УДК 004.042

V. A. Kovenko
I. V. Bogach
L. V. Ibrahimova

OPEN-SOURCE DATA SCIENCE AND MACHINE LEARNING COURSE WITH PYTHON

Vinnytsia National Technical University
Вінницький національний технічний університет

Анотація

У цій статті представлений новий курс з відкритим кодом для тих, хто хоче почати вивчати науку про дані та машинне навчання. Тут також висвітлено основні вимоги до знань, охоплення тем та натхнення для створення.

Ключові слова: наука про дані, машинне навчання, мова програмування Python, бібліотека машинного навчання *Scikit-learn*, програмна бібліотека Pandas, бібліотека Matplotlib, Numpy, розширення мови *Python*, кероване навчання, спонтанне навчання.

Abstract*: A new open-source course for those who want to start learning Data Science and Machine Learning is introduced in this article. The basic knowledge requirements, themes coverage, and inspiration behind creating are highlighted.*

Keywords: Data Science, Machine Learning, Python, Scikit-learn, Pandas, Matplotlib, Numpy, supervised learning, unsupervised learning.

Introduction

Nowadays, everyone can observe the trend of overall automation. With the developing of computer machines the level of computational power available for people is increasing, thus, giving a lift to automation through intellectual systems. This trend accompanies the other trend of people who want to be involved into AI sphere. Unfortunately, not many of them success in it. While trying to dig into some sphere one can encounter lots of obstacles, especially when this sphere is so broad and diverse like AI. Andrew Ng has already proved that the education in the form of online courses is fascinating and encouraging, as it helps to cover learners all over the world and still each of them will feel like he/she is taught individually. Nevertheless, even now when many internet resources regarding Machine Learning and Data Science exist, it's still an issue to understand from what point to start and what knowledge is required. The other issues are: the complexity of material presentation, which sometimes scares beginners; price of the course of high quality; etc. The aim of this article is to present a new course regarding Data Science and Machine Learning that covers basic concepts of it, is free, easy to extend, and is assumed to solve the emphasized problems.

Problematics

There are main problems that had an impact on creation of this course:

- A lack of basic background knowledge in statistics and Data Science of people who start learning Machine Learning.
- A lack of a toolbox for digging into AI and ineptitude of its usage.
- A complexity of understanding from what side to come into the field.
- Inappropriate price for online courses and a hard procedure of its enhancement.

Basically, lots of people start learning ML from the perspective of Neural Networks even not knowing about the need of data processing, statistical validation and other ML algorithms, and that is not correct and ineffective. A free course made concerning Machine Learning in Matlab [1] has already been constructed by Andrew Ng and helped to encourage people to use online courses as the main educational tool. The only downside of it was in the fact that the main tool used in it was Matlab, which is not very up to date. All the considered facts showed the need for open-source Data Science and Machine Learning basic course using Python.

Course creation

The course covers such basic themes in Data Science as:

- Exploratory data analysis.
- Understanding data through visualizations.
- Cleaning and getting data.
- Hands on Pandas [2], Numpy [3], Matplotlib [4], etc.

And the following themes in Machine Learning:

- Regression models.
- Classification models.
- Techniques of models' evaluation.
- Unsupervised learning techniques.
- Hands on introduction to Scikit-learn [5].

It was decided to pick exactly those ones, as they tend to be a "should know" data for any AI practitioner. The course was created using a famous platform for online wiki named readthedocs, which makes it easier to extend documentation related to the course. The assignments for it were done using Jupyter Notebook [6], which is a great tool for data analysis and ML. The assignments are accessible online using a Google Colaboratory [7], which frees one from a need of running everything on his or her machine and gives an opportunity to work on the same task independently in playground mode, meaning that your changes to assignment will be visible only to you. The only thing required for learning is connection to the internet. If a student wants to run everything on his/her local machine, the course contains an instruction of how to achieve it. As the course is an open-source one, anybody can contribute to it to make it better. There is also a possibility to discuss a course on github issue of it. What is more, the course itself is a "hub" in terms of linking students to more advanced courses, the creation was inspired by the mentioned ones. This complex of advantages makes the course a starting point for anyone who has a will to dig into AI.

Summary

In our fast-running world it's beneficial to learn new things fast. Online education has opened lots of possibilities for students all over the world. However, it still remains a challenge to find good materials that cover the basics of this topic. The created course tends to solve issues of learners trying to enter the world of AI. By being an open-source one, this course is easy to be extended and enhanced. What is more, it's online structure and non-dependence on a type of a machine helps to cover more students. To sum up, the mentioned course is a started point for any learner interested in AI and should help to choose the way in this field.

REFERENCES

1. Matlab for Machine Learning [Electronic resource]. URL : https://www.mathworks.com/solutions/machine-learning.html
2. Exploratory Data Analysis with pandas [Electronic resource]. URL: https://towardsdatascience.com/exploratory-data-analysis-with-pandas-508a5e8a5964 - Title from the screen.
3. Python Numpy tutorial [Electronic resource]. URL: http://cs231n.github.io/python-numpy-tutorial/ - Title from the screen.
4. Data Visualization using Matplotlib [Electronic resource]. URL : https://towardsdatascience.com/data-visualization-using-matplotlib-16f1aae5ce70- Title from the screen.
5. Hands-On Introduction To Scikit-learn [Electronic resource]. URL : https://towardsdatascience.com/hands-on-introduction-to-scikit-learn-sklearn-f3df652ff8f2 - Title from the screen.
6. Jupyter Notebook: An Introduction [Electronic resource]. URL : https://realpython.com/jupyter-notebook-introduction/ - Title from the screen.
7. Introduction To Google Colab [Electronic resource]. URL : https://medium.com/@animaze97/introduction-to-google-colab-9b2e28fe691a - Title from the screen.

Ковенко Володимир Андрійович — студент групи 1ІСТ-18б, кафедра автоматизації та інтелектуальних інформаційних технологій, Факультет комп'ютерних систем і автоматики, Вінницький національний технічний університет, м.Вінниця, e-mail: urumipainblackreaper@gmail.com

Богач Ілона Віталіївна — кандидат технічних наук, доцент кафедри автоматизації та інтелектуальних інформаційних технологій, Вінницький національний технічний університет, м.Вінниця.

Ібрагімова Людмила Володимирівна — старший викладач кафедри іноземних мов, Вінницький національний технічний університет, м. Вінниця, e-mail: milatvin@ukr.net

Kovenko Volodymyr A. — Faculty of Computer Systems and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: urumipainblackreaper@gmail.com

Bogach Ilona V. — PhD, Associate Professor of the Department of Automation and Intelligent Information Technologies, Vinnytsia National Technical University, Vinnytsia.

Ibrahimova Liudmyla V. — Senior Lecture, Chair of Foreign Languages, Vinnytsia National Technical University, Vinnytsia, e-mail: milatvin@ukr.net