Python and Machine Learning

Mrs. Smita Desai & Ms. Shreya Desai
1 Department of Computer Science, Bharatesh College of Computer Applications, Belagavi, Karnataka, India
2 Electronic and communications, Gogte Institute of Technology, Belagavi, Karnataka, India

Abstract—*Python is gaining the popularity. In this paper we have tried to find the characteristics of Python language that helps it gain the attention of the programmers. This paper is an initiative to review the role of Python in machine learning. Machine learning is the most happening technology of todays'sworld.The main aim of Machine Learning is to allow the computers learn automatically without human intervention and adjust its actions accordingly. In this paper main focus is on the popularity of the Python as a language preferred by the developers forMachine Learning. We have included the statistics of other computer languages and Python to support the popularity of Python in machine learning. The main focus is on the topic " Python - the ideal language for Machine Learning"*

Index Terms—Machine Learning with Python, Popular language for Machine Learning, Python programming, Python then and now, Python implementations.

1. Introduction

Python's role in Machine Learning is gaining more and more attention. The main reason could be the simplicity of python language and its rich set of libraries.

Python is a popularly used general purpose high level language. Python offers three main paradigm to its developers namely, object orientation, functional and structured programming. Python has multiple implementations, the most popular one is Cpython- a default implementation. Other python implementations are Jpython – scripted in Java, PyPy – written in Rpython, Iron-Pyhton written in C#. These implementations work in the native language they are written in but are capable of interacting with other languages through the use of modules. Most of these implementations are pen source and free.

Python is the most preferred language over other languages like Java, C and C++ by the programmers.Most of the software development companies prefer Python as a programming language because of its fewer coding lines and versatile features. Some applications of Python are:

- Image processing and Graphic design
- Game development
- Web frameworks and web applications
- Enterprise applications
- Language development

Machine Learning is an application of Artificial Intelligence (AI). It gives the ability to learn automatically from the experience and improve accordingly without being explicitly programmed. The process of learning starts with the data or observations. The iterative nature allows models to be able to adapt the new data exposed to them independently.

2. Merits and Demerits of Python

2.1 Advantages

- **Extensive Support Libraries**

Most of the frequently used tasks are already scripted into its standard libraries making the Python code petite.

- **Open Source and Community Development**

Python is developed under OSI approved open source license, thus making it available free to use and distribute even for commercial purpose.Python development is driven by community which collaborates for its code through hosting conferences, mailing lists etc.

- **Third Party Module Support**

Third party modules present in Python Package Index (PyPI) makes it possible to interact with most of the other languages and platforms.

- **User friendly Data Structure**

Python has built-in dictionary ad list data structure, which can be used to construct run time data structures quickly.

- **Simple Code**

Python's simple to learn syntax and excellent readability style makes it very easy for the beginners to understand and code.

- **Integration Feature**

Python integrates the Enterprise Application Integration which makes it easy to develop web services by invoking COM or COBRA components. It also has powerful control capabilities as it calls directly through C, C++ of Java via Jython. Python also processes XML and other markup languages as it can run on all modern operating systems through same byte code.

2.2 Disadvantages

- Low Speed
- Week Mobile Computing
- Under developed Database Access layers

3. Machine Learning

Machine learning is the technology that allows machines to learn on its own from the past experience, data or examples. There has been significant advances in machine learning due to the advancements in technology, easy and increased availability of data and the computing power.

In addition to image processing and voice recognition systems which are machine learning driven systems, it holds the developments in diverse range of fields that includes education, health care and many more. It could also support scientific advances by handling the large datasets.

3.1 Machine Learning Types

The three key branches of machine learning are:

- **Supervised learning**

In this type, the machine is trained with labelled data. The label categories each data points into one or more groups, such as ‘fruits’ or ’vegetables’. The machine learns how this data is structured(known as training data) and uses this to predict the categories of new or ‘test’ data.

