Eurasian Journal of Analytical Chemistry
ISSN: 1306-3057 OPEN ACCESS 2017 12 (3): 51-63

IRIS - Flower Classification

C. Geetha, Raghu Ram, Nazeer Vali

Received: 04 November 2016 ▪ Revised: 07 December 2016 ▪ Accepted: 06 January 2017

Abstract: The Machine learning is the subfield of computer science, "computers are having the ability to learn without being explicitly programmed". Evolved from the study of pattern recognition and computational learning theory in artificial intelligence machine learning explores the study and construction of algorithms that can learn from and make predictions on data such algorithms overcome following strictly static program instructions by making data-driven predictions or decisions, through building a model from sample inputs.

Keywords: Artificial Intelligence, IRIS Flower Species, Scikit Tool.

INTRODUCTION

Introduction about Machine Learning

The Machine learning is the subfield of computer science, according to Arthur Samuel in 1959 told "computers are having the ability to learn without being explicitly programmed". Evolved from the study of pattern recognition and computational learning theory in artificial intelligence machine learning explores the study and construction of algorithms that can learn from and make predictions on data such algorithms overcome following strictly static program instructions by making data-driven predictions or decisions, through building a model from sample inputs. Machine learning is employed in a range of computing tasks where designing and programming explicitly algorithms with good performance is difficult or unfeasible; example applications include email filtering, detection of network intruders, learning to rank and computer vision. Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data. It is a research field at the intersection of statistics, artificial intelligence and computer science and is also known as predictive analytics or statistical learning. There are two main categories of Machine learning.

They are Supervised and Unsupervised learning and here in this, the paper focuses on supervised learning. Supervised learning is a task of inferring a function from labeled training data. The training data consists of set of training examples. In supervised learning, each example is a pair of an input object and desired output value. A supervised learning algorithm analyze the training data and produces an inferred function, which can be used for mapping new examples. Supervised learning problems can be further grouped into regression and classification problems. Classification problem is when the output variable is a category, such as "red" or "blue" or "disease" and "no disease". Regression problem is when the output variable is a real value, such as "dollars" or "weight".

Scope of the Project

In this paper a novel method for Identification of Iris flower species is presented. It works in two phases, namely training and testing. During training the training dataset are loaded into Machine Learning Model and Labels are assigned. Further the predictive model, predicts to which species the Iris flower belongs to. Hence, the expected Iris species is labeled. This paper focuses on IRIS flower classification using Machine Learning with scikit tools. The problem statement concerns the identification of IRIS flower species on the basic of flower attribute measurements. Classification of IRIS data set would be discovering pattern from examining petal and sepal size of the IRIS flower and how the prediction was made from analyzing the pattern to form the class of IRIS flower. In this paper we train the Machine Learning Model with data and when unseen data is discovered the predictive model predicts the species using what it has learn from trained data scikit tools.

C. Geetha, Assistant Professor, Department of Computer Science and Engineering, BIST, BIHER, Bharath Institute of Higher Education & Research, Selaiyur, Chennai. E-mail: author@bharathuniv.ac.in

Raghu Ram, UG Scholar, Department of Computer Science and Engineering, BIST, BIHER, Bharath Institute of Higher Education & Research, Selaiyur, Chennai.

Nazeer Vali, UG Scholar, Department of Computer Science and Engineering, BIST, BIHER, Bharath Institute of Higher Education & Research, Selaiyur, Chennai.

The problem statement concerns the identification of IRIS flower species on the basic of flower attribute measurements. Classification of IRIS data set would be discovering pattern from examining petal and sepal size of the IRIS flower and how the prediction was made from analyzing the pattern to form the class of IRIS flower. In this paper we train the Machine Learning Model with data and when unseen data is discovered the predictive model predicts the species using what it has learn from trained data.

LITERATURE SURVEY

Basic Introduction to Machine Learning

Learning is a very important feature of Artificial Intelligence. Many scientist stried to explain and give a proper definition for learning. However, learning is not that easy to cover with few simple sentences. Many computer scientists, sociologists, logicians and other scientists discussed about this for a long time. Some scientists think learning is an adaptive skill so that the system can perform the similar task better in the next time (Simon 1987). Others claim that learning is a process of collecting knowledge (Feigenbaum 1977). Even though there is no proper definition for learning skill, we still need to give a definition for machine learning. In general, machine learning aims to find out how the computer algorithms can be improved automatically through experience (Mitchell1997). Machine learning has an important position in the field of Artificial Intelligence. At the beginning of development of Artificial Intelligence (AI), the AI system does not have a thorough learning ability so the whole system is not perfect. For instance, a computer cannot do self-adjustment when it faces problems. Moreover, the computer cannot automatically collect and discover new knowledge. The inference of the program needs more induction than deduction. Therefore, computer only can figure out already existing truths. It does not have the ability to discover a new logical theory, rules and so on.

Fundamental Structure of Machine Learning System

The environment represents a combination of information from external information source. That would include any information from persons or references materials and so on. It is the learning source for the whole machine learning system. The environment is responsible for transferring data to the system. The quality of the data is very important. In the reality, the data can be complex so it will be difficult for computer to process. In addition, the data can be incomplete, therefore the illation from the learning system is unauthentic. Learning is the procedure of transferring the information from the environment to knowledge. The environment will give the computer external information, and then the computer will go through all the information by using analysis, comprehensive induction and analogy to process this information to knowledge. At last, all the knowledge would be imported to the knowledge base. The knowledge base can be treated as the brain of the whole machine learning system. Different kinds of form and content of knowledge can have different influence on the designing of a machine learning system. Knowledge representation modes are eigenvector, First-order logic statements, production rule, and semantic system. Every mode has its own advantages and disadvantages. Therefore, when users want to design a machine leaning system, a good knowledge representation mode is very important for the whole system.

