Программа учебной дисциплины

«Введение в машинное обучение на языке Python»

Утверждена

Академическим советом ООП

Протокол № ___ от ___ августа 2018 г.

Автор	Горбунов А.А., преп.
Число кредитов	4
Контактная работа (час.)	52
Самостоятельная работа (час.)	52
Курс	3
Формат изучения дисциплины	Без использования онлайн курса

1. Course description

When solving various business tasks, it has to deal with the need to process large amounts of information. To work with” Big data " it needs to own a variety of technologies that allows to use machine learning algorithms. Python language contains a number of built-in libraries for working with data, developing machine learning algorithms, and is supported by many modern platforms Apache Spark, Microsoft Azure, etc.

The objectives of the course is to develop students ' complex theoretical knowledge and methodological foundations in the field of machine learning, as well as practical skills for working with big data using Python.

To achieve these goals, it needs to solve the following tasks:

- to learn the basic terms and concepts of the Python language;
- to learn basic machine learning terms and concepts;
- to learn modern software platforms that support the implementation of machine learning algorithms;
- to learn practical examples of machine learning algorithms implementation in Python

The peculiarity of the discipline is a combination of theory and practice, the use of modern software platforms for the study of machine learning.

2. Learning outcomes

On successful completion of the course students are expected

To know:

- basic terms and concepts of Python language;
- basic terms and concepts of machine learning;

To have practical skills of:

- using built-in Python libraries;
- developing machine learning algorithms in Python;

- solving various business tasks for processing large amounts of information.

To acquire basic knowledge of:

- Tools and modern software platforms that support the implementation of machine learning algorithms;

For successful completion of this course, students should have a basic knowledge of probability theory, mathematical statistics, information technology.

The knowledge, skills gained in the course of studying can be used in solving various business problems in the processing and analysis of large amounts of information in the areas of design and implementation of intelligent systems and services.

3. Topic-wise course plan

№	Topic	Total hours	In-class		Self-study
			Lectures	Practice	
1	Introduction . Python basics	16	4	4	8
2	Statistics and Probability Refresher, and Python Practise	16	4	4	8
3	Predictive Models	16	4	4	8
4	Machine Learning with Python	16	4	4	8
5	Recommender Systems	16	4	4	8
6	Dealing with Real-World Data	8	-	4	4
7	Apache Spark Machine Learning on Big Data	16	4	4	8
	Total	**104**	**24**	**28**	**52**

4. Requirements and grading

Type of control	Form of control	Module s		Characteristics
		1	2	
Current control	Homework assignment	X		Short coding problems split throughout the course
	Test		X	Test on theoretical aspects of the course with multiple choice and open response questions
Final control	Exam		X	One or two coding problems

Final grading

The teacher evaluates the work of students in practical classes.

The current $O_{current}$ is calculated as the weighted sum of all current control forms.

$$O_{current} = 0.4\cdot{*}O_{hw} + 0.6\cdot{*}O_{test};$$

The cumulative rating O_{cum} is calculated by the formula:

$$\mathrm{O}_{cum} = 0.6{*}O_{current} + 0.4{*}\ O_{seminars}$$

Otherwise, if you perform all the substitutions, we get that the accumulated score is calculated by the formula

$$\mathrm{O}_{cum} = 0.24\cdot{*}\ O_{hw} + 0.36\cdot{*}\ O_{test} + 0.4{*}\ O_{seminars}$$

The resulting score for the discipline is calculated as follows:

$$O_{final} = 0.5{*}\ \mathrm{O}_{cum} + 0.5\ {*}\cdot O_{exam}$$

The method of rounding the accumulated assessment of the final control in the form of an exam– arithmetic.

The course does not provide for the exemption of the student from passing the exam, with the issuance of a grade on the interim certification corresponding to the accumulated assessment without taking into account the weight of the exam.

The last element of Discipline control or Exam (provided for in the current period and conducted during the session or within 10 calendar days before the session) is carried out in writing and has a blocking character. This is the only blocking element of control within the discipline.

