anadrol

053.314.328-40 Identificação

custom\_dictionaries\_only

custom\_terms\_only

Jane Smith

5446882375129691

Exp. 09/16

John Smith

4532014495343

Exp. 02/17A question I get asked a lot on my email list is: what is the best programming language for machine learning?

I’ve replied to this question many times now it’s about time to explore this further in a blog post. Ultimately, the programming language you use for machine learning should consider your own requirements and predilections. No one can meaningfully address those concerns for you.

What Languages Are Being Used

Before I give you my opinion, it is good to have  a look around to see what languages and platforms are popular in self-selected communities of data analysis and machine learning professionals.

KDnuggets has had language polls forever. A recent poll is titled “What programming/statistics languages you used for an analytics / data mining / data science work in 2013“. The trends are almost identical to the previous year. The results suggest heavy use of R and Python and SQL for data access. SAS and MATLAB rank higher than I would have expected. I’d expect SAS accounts for larger corporate (Fortune 500) data analysis and MATLAB for engineering, research and student use.

Kaggle offer machine learning competitions and have polled their user base as to the tools and programming languages used by participants in competitions. They posted results in 2011 titledKagglers’ Favorite Tools (also see the forum discussion). The results suggested the abundant use of R. The results also show good use of MATLAB and SAS with much lower Python representation. I can attest that I prefer R over Python for competition work. It just feels though it has more on offer in terms of data analysis and algorithm selection.

Ben Hamner, Kaggle Admin and author of the blog post above on the Kaggle blog goes into more detail on the options when it comes to programming languages for machine learning in a forum post titled “What tools do people generally use to solve problems“.

Ben comments that MATLAB/Octave is a good language for matrix operations and can be good when working with a well defined feature matrix. Python is fragmented by comprehensive and can be very slow unless you drop into C. He prefers Python when not working with a well defined feature matrix and uses Pandas and NLTK. Ben comments that “As a general rule, if it’s found to be interesting for statisticians, it’s been implemented in R” (well said). He also complains about the language itself being ugly and painful to work with. Finally, Ben comments on Julia that doesn’t have much to offer in the way of libraries but is his new favorite language. He comments that it has the conciseness of languages like MATLAB and Python with the speed of C.

Anthony Goldbloom, the CEO of Kaggle gave a presentation to the Bay Area R user group in 2011 on the popularity of R in Kaggle competitions titled Predictive modeling competitions: making data science a sport (see the powerpoint slides). The presentation slides give more detail on the use of programming languages and suggest an Other category that is as close to as large as large as the usage of R. It would be nice to have the raw data that was collected (why didn’t they release it to their own data community, seriously!?).

John Langford on his blog Hunch has an excellent article on the properties of a programming language to consider when working with machine learning algorithms titled “Programming Languages for Machine Learning Implementations“. He divides the properties into concerns of speed and the concerns of programability (programming ease). He points to powerful industry standard implementations of algorithms, all in C and comments that he has not used R or MATLAB (the post was written 8 years ago). Take some time and read some of the comments by academics and industry specialists alike. This is a deep and nuanced problem that really comes down to the specifics of the problem you are solving and the environment in which you are solving it.

Machine Learning Languages

I think of programming languages in the context of the machine learning activities I want to perform.

MATLAB/Octave

I think MATLAB is excellent for representing and working with matrices. As such, I think it’s an excellent language or platform to use when climbing into the linear algebra of a given method. I think it’s suited to learning about algorithms both superficially the first time around and deeply when you are trying to figure something out or go deep into the method. For example, it’s popular in university courses for beginners, like Andrew Ng’s Coursera Machine Learning course.

R

R is a workhorse for statistical analysis and by extension machine learning. Much talk is given to the learning curve, I didn’t really see the problem. It is the platform to use to understand and explore your data using statistical methods and graphs. It has a n enormous number of machine learning algorithms, and advanced implementations too written by the developers of the algorithm.

I think you can explore, model and prototype with R. I think it suits one-off projects with an artefact like a set of predictions, report or research paper. For example, it is the most popular platform for machine learning competitors such as Kaggle.

Python

Python if a popular scientific language and a rising star for machine learning. I’d be surprised if it can take the data analysis mantle from R, but matrix handling in NumPy may challenge MATLAB and communication tools like IPython are very attractive and a step into the future of reproducibility.

I think the SciPy stack for machine learning and data analysis can be used for one-off projects (like papers), and frameworks like scikit-learn may be mature enough to be used in production systems.

Java-family/C-family

Implementing a system that uses machine learning is an engineering challenge like any other. You need good design and developed requirements. Machine learning is algorithms, not magic. When it comes to serious production implementations, you need a robust library or you customize an implementation of the algorithm for your needs.

There are robust libraries, for example Java has Weka and Mahout. Also note that the deeper implementations of core algorithms like regression (LIBLINEAR) and SVM (LIBSVM) are written in C and leveraged by Python and other toolkits. I think you are serious you may prototype in R or Python, but you will implement in a heavier language for reasons such as execution speed and system reliability. For example, the backend of BigML is implemented in Clojure.

Other Concerns

Not a Programmer: If you are not a programmer (or not a confident programmer) I recommend playing machine learning via a GUI interface like Weka.

One Language for Research and Ops: You may want to use the same language for prototyping and for production to reduce risk of not effectively transferring the results.

Pet Language: You may have a pet language of favorite language and want to stick to that. You can implement algorithms yourself or leverage libraries. Most languages have some form of machine learning package, however primitive.

Form 1099-R

OMB No. 1545-0115

2016

PAYER’S name, street address, city or town, state or province, country, ZIP

or foreign postal code, and telephone no :

Z Builders

934 Cobblestone Court, Kingston, New York 12401

PAYER’S federal identification number : 846-6220742

RECIPIENT’S identification number : 846-6221234

RECIPIENT’S name : ALex Smith

Street address (including apt. no.) : 10th Floor, 100 South Main Street

City or town, state or province, country, and ZIP or foreign postal code :

Los Angeles, California 90012

Account number (see instructions) : 1284930494

FATCA filing 2nd TIN not. 13 Excess golden parachute

requirement

1 Rents : $ 1200.00

2 Royalties : $ 500.00

3 Other income : $ 1000.00

4 Federal income tax withheld : 230.00

5 Fishing boat proceeds : 900.00

6 Medical and health care payments : $ 560.00

7 Nonemployee compensation

8 Substitute payments in lieu of dividends or interest : $ 100.00

9 Payer made direct sales of $5,000 or more of consumer products to a buyer (recipient) for resale

10 Crop insurance proceeds : $ 100.00

13 Excess golden parachute payments $ 100.00

14 Gross proceeds paid to an attorney $ 600

15a Section 409A deferrals

15b Section 409A income

15b Section 409A income

16 State tax withheld

17 State/Payer’s state no. PA 3333333

18 State Income 4,567.00

Form 1099-MISC www.irs.gov/form1099misc Department of the Treasury - Internal Revenue Service