A proper knowledge representation mode should satisfy four basic requirements:

1. Strong expression
2. Easy theorization
3. Easy to modify the knowledge base
4. Easy to expand the knowledge representation

Moreover, a machine learning system cannot create new knowledge from nothing. It always needs original knowledge to understand the information from environment. The complexity of knowledge is different depending on the different learning tasks. Some tasks are quite easy, so the system does not need too much information. If the tasks are quite difficult, the system will need more information to learn.

Feedback

After the execution, the execution system can evaluate the leaning task, and then give feedback information to the learning process. The learning process will try to decide whether to collect information from environment to modify or improve the knowledge in knowledge base or not based on the feedback.

The applications of Machine Learning

Machine learning as a very likely approach to achieve human-computer integration and can be applied in many computer fields. Machine learning is not a typical method as it contains many different computer

algorithms. Different algorithms aim to solve different machine learning tasks. At last, all the algorithms can help the computer to act more like a human. Machine learning is already applied in many fields, for instance, pattern recognition, Artificial Intelligence, computer vision, data mining, text categorization and so on. Machine learning gives a new way to develop the intelligence of the machines. It also becomes an easier way to help people to analyse data from huge data sets.

The description of Machine Learning forms

A learning method is a complicated topic which has many different kinds of forms. Everyone has different methods to study, so does the machine. We can categorize various machine learning systems by different conditions. In general, we can separate learning problems in two main categories: supervised learning and unsupervised learning.

Supervised Learning

Supervised learning is a commonly used machine learning algorithm which appears in many different fields of computer science. In the supervised learning method, the computer can establish a learning model based on the training dataset. According to this learning model, a computer can use the algorithm to predict or analyze new information. By using special algorithms, a computer can find the best result and reduce the error rate all by itself. Supervised learning is mainly used for two different patterns: classification and regression. In supervised learning, when a developer gives the computer some samples, each sample is always attached with some classification information. The computer will analyze these samples to get learning experiences so that the error rate would be reduced when a classifier does recognitions for each patterns. Each classifier has a different machine learning algorithm. For instance, a neutral network algorithm and a decision tree learning algorithm suit to two different classifiers. They have their own advantages and disadvantages so that they can accomplish different learning objectives.

Unsupervised Learning

Unsupervised learning is also used for classification of original data. The classifier in the unsupervised learning method aims to find the classification information for unlabeled samples. The objective of unsupervised learning is to let the computer learn it by itself. We do not teach the computer how to do it. The computer is supposed to do analyzing from the given samples. In unsupervised learning, the computer is not able to find the best result to take and also the computer does not know if the result is correct or not. When the computer receives the original data, it can find the potential regulation within the information automatically and then the computer will adopt this regulation to the new case. That makes the difference between supervised learning and unsupervised learning. In some cases, this method is more powerful than supervised learning. That is because there is no need to do the classification for samples in advance. Sometimes, our classification method may not be the best one. On the other hand, a computer may find out the best method after it learns it from samples again and again.

Machine Learning in Pattern Recognition

As mentioned above, the method of machine learning can also be used in pattern recognition. In fact, pattern recognition really needs machine learning to achieve its objective. Both supervised learning and unsupervised learning are useful for pattern recognition, for example, in this thesis, K-means clustering algorithm in unsupervised learning. The K-means clustering algorithm is always used for image segmentation. The image segmentation is so important for image pattern recognition. Because of the technology of image segmentation, it is easier to do the image analyzing so that it will achieve much better results for image pattern recognition. Moreover, the technology of machine learning has been used in almost every field in pattern recognition. For example, image pattern recognition, voice recognition, fingerprint recognition, character recognition and so on. They all need machine learning algorithms to select features from the objects and to do the analyzing.

Basic Introduction to Pattern Recognition

Pattern Recognition is a fundamental human intelligence. In our daily life, we always do ‘pattern recognition’, for instance, we recognize faces and images. Basically, pattern recognition refers to analyzing information and identifying for any kind of forms of visual or phenomenon information. Pattern recognition can describe, recognize, classify and explain the objects or the visual information. As machine learning, pattern recognition, can be treated as two different classification methods: supervised classification and unsupervised classification. They are quite similar to supervised learning and unsupervised learning. As supervised classification needs a teacher that gives the category of samples, the unsupervised classification is doing it the other way around. Pattern recognition is related to statistics, psychology, linguistics, computer science.

OBJECTIVES OF THE PROJECT

Objectives

It is observed from the literature survey that the existing algorithms face several difficulties like the computational power is increases when run Deep Learning on latest computation, requires a large amount of data, is extremely computationally expensive to train, they do not have explanatory power that is they may extract the best signals to accurately classify and cluster data, but cannot get how they reached a certain conclusion.

Neural Networks cannot be retrained that is it is impossible to add data later. To address these problems the current work is taken up to develop a new technique for Identification of Iris Flower Species using Machine Learning

Problem Statement

To design and implement the Identification of Iris Flower species using machine learning using Python and the tool Scikit-Learn biology. It plays an important role in Artificial Intelligence and image processing

Python 3.6.0

Python is an interpreted, high-level, general-purpos programming language. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales. Van Rossum led the language community until stepping down as leader in July 2018.

Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, it also has a comprehensive standard library. Python interpreters are available for many operating systems. C Python, the reference implementation of Python, is open source software and has a community-based development model, as do nearly all of Python's other implementations. Python and C Python are managed by the non-profit Python Software Foundation.

Installation

https://www.python.org/downloads/release/python-360/

The above link provide the installation procedure to install python in pc.

Anaconda

The open-source Anaconda Distribution is the easiest way to perform Python/R data science and machine learning on Linux, Windows, and Mac OS X. With over 11 million users worldwide, it is the industry standard for developing, testing, and training on a single machine, enabling individual data scientists to:

Quickly download 1,500+ Python/R data science packages Manage libraries, dependencies, and environments with Conda Develop and train machine learning and deep learning models with scikit-learn, Tensor Flow, and Theano Analyze data with scalability and performance with Dask, NumPy, pandas, and Numba Visualize results with Matplotlib, Bokeh, Datashader, and Holoviews

Installation

The below link will provide the installation of anaconda. https://www.anaconda.com/distribution/Scikit-Learn 0.18.1:

Scikit-learn (formerly scikits.learn) is a free software machine learning library for the python programming language. It features various classification, regression and clustering algorithms including support vector machines, randon forests, gradient boosting, k-means and DBSCAN and is designed to interoperate with the python numerical and specific libraries NumPy.

EXISTING SYSTEM

The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper. The use of multiple measurements in taxonomic problems as an example of linear discriminant analysis.

It is sometimes called Anderson’s Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris Flower of three related species.

Two of the three species were collected in Gaspe Peninsula all from the same pasture, and picked on the same day and measured at the same time by the same person with same apparatus. The data set consists of 50 samples from each of three species of Iris that is

1. Iris Setosa
2. Iris Virginica
3. Iris Versicolor.

Four features were measured from each sample. They are

1. Sepal Length
2. Sepal Width
3. Petal Length
4. Petal Width.

All these four parameters are measured in Centimeters. Based on the combination of these four features, the species among three can be predicted.

Summary Statistics

Min Max Mean SD Class Correlation

Sepal length: 4.3 7.9 5.84 0.83 0.7826

Sepal width: 2.0 4.4 3.05 0.43 -0.4194

Petal length: 1.0 6.9 3.76 1.76 0.9490 (high!)

Petal width: 0.1 2.5 1.20 0.76 0.9565 (high!)

PROPOSED SYSTEM

To design and implement the Identification of Iris Flower species using machine learning using Python and the tool Scikit-Learn.

Advantages of Proposed System

The below implementation using Python and its libraries make the system fast, accurate and help for further Advancements.

Work Carried Out

- **Data collection:** Various datasets of Iris Flower are collected. There are totally 150 datasets belonging to three different species of Iris Flower that is Setosa, Versicolor and Virginca.
- **Literature survey:** Studied various papers related to proposed work.
- **Algorithms developed**

1. A K-Nearest Neighbor Algorithm to predict the species of Iris Flower.
2. A Logistic Regression Algorithm to predict the species of Iris Flower.

BLOCK DIAGRAM OF THE PROPOSED WORK

The proposed method comprises of sub-phases that is Loading and Modeling as schematic diagram of the proposed model is given in figure1.0.

The detailed description of each processing step is presented in the following sub sections.

Loading

Various datasets of Iris Flower are collected. There are totally 150 datasets belonging to three different species of Iris Flower that is Setosa, Versicolor and Virginca. The collected Iris Datasets are loaded into the Machine Learning Model. Scikit-learn comes with a few standard datasets, for instance the Iris Dataset for Classification. The load_iris function is imported from Scikit-learn. The load_iris function is run and save the return value in an object called “iris”. The iris object is of type “sklearn.datasets.base.bunch”, here bunch is a special object in scikit-learn to store datasets and attributes. The few attributes of iris object are data, feature names, target, target names etc. The iris.data is of type numpy.ndarray and is stored in a Feature Matrix say “X”. Here X is a two dimensional array, the two dimensions of it are number of observations and number of features. The iris.target is of type numpy.ndarray and is stored in a Response Vector say “y”. Here y is aone dimensional array, the one dimension of it is number of observations. In Scikit-learn, each Row is an observation that is sample,

example, instance, record and each Column is a feature that is predictor, attribute, independent variable, input, regressor, covariate.

Four key requirements for working with data in Scikit-learn, Features and Response are separate objects.Features and Response should be numeric. Features and Response should be NumPy arrays. Features and Response should have specific shapes. The shapes of X and y here is (150, 4) and (150,) respectively that is 150 observations and 4 features. The Target Names for Iris dataset are ['setosa','versicolor','virginica'] and the Feature Names for Iris dataset are ['sepal length (cm)', 'sepal width (cm)', 'petal length (cm)', 'petal width (cm)'].

Modeling

Scikit-learn has four step Modeling Pattern.

Step 1: Import the class which is needed from Scikit-learn

In first case, we import KNeighbors Classifier from Sklearn Neighbors. Sklearn Neighbors provides functionality for supervised neighbors-based learning methods. The principle behind nearest neighbor methods is to find a predefined number of training samples closest in distance to the new point, and predict the label from these. In second case, we import Logistic Regression from Sklearn Linear Model module. The module implements generalized linear models. It includes Ridge regression, Bayesian Regression Lasso and Elastic Net estimators computed with Least Angle Regression and coordinate descent. It also implements Stochastic Gradient Descent related algorithms.