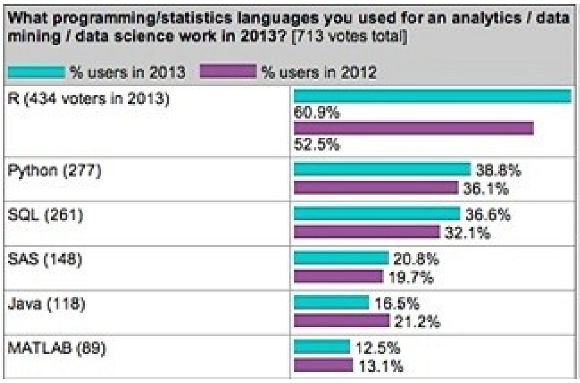
- **Unsupervised learning**

Unsupervised learning is learning without labels. It aims to detect the characteristics that make the data more or less similar to each other. For example, creating clusters and assigning data to these clusters.

- **Reinforcement learning**

Reinforcement learning lies between unsupervised and supervised learning. It focuses on learning from experience.

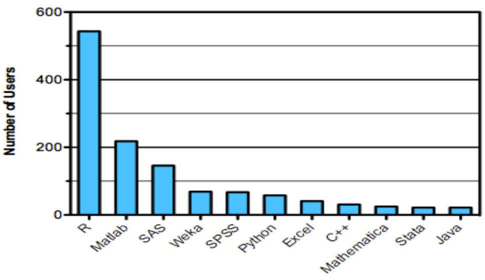
Following is the poll of languages by kdNuggets



What programming/statistics languages you used for an analytics / data mining / data science work in 2013? [713 votes total]

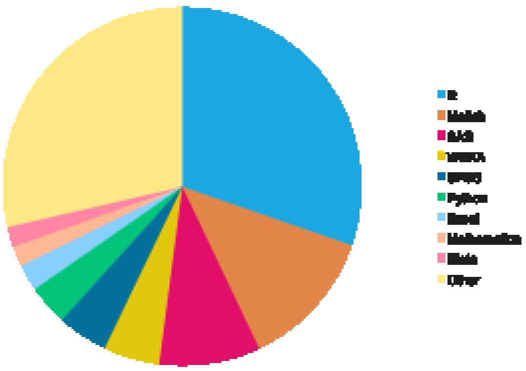
▬ % users in 2013 ▬ % users in 2012

Language	% users in 2013	% users in 2012
R (434 voters in 2013)	60.9%	52.5%
Python (277)	38.8%	36.1%
SQL (261)	36.6%	32.1%
SAS (148)	20.8%	19.7%
Java (118)	16.5%	21.2%
MATLAB (89)	12.5%	13.1%



Kaggle offer machine learning completions and have polled their user base as to the tools and programming languages used by partisans in competition.

Kaggle’s poll results in 2011 is as follows:



Popular programming languages on Kaggle

While many machine learning algorithms have been around for a long time, the ability to automatically apply complex mathematical calculations to big data – over and over, faster and faster is a recent development.

Few examples of machine learning applications:

- The self-driving Google car
- Online recommendation offers such as Amazon
- Knowing what customers are saying about you on Twitter – Machine learning combined with linguistic rule creation
- Fraud detection

3.2 Why is Machine Learning Important?

The popularity of machine learning is due to the same factors that have made data mining and Bayesian analysis more popular than ever. Things like growing volumes and varieties of available data, computational processing that is cheaper and more powerful, and affordable data storage.

All of these makes me possible to quickly and automatically produce models that can analyze bigger, more complex data and deliver faster, more accurate results- even on a very large scale. Building precise models, an organization has better chance of identifying profitable opportunities or avoiding unknown risks.

3.3 What is Required to Create Good Machine Learning Systems?

- Data preparation capabilities
- Algorithms – basic and advanced
- Automation and iterative processes
- Scalability
- Ensemble modeling

3.4 Who is Using Machine Learning?

Most industries working with large amounts of data have recognized the importance of machine learning technology. By gleaning insights from this data (often in real time) organizations are able to work more efficiently or gain advantage over competitors.

