Національний технічний університет України
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

Emblem departments (if available)

Department of computer technology

STATISTICAL METHODS OF MACHINE LEARNING

Working program of the academic discipline (Syllabus)

Details of the academic discipline

Level of higher education	***First (bachelor's degree)***
Branch of knowledge	*12 Information Technology*
Specialty	*123 Computer Engineering*
Educational program	*Computer systems and networks*
Discipline status	*Selective*
Form of education	*Daytime*
Year of training, semester	*4th year, spring semester*
Scope of the discipline	*4 credits, 120 hours*
Semester control/ control measures	*Test*
Lessons schedule	*Lectures - 36 hours. Laboratory work - 36 hours. Independent work of students - 48 hours*
Language of teaching	*Ukrainian*
Information about head of the course / teachers	Lecturer: Ph.D, professor, Mykhailo Anatoliyovych Novotarskyi novotar@gmail.com, http://novotarsky.pp.ua Laboratory: Ph.D, professor, Mykhailo Anatoliyovych Novotarskyi novotar@gmail.com, http://novotarsky.pp.ua
Placement of the course	https://ecampus.kpi.ua

Program of educational discipline

1. Description of the academic discipline, its purpose, subject of study and learning outcomes

The syllabus of the discipline "Statistical methods of machine learning" is compiled in accordance with the educational and professional training program for bachelors of specialty 123 - "Computer engineering".

2. The educational discipline "Statistical methods of machine learning" is an elective discipline.

The program of the educational discipline "Statistical methods of machine learning" establishes the requirements for the student's knowledge and skills and determines the content and types of training sessions.

The study of the educational discipline "Statistical methods of machine learning" is based on the following disciplines:

- Mathematical analysis
- Linear algebra and analytic geometry
- Probability theory and mathematical statistics

To master the academic discipline, students must possess:

- knowledge of basic definitions and theorems of the above disciplines;
- skills in solving typical problems of these disciplines.

1.1. The purpose and tasks of the educational discipline

1.1. The purpose of the educational discipline

The goal of mastering the discipline "Statistical methods of machine learning" is to study the basic methods of machine learning for the tasks of classification, clustering and regression (forecasting). Within the framework of this course, students get an idea of the tasks that are solved with the help of this theory, and the principles of building some basic classifiers. When studying the discipline, significant attention is paid to algorithmic and computational aspects, so students additionally gain knowledge of the Python programming language and the THEANO library as tools for practicing practical skills.

1.2. Основні завдання навчальної дисципліни

According to the requirements of the educational and professional program, after mastering the academic discipline, students must demonstrate the following learning outcomes:

Knowledge of: principles of construction of feature vectors, decision rules and classification;

main types of classifiers; principles of building linear classifiers; principles of building nonlinear classifiers; features of the selection of features of classification and pre-processing of data.

Ability: to choose the appropriate type of classifier depending on the problem to be solved; choose a set of features for classification and carry out preliminary data processing; to be able to apply algorithms for building and training a classifier by sample; perform calculations related to the training and operation of the classifier in the Python environment using the THEANO library.

Experience: student must

have the skills to: choose, build, learn and use basic classifiers when solving tasks; independent work in modern software complexes; mastering a large amount of information; programming for solving data analysis problems; have a culture of setting tasks and conducting an experiment; to have visualization tools to demonstrate the obtained results.

2. The structure of the academic discipline

120 hours/4 ECTS credits are allocated to the study of the academic discipline.

The educational discipline includes 1 credit module: "Statistical methods of machine learning»

Distribution of study time

Form of education	Credit modules	In total		Distribution of study time by types of classes			Semester certification
		Credits	Hour	Lectures	Laboratory work	IWS	
Daytime	1	4	120	36	36	48	Test

3. Content of the academic discipline

Topic 1. Introductory provisions of the theory of machine learning

The concept of machine learning. Fields of application of machine learning. Types of training. Deductive and inductive learning. Connection with other areas of science. Symbolic learning. Modern implementations of symbolic learning.

Topic 2. Statement of the problem and basic concepts of statistical learning

Objects and signs. Types of tasks. Algorithm model and learning method. Examples of models. Functionality of quality. Minimization of empirical risk. The problem of retraining. Applied learning problems: classification problems, regression estimation problems, ranking problems; clustering tasks; association search tasks. Methods of testing learning algorithms. Techniques for generating model data.

Topic 3. Metric methods of classification

Nearest neighbor method and its generalization: generalized metric classifier, k-nearest neighbor method, Parsen window method, potential function method. Selection of reference objects: concept of object offset, STOLP algorithm for selection of reference objects.

Topic 4. Linear classification methods

An artificial neuron as a model of a nerve cell. Connectivism and neural networks. Approximation and regularization of empirical risk. Linear classification model. Stochastic gradient method: classic cases, heuristics for improving gradient learning methods.