Step 2: Here we Instantiate the Estimator

Scikit-learn refers its model as Estimator. A estimator is an object that fits a model based on some training data and is capable of inferring some properties on new data. It can be, for instance, a classifier or a regressor. Instantiation concerns the creation of an object that is Instantiate the object “Estimator”.

Here in first case, Instantiate the Estimator means make instance of KNeighborsClassifier Class. The object here has various parameters that is KNeighborsClassifier (algorithm='auto', leaf_size=30, metric='minkowski',metric_params=None, n_jobs=1, n_neighbors=5, p=2, weights='uniform').

Here in first case, Instantiate the Estimator means make instance of Logistic Regression Class. The object here has various parameters that is LogisticRegression (C=1.0, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, max_iter=100, multi_ class='ovr', n_jobs=1, penalty='l2', random_state=None, solver='liblinear', tol=0.0001, verbose=0, warm_start=False).

Now, there are Objects that knows how to do K-Nearest Neighbor and Logistic Regression and waiting for user to give data. The name of the Estimator object can be anything, we can tend to choose the name that reflex the model it represents, “est” short of estimator or “clf” short of classifier. The Tuning Parameter that is Hyper Parameter can be specified at this step. For example, n_neighbors is a tuning parameter. All the other parameters which are not specified here are set to their default values. By printing the Estimator object we can get all the parameters and its values.

Step 3: Fit the Model with Data

This is the model training step. Here the Model learns the relationship between the features X and response y. Here fit method is used on the object of type KNeighborsClassifier Class and Logistic Regression Class. The fit method takes two parameters that is the feature matrix X and response vector y.

The model is under fitting or over fitting the training data. The model is under fitting the training data when the model performs poorly on the training data. This is because the model is unable to capture the relationship between the input examples (often called X) and the target values (often called Y). The model is over fitting your training data when you see that the model performs well on the training data but does not perform well on the evaluation data. This is because the model is memorizing the data it has seen and is unable to generalize to unseen examples.

Step 4: Predict the response for a new observation

In this step, the response is predicted for a new observation. Here a new observation means “out-of-sample” data. Here, it’s inputting the measurements for unknown iris and asking the fitted model to predict the iris species based on what it has learnt in previous step.

The predict method is used on the KNeighbors Classifier Class object and Logistic Regression Class object and pass the features of Unknown iris as a Python list. Actually, expects numpy array but it still works with list since numpy automatically converts it to an array of appropriate.

UML DIAGRAMS

The following images are Use Case and Class Diagram:

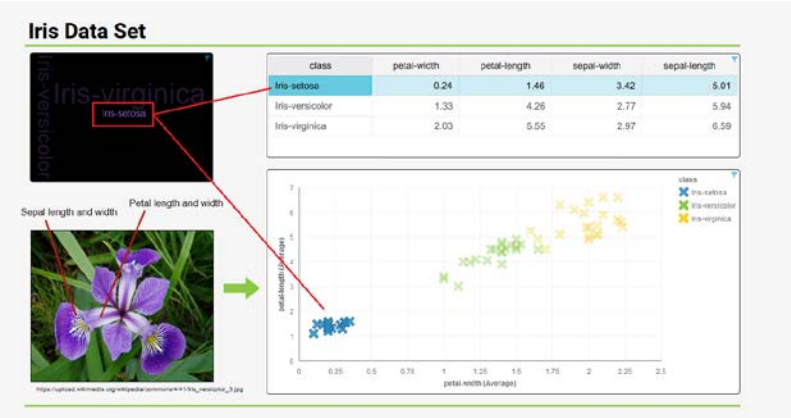

Figure 1

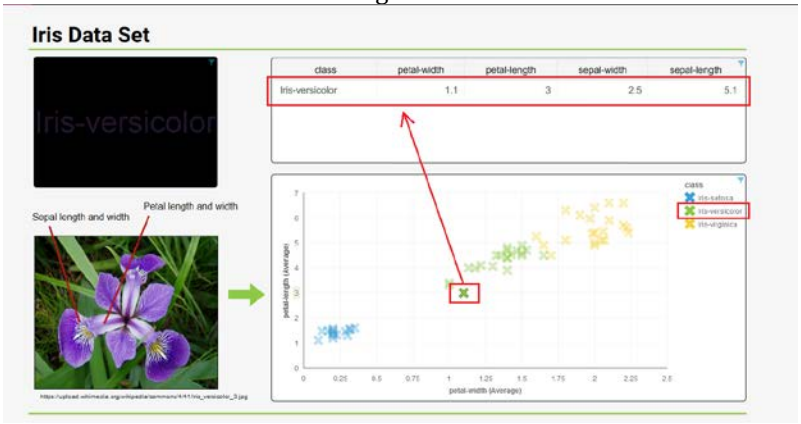

Figure 2

MODULES SPECIFICATION

List of Modules

- Training Dataset
- Machine Learning Algorithms
- Classification
- New Dataset
- Prediction

Screenshots

Sample Data for Iris classification in figure 1.4

Sepal length	Sepal width	Petal length	Petal width	Species
5.0	3.5	1.6	0.6	*I. setosa*
5.1	3.3	1.7	0.5	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
5.7	4.4	1.5	0.4	*I. setosa*
5.4	3.9	1.3	0.4	*I. setosa*

Description

The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper the use of multiple measurements in taxonomic problems as an example of linear discriminant analysis. It is sometimes called Anderson's *Iris* data set because Edgar Anderson collected the data to quantify the morphologic variation of *Iris* flowers of three related species. Two of the three species were collected

in the Gaspé Peninsula "all from the same pasture, and picked on the same day and measured at the same time by the same person with the same apparatus".