- **Financial services**

Banks and other businesses in the financial industry use machine learning technology for two key purposes: to identify important insights in data, and prevent fraud. The insights can identify investment opportunities, help investors know when to trade. Data mining can also identify clients with high risk profiles, use cyber surveillance to pinpoint warning signs of fraud.

- **Government**

Government agencies such as public safety and utilities have a particular need for machine learning since hay have multiple sources of data that can be mined for insights. Analyzing sensor data, for example, identifies ways to increase efficiency and save money. Machine learning can also help detect fraud and minimizing identity theft.

- **Health care**

Machine learning is a fast growing trend in the health care industry, wearable devices and sensors can use data asses a patient's health in real time. The technology can also help medical exerts analyze data to identify trends or red flags that may lead to improved diagnoses and treatment.

- **Oil and gas**

Finding new energy sources. Analyzing minerals in the ground. Predicting refinery sensor failure. Streamlining oil distribution to make it more efficient and cost effective. The number of machine learning use cases for this industry is vast and still expanding.

- **Marketing and Sales**

Websites recommending items you might like based on previous purchases are using machine leavening to analyze your buying history – and promote other items you'd be interested in. This ability to capture data, analyze it and use it to personize a shopping experience or implementing a marketing campaign is the future of retail.

- **Transportation**

Analyzing data to identify patterns and trends is key to the transportation industry, which relies on making routes more efficient and predicting potential problems to increase profitability. The data analysis and modeling aspects of machine learning are important tools to delivery companies, public transportation and other transportation organizations.

4. PYTHON'S PYBRAIN

PyBrain is a machine learning library written in Python. PyBrain is implemented in Python, with the scientific library SciPy. Pybrain is designed to be able to connect various types of architectures and algorithms. PyBrain provides a toolbox for supervised, unsupervised and reinforcement learning as well as black box and multi objective optimization.

The library includes different types of training algorithms, specialized data sets, trainable architectural components etc. Appropriate data handling tools have been developed for special applications reinforcement learning, handwriting recognition applications.

4.1 What's Unique in Pybrain

PyBrain is short forPython-Based Reinforcement Learning, Artificial Intelligence and Neural Network Library. Amongst few machine learning libraries available, PyBrain aims to be very easy-to-use modular library that offers flexibility and algorithms for research.

PyBrain contains algorithms for neural networks, for reinforcement learning (and the combination of the two), for unsupervised learning and evolution. The library is used around neural networks in the kernel and all of the training methods accept a neural network as the to-be-trained instance. This makes PyBrain powerful tool for real-life tasks.

4.2 Using PyBrain

PyBrain is open source and free to use for everyone (it is licensed under the BSD Software License). We can download it and start using the algorithms and modules in our project.

5. USER FRIENDLY DATA STRUCTURES

Python has built-in list and dictionary data structures which can be used to construct fast runtime data structures. Python also provides the option of dynamic high-level data typing that reduces the length of support code that is needed.

5.1 Python Pandas

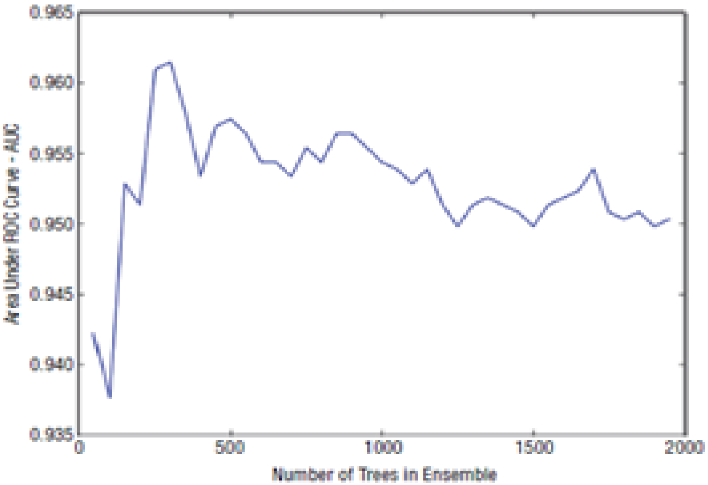
The python package Pandas can help automate the process of data inspection and handling. It proves particularly useful for the early stages of data inspection and preprocessing.