Topic 5. Logistic regression. Method of support vectors

Justification of logistic regression. Stochastic gradient method for logistic regression: principle of maximum likelihood, gradient step, analogy with Hebb's rule, advantages and disadvantages of logistic regression. Scoring and evaluation of a priori probabilities, probabilistic output and risk assessment, probabilistic calibration. Method of support vectors. Sampling with linear resolution: optimal separating hyperplane, normalization, width of the separating band. Linearly non-separate sampling: regularization of empirical risk, a dual problem. Cores and straightening spaces: constructive ways of building cores. Two-layer neural network for implementation of the support vector method. Advantages and disadvantages of the method of support vectors.

Topic 6. Bayesian theory of classification. Substantive provisions. Probabilistic formulation of the problem. Non-parametric classification.

Basic provisions of probability theory, necessary for use within the course: probability distribution function of a discrete random variable and its properties, continuous random variables, conditional probability, complete probability, Bayes' formula. Probabilistic formulation of the classification problem. Medium risk functional. Maximization and minimization arguments. Optimal Bayes classifier. "Naive" Bayesian classifier. Nonparametric classification: one-dimensional continuous case, multidimensional continuous case. The Parsen window method.

Topic 7. Bayesian theory of classification. Normal discriminant analysis. Separation of a mixture of distributions.

Multivariate normal distribution. Quadratic discriminant. Fisher's linear discriminant. EM algorithm. Mixtures of multivariate normal distributions.

Topic 8. Regression recovery methods (regression estimates)

The method of least squares. Nonparametric regression: kernel smoothing, Nadaray-Watson formula, selection of kernel and window width, outliers problem (robust nonparametric regression), LOWESS algorithm (locally weighted smoothing), edge effects problem. Linear regression: singular decomposition, multicollinearity problem, ridge regression. Nonlinear regression recovery methods: nonlinear regression model, nonlinear one-dimensional feature transformations.

Topic 9. Logical classification algorithms

Problem formulation for logical classification methods. Statistical definition of informativeness. Entropy definition of informativeness. Multi-class informativeness Weighted informativeness. Methods for searching for informative regularities: binarization of quantitative features, "greedy" zone merging algorithm, search for regularities in the form of conjunctions, "gradient" conjunction synthesis algorithm, "greedy" conjunction synthesis algorithm, stochastic local search, emergent conjunction synthesis algorithm junctions, forms of regularities.

Topic 10. Lists and decision trees

Decision lists: a greedy decision list construction algorithm, examples of decision lists. Decision-making trees: synthesis of decision-making trees, ID3 decision-making tree construction algorithm, gap processing, probability estimation, complexity of the ID3 algorithm, advantages and disadvantages of the ID3 algorithm, reduction of decision-making trees. Converting a decision tree to a decision list. Looking ahead.

Topic 11. Weighted voting of rules

Voting principles: simple voting algorithm, weighted voting algorithm, adjustment of weights, diversification of rules, refusal of classification. The KORA algorithm: principles of construction of various modifications of the KORA algorithm, advantages and disadvantages of the KORA algorithm. The TEMP algorithm: principles of the algorithm, advantages and disadvantages of the TEMP algorithm. Boosting algorithm: exponential approximation of the marginal loss function, principles of the boosting algorithm, advantages and disadvantages of the boosting algorithm.

Topic 12. Artificial neural networks

The completeness problem and the "XOR" problem. Computing capabilities of neural networks. Multilayer neural networks and the method of error backpropagation. Advantages and disadvantages of the error backpropagation method. Heuristics for improving convergence: choice of initial approximation, choice of gradient optimization method. Optimization of the structure of the neural network: selection of the number of layers, selection of the number of neurons in the hidden layer, dynamic addition of neurons, removal of redundant connections.

Topic 13. Clustering. Clustering algorithms

Objectives of clustering. Types of cluster structures. Heuristic graph algorithms: the algorithm for selecting connected components, the shortest non-closed path algorithm, the FOREL algorithm. Clustering quality functionals. Statistical algorithms: hypothesis about the space of objects and the shape of clusters, the method of averages, clustering with partial learning. Hierarchical clustering: monotonicity property, Milligan's theorem, properties of stretching and compression, reducibility property, Didet and Moreau's theorem, determination of the number of clusters, advantages and disadvantages of clustering.

Topic 14. Kohonen networks

Models of competitive learning: the rule of hard competition WTA, the rule of fair competition CWTA, the rule of soft competition WTM. Kohonen self-organizing maps, the art of interpreting Kohonen maps. Disadvantages of Kohonen cards. Hybrid backpropagation networks: piecewise-continuous approximation, smooth approximation. Multidimensional scaling: placement of one object by Newton-Raphson method, sub-quadratic multidimensional scaling algorithm, similarity map, Shepard diagram.