K-Nearest Neighbor Algorithm

K-Nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection. It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data).We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.

As an example, consider the following table of data points containing two features in figure 3:

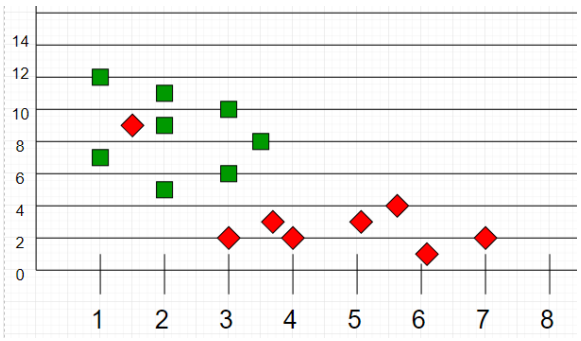

Figure 3

Now, given another set of data points (also called testing data), allocate these points a group by analyzing the training set. Note that the unclassified points are marked as 'White'.

Algorithm

Let m be the number of training data samples. Let p be an unknown point.

1. Store the training samples in an array of data points arr[]. This means each element of this array represents a tuple (x, y).
2. for i=0 to m:
3. Calculate Euclidean distance d(arr[i], p).
4. Make set S of K smallest distances obtained. Each of these distances correspond to an already classified data point.
5. Return the majority label

ALGORITHM SPECIFICATION

Used Algorithms:

IMPLEMENTATION OF ALGORITHMS

K-Nearest Neighbors Algorithm

The k-Nearest Neighbors algorithm (or kNN for short) is an easy algorithm to understand and to implement, and a powerful tool to have at your disposal. The implementation will be specific for classification problems and will be demonstrated using the Iris flowers classification problem.

What is k-Nearest Neighbors

The model for kNN is the entire training dataset. When a prediction is required for a unseen data instance, the kNN algorithm will search through the training dataset for the k-most similar instances. The prediction attribute of the most similar instances is summarized and returned as the prediction for the unseen instance. The similarity measure is dependent on the type of data. For real-valued data, the Euclidean distance can be used. Other types of data such as categorical or binary data, Hamming distance can be used. In the case of regression problems, the average of the predicted attribute may be returned. In the case of classification, the most prevalent class may be returned.

How does k-Nearest Neighbors Work

The kNN algorithm is belongs to the family of instance-based, competitive learning and lazy learning algorithms. Instance-based algorithms are those algorithms that model the problem using data instances

(or rows) in order to make predictive decisions. The kNN algorithm is an extreme form of instance-based methods because all training observations are retained as part of the model. It is a competitive learning algorithm, because it internally uses competition between model elements (data instances) in order to make a predictive decision. The objective similarity measure between data instances causes each data instance to compete to "win" or be most similar to a given unseen data instance and contribute to a prediction. Lazy learning refers to the fact that the algorithm does not build a model until the time that a prediction is required. It is lazy because it only does work at the last second. This has the benefit of only including data relevant to the unseen data, called a localized model. A disadvantage is that it can be computationally expensive to repeat the same or similar searches over larger training datasets. Finally, kNN is powerful because it does not assume anything about the data, other than a distance measure can be calculated consistently between any two instances. As such, it is called non-parametric or non-linear as it does not assume a functional form. First let us try to understand what exactly does K influence in the algorithm. If we see the last example, given that all the 6 training observation remain constant, with a given K value we can make boundaries of each class. These boundaries will segregate RC from GS. The same way, let's try to see the effect of value "K" on the class boundaries. Following are the different boundaries separating the two classes with different values of K in figure 4.

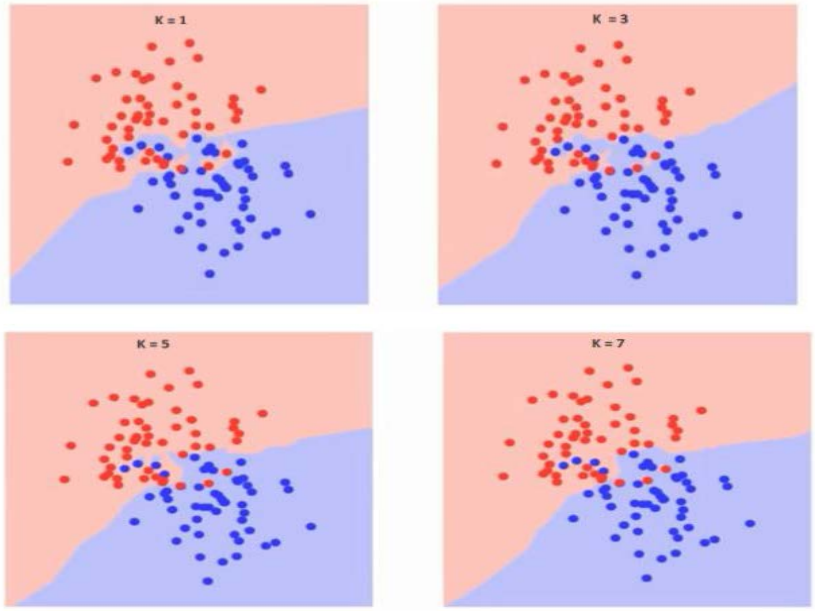

Figure 4

If you watch carefully, you can see that the boundary becomes smoother with increasing value of K. With K increasing to infinity it finally becomes all blue or all red depending on the total majority. The training error rate and the validation error rate are two parameters we need to access on different K-value.