The Pandas package makes it possible to read data into a specialized data structure called **data frame.** The data frame is modeled after the CRAN-R data structure of the same name.

The Pandas can be difficult to install as it has number of dependencies that need to be correctly versioned. The installation procedures are easy to follow and result in compatible installations of wide variety of packages for data analysis and **machine learning.**

5.2 Python's Ensemble Package

The Python scikit-learn ensemble module houses a Radom Forest algorithm and a Gradient Boosting algorithm for **regression problem.** *RandomForestRegressor*has several attributes, including the trained trees that makeup the ensemble. The *predict* method will use the trained trees to make predictions, so you will not generally access those directly.



6. CONCLUSION

In this paper the recent trends of research on Python's rich set of libraries used in machine learning was reviewed. The focus was on the role of Python as an upcoming machine learning language preferred by the developers. In this paper we discussed the importance, new trends and the vital role of machine learning in real life applications. This paper reviews the Python libraries like PyBrain, Python Pandas and Python's Ensemble Package that are used for machine learning project. We have also discussed the simplicity and flexibility of Python that makes it programmer's first choice. Though the research on Machine learning languages shows R as the first preferred language over Python, Python is picking up the attention of the programmers. This research paper on the Python's role as a programming language in machine learning will help serve as a reference for researchers and machine learning programmers and also be an introduction for those who are less familiar with the subject.

REFERENCES

[1] J.S. Bridle, “Probabilistic Interpretation of Feed forward Classification Network Outputs, with Relationships to Statistical Pattern Recognition,” *Neurocomputing—Algorithms, Architectures and Applications,* F. Fogelman-Soulie and J. Herault, eds., NATO ASI Series F68, Berlin: Springer-Verlag, pp. 227-236, 1989. (Book style with paper title and editor)

[2] Mrs. Smita Desai, Miss. Shreya Desai “Smart Vehicle Automation”,IJCSMC, Vol. 6,Issue. 9, September2017, pg.46 – 50

[3] W.-K. Chen, *Linear Networks and Systems.* Belmont, Calif.: Wadsworth, pp. 123-135, 1993. (Book style)

[4] D.S. Coming and O.G. Staadt, "Velocity-Aligned Discrete Oriented Polytopes for Dynamic Collision Detection," *IEEE Trans. Visualization and Computer Graphics*, vol. 14, no. 1, pp. 1-12, Jan/Feb 2008, doi:10.1109/TVCG.2007.70405. (IEEE Transactions)

[5] S.P. Bingulac, “On the Compatibility of Adaptive Controllers,” *Proc. Fourth Ann. Allerton Conf. Circuits and Systems Theory*, pp. 8-16, 1994. (Conference proceedings)

[6] H. Goto, Y. Hasegawa, and M. Tanaka, “Efficient Scheduling Focusing on the Duality of MPL Representation,” *Proc. IEEE Symp. Computational Intelligence in Scheduling(SCIS ’07)*, pp. 57-64, Apr. 2007, doi:10.1109/SCIS.2007.367670. (Conference proceedings)

[7] J. Williams, “Narrow-Band Analyzer,” PhD dissertation, Dept. of Electrical Eng., Harvard Univ., Cambridge, Mass., 1993. (Thesis or dissertation)

[8] J.M.P. Martinez, R.B. Llavori, M.J.A. Cabo, and T.B. Pedersen, "Integrating Data Warehouses with Web Data: A Survey," *IEEE Trans. Knowledge and Data Eng.*, preprint, 21 Dec. 2007, doi:10.1109/TKDE.2007.190746.(PrePrint)