5. Course content

Topic 1. Introduction. Python basics.

Enthought Canopy Express development environment. Basic concepts of the Python language. Run Python scripts.

Main readings:

Wes McKinney - Python and data analysis, DMK, 2015

Additional readings:

L. P. Coelho, V. Richard - building machine learning systems in Python, DMK, 2016 Multimedia presentation for lectures on the topic

***Topic 2* Statistics and Probability Refresher, and Python Practise**

Data type. Expectation, median, mode, standard deviation, variance. Distribution functions, probability density. Percentiles and moments. Covariance and correlation. Conditional probability. Bayes theorem.

Main readings:

Wes McKinney - Python and data analysis, DMK, 2015

Additional readings:

L. P. Coelho, V. Richard - building machine learning systems in Python, DMK, 2016 Multimedia presentation for lectures on the topic

Topic 3 Predictive Models

Regression Algorithms. Multilevel models.

Main readings:

Wes McKinney - Python and data analysis, DMK, 2015

Additional readings:

L. P. Coelho, V. Richard - building machine learning systems in Python, DMK, 2016 Multimedia presentation for lectures on the topic

***Topic 4* Machine Learning with Python**

Training with a teacher and without a teacher. Overfitting. Bayesian methods. Clustering. Entropy change. Decision tree. Ensemble learning. SVM. K-nearest neighbor method. Dimension reduction. Principal components analysis method.

Main readings:

Wes McKinney - Python and data analysis, DMK, 2015

Additional readings:

L. P. Coelho, V. Richard - building machine learning systems in Python, DMK, 2016 Multimedia presentation for lectures on the topic

Topic 5. Recommender Systems

User-Based Collaborative Filtering. Item-Based Collaborative Filtering

Main readings:

Wes McKinney - Python and data analysis, DMK, 2015

Additional readings:

L. P. Coelho, V. Richard - building machine learning systems in Python, DMK, 2016

Multimedia presentation for lectures on the topic

Topic 6. Dealing with Real-World Data

Cross-validation for K blocks. Data cleaning and normalization. The detection of outliers.Main readings:

Main readings:

Wes McKinney - Python and data analysis, DMK, 2015

Additional readings:

L. P. Coelho, V. Richard - building machine learning systems in Python, DMK, 2016 Multimedia presentation for lectures on the topic

Topic 7. Apache Spark Machine Learning on Big Data

The Concept Of Apache Spark. RDD. Introduction to MLLib

Main readings:

Wes McKinney - Python and data analysis, DMK, 2015

Additional readings:

L. P. Coelho, V. Richard - building machine learning systems in Python, DMK, 2016 Multimedia presentation for lectures on the topic

6. Assignment topics

In the course of the classes, students perform practical classes on developing their own scenarios of data analysis using the built-in Python libraries.

7. Assignment topics

Examples of test questions

1. The basic design of the Python language.
2. Library of Python to work with forecast models.
3. Python libraries for working with machine learning algorithms 4. Algorithms of work with regression. 5. Multilevel model.
6. Machine learning algorithms
7. Data type.
8. Expectation, median, mode, standard deviation, variance. Distribution functions, probability density. Percentiles and moments. Covariance and correlation.
9. Conditional probability. bayes theorem.
10. Training with a teacher and without a teacher.

11. Retraining.
12. Bayesian methods.
13. Clustering.
14. Decision tree.
15. Ensemble learning.
16. SVM.
17. K-nearest neighbor method.
18. Dimension reduction. Principal components analysis method.
19. User-Based Collaborative Filtering.
20. Item-Based Collaborative Filteringstandard
21. The Concept Of Apache Spark.
22. What is RDD?
23. What is MLLib?
24. Cross-validation for K blocks.
25. Data cleaning and normalization.
26. The detection of outliers.

7. Technical equipment

Lectures require an internet-connected PC and a projector.