LOGISTIC REGRESSION

Algorithm

Logistic Regression is a type of regression that predicts the probability of occurrence of an event by fitting data to a logit function (logistic function). Like many forms of regression analysis, it makes use of several predictor variables that may be either numerical or categorical. For instance, the probability that a person has a heart attack within a specified time period might be predicted from knowledge of the person's age, sex and body mass index. This regression is quite used in several scenarios such as prediction of customer's propensity to purchase a product or cease a subscription in marketing applications and many others

What is Logistic Regression?

Logistic Regression, also known as Logit Regression or Logit Model, is a mathematical model used in statistics to estimate (guess) the probability of an event occurring having been given some previous data. Logistic Regression works with binary data, where either the event happens (1) or t(0). So given some feature x it tries to he event does not happen find out whether some or not. So y can either be 0 or 1. In the case where the event happens, y is given the value 1. If the event does not happen, then y is given the

value of 0. For example, if y represents whether a sports teams wins a match, then y will be 1 if they win the match or y will be 0 if they do not. This is known as Binomial Logistic Regression. There is also another form of Logistic Regression which uses multiple values for the variable y.This form of Logistic Regression is known as Multinomial Logistic Regression.

How does logistic Regression work?

Logistic Regression uses the logistic function to find a model that fits with the data points. The function gives a 'S'shaped curve to model the data. The curve is restricted between 0 and 1, so it is easy to apply when y is binary. Logistic Regression can then model events better than linear regression.

CONCLUSION

With the rapid development of technology, AI has been applied in many fields. Machine learning is the most fundamental approach to achieve AI. This thesis describes the work principle of machine learning, two different learning forms of machine learning and an application of machine learning. In addition, a case study of Iris flower recognition to introduce the workflow of machine learning in pattern recognition is shown. In this case, the meaning of pattern recognition and how the machine learning works in pattern recognition has been described. The K-means algorithm, which is a very simple machine learning algorithm from the unsupervised learning method is used. The work also shows how to use software to learn machine learning.

FUTURE ENHANCEMENT

The Iris recognition case study above shows that the Machine Learning algorithm works well in this pattern recognition. The speed of computing is fast and the result is acceptable. However, the K-means clustering algorithm is just one of the clustering algorithm in unsupervised learning. There are more algorithms for different work objectives in different scientific fields. As it is mentioned above, Machine Learning can be separated into supervised learning and unsupervised learning. However, sometimes, a whole dataset have both labeled data and unlabeled data. In order to process this kind of dataset, a new learning method called Semi-supervised (SSL) Learning has become a research hotspot. Because of this learning method, both machine learning and pattern recognition have a new research direction. It saves a lot of time and human resource to label those large amounts of unlabeled data. The SSL is also significant on improving learning performance of a computer. Moreover, a learning system always consists of two parts, learning and environment. The environment gives knowledge to the computer and the computer will transfer this knowledge and store them and select useful information to implements different learning objectives. Therefore, different learning strategies can also be separated into rote learning, learning from instruction, learning by deduction, learning by analog, explanation-based learning and learning from induction. All of them have different algorithms to process different work objectives. The implemented case in this thesis is only a simple example of machine learning and pattern recognition. Moreover, the K-means algorithm used in this thesis is a basic algorithm. However, if the data set has many feature dimensions and it is complicated, and if the learning objective is not that simple, the K-means algorithm cannot be used. Nowadays, GA (Genetic Algorithm), Artificial neural network and other machine learning algorithms have become more and more stable and useful. Many scientists are working on improving the performance of machine learning algorithms. The K-means has also its own improved parts. The K-means can also be used along with other algorithms, such as ISODATA, EM and K means++. A better machine learning algorithm can obtain better results for pattern recognition. As the technology of pattern recognition develops, it requires more professional and more perfect machine learning algorithms. In this case, machine learning has a huge potential for growth. In general, besides pattern recognition, machine learning can also be widely used in many fields of computer science and Artificial Intelligence. More and more.

REFERENCES

[1] Khanaa V., & Thooyamani K.P. (2013). Using triangular shaped stepped impedance resonators design of compact microstrip quad-band, *Middle - East Journal of Scientific Research,* 18(12), 1842-1844.

[2] Asiri S., Sertkol M., Güngüneş H., Amir M., Manikandan A., Ercan I., & Baykal A. (2018). The Temperature Effect on Magnetic Properties of NiFe 2 O 4 Nanoparticles. *Journal of Inorganic and Organometallic Polymers and Materials,* 28(4), 1587-1597.

[3] Thaya, R., Malaikozhundan, B., Vijayakumar, S., Sivakamavalli, J., Jeyasekar, R., Shanthi, S., Vaseeharan B., Ramasamy P., & Sonawane, A. (2016). Chitosan coated Ag/ZnO nanocomposite

and their antibiofilm, antifungal and cytotoxic effects on murine macrophages. *Microbial pathogenesis*, *100*, 124-132.

[4] Kolanthai, E., Ganesan, K., Epple, M., & Kalkura, S.N. (2016). Synthesis of nanosized hydroxyapatite/agarose powders for bone filler and drug delivery application. *Materials Today Communications*, *8*, 31-40.

[5] Thilagavathi, P., Manikandan, A., Sujatha, S., Jaganathan, S.K., & Arul Antony, S. (2016). Sol–Gel Synthesis and Characterization Studies of NiMoO4 Nanostructures for Photocatalytic Degradation of Methylene Blue Dye. *Nanoscience and Nanotechnology Letters*, *8*(5), 438-443.

