of India and create a predictive Analysis for the progress of farmers. The proposed system will give the following predictive analysis,

1) Suitable Weather for a particular crop
2) Suggestions regarding which crop should be grown
3) Suggestions for suitable pesticides and fertilizers

All rights reserved by *www.ijsrd.com*

REFERENCES

[1] Applying Machine Learning to Agricultural Data - ROBERT J. McQUEEN
[2] Use of Machine Learning Techniques to Help in Predicting Fertilizer Usage in Agriculture Production - H D Aparna, Dr. Kavitha K S, Dr. Kavitha Ca
[3] Robnik-Šikonja, Marko, and Igor Kononenko. "Theoretical and empirical analysis of ReliefF and RReliefF." Machine learning 53.1-2 (2003): 23-69.
[4] Siegel, Eric. Predictive analytics: The power to predict who will click, buy, lie, or die. Hoboken (NJ): Wiley, 2016.
[5] Demšar, Janez, et al. "Orange: From experimental machine learning to interactive data mining." Knowledge discovery in databases: PKDD 2004 (2004): 537-539.

All rights reserved by *www.ijsrd.com*