[6] Thamotharan C., Prabhakar S., Vanangamudi S., & Anbazhagan R. (2014). Anti-lock braking system in two wheelers. *Middle - East Journal of Scientific Research, 20*(12), 2274-2278.

[7] Thamotharan C., Prabhakar S., Vanangamudi S., Anbazhagan R., & Coomarasamy C. (2014). Hydraulic rear drum brake system in two wheeler. *Middle - East Journal of Scientific Research, 20*(12), 1826-1833.

[8] Vanangamudi S., Prabhakar S., Thamotharan C., & Anbazhagan R. (2014). Collision control system in cars. *Middle - East Journal of Scientific Research, 20*(12), 1799-1809.

[9] Vanangamudi S., Prabhakar S., Thamotharan C., & Anbazhagan R. (2014). Drive shaft mechanism in motor cycle. *Middle - East Journal of Scientific Research, 20*(12), 1810-1815.

[10] Anbazhagan R., Prabhakar S., Vanangamudi S., & Thamotharan C. (2014). Electromagnetic engine. *Middle - East Journal of Scientific Research, 20*(3), 385-387.

[11] Kalaiselvi, V.S., Prabhu, K., & Mani Ramesh, V.V. (2013). The association of serum osteocalcin with the bone mineral density in post-menopausal women. *Journal of clinical and diagnostic research: JCDR*, *7*(5), 814-816.

[12] Kalaiselvi, V.S., Saikumar, P., & Prabhu, K. (2012). The anti mullerian hormone-a novel marker for assessing the ovarian reserve in women with regular menstrual cycles. *Journal of clinical and diagnostic research: JCDR*, *6*(10), 1636-1639.

[13] Arul, T.K., Manikandan, E., Ladchumananandasivam, R., & Maaza, M. (2016). Novel polyvinyl alcohol polymer based nanostructure with ferrites co-doped with nickel and cobalt ions for magneto-sensor application. *Polymer International*, *65*(12), 1482-1485.

[14] Das, M.P., & Kumar, S. (2015). An approach to low-density polyethylene biodegradation by Bacillus amyloliquefaciens. *3 Biotech*, *5*(1), 81-86.

[15] Vanangamudi S., Prabhakar S., Thamotharan C. & Anbazhagan R., (2014). Turbo charger in two wheeler engine. *Middle - East Journal of Scientific Research, 20*(12), 1841-1847, 2014.

[16] Vanangamudi S., Prabhakar S., Thamotharan C., & Anbazhagan R. (2014). Design and calculation with fabrication of an aero hydraulwicclutch. *Middle - East Journal of Scientific Research, 20*(12), 1796-1798.

[17] Saravanan T., Raj M.S., & Gopalakrishnan K. (2014). VLSI based 1-D ICT processor for image coding. *Middle - East Journal of Scientific Research, 20*(11), 1511-1516.

[18] Ajona M., & Kaviya B. (2014). An environmental friendly self-healing microbial concrete. *International Journal of Applied Engineering Research, 9*(22), 5457-5462.

[19] Hemalatha, R., & Anbuselvi, S. (2013). Physicochemical constituents of pineapple pulp and waste. *Journal of Chemical and Pharmaceutical Research*, *5*(2), 240-242.

[20] Langeswaran, K., Revathy, R., Kumar, S.G., Vijayaprakash, S., & Balasubramanian, M.P. (2012). Kaempferol ameliorates aflatoxin B1 (AFB1) induced hepatocellular carcinoma through modifying metabolizing enzymes, membrane bound ATPases and mitochondrial TCA cycle enzymes. *Asian Pacific Journal of Tropical Biomedicine*, *2*(3), S1653-S1659.

[21] Masthan, K.M.K., Babu, N.A., Dash, K.C., & Elumalai, M. (2012). Advanced diagnostic aids in oral cancer. *Asian Pacific Journal of Cancer Prevention*, *13*(8), 3573-3576.

[22] Asiri S., Güner S., Demir A., Yildiz A., Manikandan A., & Baykal A. (2018). Synthesis and Magnetic Characterization of Cu Substituted Barium Hexaferrites. *Journal of Inorganic and Organometallic Polymers and Materials, 28*(3), 1065-1071.

[23] Vellayappan, M.V., Jaganathan, S.K., & Manikandan, A. (2016). Nanomaterials as a game changer in the management and treatment of diabetic foot ulcers. *RSC Advances*, *6*(115), 114859-114878.

[24] Vellayappan, M.V., Venugopal, J.R., Ramakrishna, S., Ray, S., Ismail, A.F., Mandal, M., Manikandan A., Seal S., & Jaganathan, S.K. (2016). Electrospinning applications from diagnosis to treatment of diabetes. *RSC Advances*, *6*(87), 83638-83655.

[25] Bavitra, K., Sinthuja, S., Manoharan, N., & Rajesh, S. (2015). The high efficiency renewable PV inverter topology. *Indian Journal of Science and Technology*, *8*(14), 1.

[26] Vanangamudi S., Prabhakar S., Thamotharan C., & Anbazhagan R. (2014). Design and fabrication of dual clutch. *Middle - East Journal of Scientific Research, 20*(12), 1816-1818.

[27] Sandhiya K., & Kaviya B. (2014). Safe bus stop location in Trichy city by using gis. *International Journal of Applied Engineering Research, 9*(22), 5686-5691.

[28] Selva Kumar, S., Ram Krishna Rao, M., Deepak Kumar, R., Panwar, S., & Prasad, C. S. (2013). Biocontrol by plant growth promoting rhizobacteria against black scurf and stem canker disease of potato caused by Rhizoctonia solani. *Archives of Phytopathology and Plant Protection*, *46*(4), 487-502.

[29] Sharmila, S., & Jeyanthi Rebecca, L. (2012). GC-MS Analysis of esters of fatty acid present in biodiesel produced from Cladophora vagabunda. *Journal of Chemical and Pharmaceutical Research*, *4*(11), 4883-4887.

[30] Ramkumar, M., Rajasankar, S., Gobi, V.V., Dhanalakshmi, C., Manivasagam, T., Thenmozhi, A.J., Essa M.M., Kalandar A., & Chidambaram, R. (2017). Neuroprotective effect of Demethoxycurcumin, a natural derivative of Curcumin on rotenone induced neurotoxicity in SH-SY 5Y Neuroblastoma cells. *BMC complementary and alternative medicine*, *17*(1), 217.

[31] Selvi S.A., & Sundararajan M. (2016). A combined framework for routing and channel allocation for dynamic spectrum sharing using cognitive radio. *International Journal of Applied Engineering Research, 11*(7), 4951-4953.

[32] Krupaa R.J., Sankari S.L., Masthan K.M.K., & Rajesh E. (2015). Oral lichen planus: An overview, *Journal of Pharmacy and Bioallied Sciences, 7,* S158-S161.

[33] Srividya T., & Saritha B. (2014). Strengthening on RC beam elements with GFRP under flexure. *International Journal of Applied Engineering Research, 9*(22), 5443-5446.

[34] Kumar J., Sathish Kumar K., & Dayakar P. (2014). Effect of microsilica on high strength concrete, *International Journal of Applied Engineering Research, 9*(22), 5427-5432.

[35] Saraswathy R., & Saritha B. Planning of integrated satellite township at Thirumazhisai. *International Journal of Applied Engineering Research, 9*(22), 5558-5560.

[36] Saritha B., Ilayaraja K., & Eqyaabal Z. Geo textiles and geo synthetics for soil reinforcement, *International Journal of Applied Engineering Research, 9*(22), 5533-5536.

[37] Iyappan L., & Dayakar P. (2014). Identification of landslide prone zone for coonoor taluk using spatialtechnology, *International Journal of Applied Engineering Research, 9*(22), 5724-5732, 2014.

[38] Arunachalam, A.R. (2014). Bringing out the effective learning process by analyzing of e-learning methodologies. *Indian Journal of Science and Technology*, *7*, 41-43.

[39] Wasy, A., Balakrishnan, G., Lee, S.H., Kim, J.K., Kim, D.G., Kim, T.G., & Song, J.I. (2014). Argon plasma treatment on metal substrates and effects on diamond-like carbon (DLC) coating properties. *Crystal Research and Technology*, *49*(1), 55-62.

[40] Jaganathan, S., Mani, M., Ismail, A., & Ayyar, M. (2017). Manufacturing and characterization of novel electrospun composite comprising polyurethane and mustard oil scaffold with enhanced blood compatibility. *Polymers*, *9*(5), 163.

[41] Yan, S., Gao, M., Qi, B., & Jiang, X. (2014). Blast Wave Propagation and Casualty Distribution Evaluation in the Subway Station Subjected to Internal Blast Loading. *The SIJ Transactions on Advances in Space Research & Earth Exploration, 2*(1), 6-11.

[42] Geetha, K., Preethy, C., and Thenmozhi, P. (2017). Simulation Model of Solar Induction Motor Drive System Using SVPWM Technique. *Bonfring International Journal of Power Systems and Integrated Circuits, 7*(1), 1-6.

[43] Archana Lal, P. (2014). A Neural Network Based Analysis of Altered Fingerprints. International Scientific *Journal on Science Engineering & Technology*, *17*(9), 863-868.

[44] AlaguPandian, P., Sakthivel, K., Sheik Alavudeen, K., & R.LakshmiPriya. R. (2017). A Low Power Efficient Design of Full Adder Using Transmission Gate. *International Journal of Communication and Computer Technologies, 5*(1), 1-5.

[45] SakthiPriya V., & Vijayan, M., (2017). Automatic Street Light Control System Using WSN Based on Vehicle Movement and Atmospheric Condition. *International Journal of Communication and Computer Technologies, 5*(1), 6-11.

[46] Sowmiya, E., Dr.Chandrasekaran, V., & Sathya, T. (2017). Sensor Node Failure Detection Using Round Trip Delay in Wireless Sensor Network. *International Journal of Communication and Computer Technologies, 5*(1), 12-16.

[47] Senthil Kumar, B., & Dr.Srivatsa, S.K.(2014). Opportunistic Channel Access Algorithm Based on Hidden Semi Markov Model for Cognitive Radio Networks. *Bonfring International Journal of Research in Communication Engineering, 4*(2), 17-21.

[48] Angeline, D.M.D., (2013). Association Rule Generation for Student Performance Analysis using Apriori Algorithm. *The SIJ Transactions on Advances in Space Research & Earth Exploration, 1*(1), 16-20.

[49] Preethi, L., & Dr.Periyasamy, S. (2018). Enhanced Scalable Learning for Identifying and Ranking for Big Data Using Social Media Factors. *Bonfring International Journal of Software Engineering and Soft Computing, 8*(1), 31-35.

[50] Saikong, W., & Kulworawanichpong, T. (2014).Voltage Stability Assessment in DC Railways with Minimum Headway Consideration. *The SIJ Transactions on Computer Networks & Communication Engineering (CNCE), 2*(4), 1-6.