Educational materials and resources

Basic:

1. Novotarskyi M.A. Lectures on the course "Statistical methods of machine learning" // https://cloud.comsys.kpi.ua/s/pjw4nAStZFPsDos

2. Novotarskyi M.A. Methodical instructions for performing laboratory work from the course "Fundamentals of science sold" // Novotarskyi M.A. Lectures from the course "Fundamentals of science sold" //
3. Asuncion A., Newman D. UCI machine learning repository: Tech. rep.: University of California, Irvine, School of Information and Computer Sciences, 2007. http://www.ics.uci.edu/^mlearn/MLRepository.html.
4. Bishop C. M. Pattern Recognition and Machine Learning. — Springer, Series: Information Science and Statistics, 2006. — 740 pp.
5. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning.— Springer, 2001. – 533 pp. http://http://www-stat.stanford.edu/^tibs/ElemStatLearn.
6. Jordan M. I., Xu L. Convergence results for the EM algorithm to mixtures of experts architectures: Tech. Rep. A.I. Memo No. 1458: MIT, Cambridge, MA, 1993.
7. LeCun Y., Bottou L., Orr G. B., Muller K.-R. Efficient BackProp // Neural Networks: tricks of the trade. – Springer, 1998.
8. Smola A., Schoelkopf B. A tutorial on support vector regression: Tech. Rep. NeuroCOLT2 NC2-TR-1998-030: Royal Holloway College, London, UK, 1998. http://citeseer.ist.psu.edu/smola98tutorial.html.
9. Новотарський М.А., Нестеренко Б.Б. Штучні нейронні мережі: обчислення. К: Ін-т математики НАН України, 2004.–408 с.

Additional:
1. J. Kelleher. Data science: Basic course / J. Kelleher, B. Tyrny, Publisher, 2020. – 157 p.
2. Wang L., Fu X. Data Mining with Computational Intelligence. –Springer, 2005. –280 p.

Educational content

4. Methods of mastering an educational discipline (educational component)

The structure of the educational discipline "Statistical methods of machine learning" in table 1.

Table 1

The structure of the educational discipline "Statistical methods of machine learning"

Names of sections, topics	Number of hours			
	In total	У тому числі		
		Lectures	Laboratory work	IWS
Topic 1. Introductory provisions of the theory of machine learning The concept of machine learning. Fields of application of machine learning. Types of training. Deductive and inductive learning. Connection with other areas of science. Symbolic learning. Modern implementations of symbolic learning.	6	2	4	0
Topic 2. Statement of the problem and basic concepts of statistical learning Objects and signs. Types of tasks. Algorithm model and learning method. Examples of models. Functionality of quality. Minimization of empirical risk. The problem of retraining. Applied learning problems: classification problems,	3	2	0	1

Names of sections, topics	Number of hours			
	In total	У тому числі		
		Lectures	Laboratory work	IWS
regression estimation problems, ranking problems; clustering tasks; association search tasks. Methods of testing learning algorithms. Techniques for generating model data.				
Topic 3. Metric methods of classification Nearest neighbor method and its generalization: generalized metric classifier, k-nearest neighbor method, Parsen window method, potential function method. Selection of reference objects: concept of object offset, STOLP algorithm for selection of reference objects.	6	2	4	0
Topic 4. Linear classification methods An artificial neuron as a model of a nerve cell. Connectivism and neural networks. Approximation and regularization of empirical risk. Linear classification model. Stochastic gradient method: classic cases, heuristics for improving gradient learning methods.	6	2	4	
Topic 5. Logistic regression. Method of support vectors Justification of logistic regression. Stochastic gradient method for logistic regression: principle of maximum likelihood, gradient step, analogy with Hebb's rule, advantages and disadvantages of logistic regression. Scoring and evaluation of a priori probabilities, probabilistic output and risk assessment, probabilistic calibration. Method of support vectors. Sampling with linear resolution: optimal separating hyperplane, normalization, width of the separating band. Linearly non-separate sampling: regularization of empirical risk, a dual problem. Cores and straightening spaces: constructive ways of building cores. Two-layer neural network for implementation of the support vector method. Advantages and disadvantages of the method of support vectors.	8	4	4	0
Topic 6. Bayesian theory of classification. Substantive provisions. Probabilistic formulation of the problem. Non-parametric classification. Basic provisions of probability theory, necessary for use within the course: probability distribution function of a discrete random variable and its properties, continuous random variables, conditional probability, complete probability, Bayes' formula. Probabilistic formulation of the classification problem. Medium risk functional. Maximization and minimization arguments. Optimal Bayes classifier. "Naive" Bayesian classifier. Nonparametric classification: one-	6	2	4	0

Names of sections, topics	Number of hours			
	In total	У тому числі		
		Lectures	Laboratory work	IWS
dimensional continuous case, multidimensional continuous case. The Parsen window method.				
Topic 7. Bayesian theory of classification. Normal discriminant analysis. Separation of a mixture of distributions Multivariate normal distribution. Quadratic discriminant. Fisher's linear discriminant. EM algorithm. Mixtures of multivariate normal distributions.	10	2	8	0
Topic 8. Regression recovery methods (regression estimates) The method of least squares. Nonparametric regression: kernel smoothing, Nadaray-Watson formula, selection of kernel and window width, outliers problem (robust nonparametric regression), LOWESS algorithm (locally weighted smoothing), edge effects problem. Linear regression: singular decomposition, multicollinearity problem, ridge regression. Nonlinear regression recovery methods: nonlinear regression model, nonlinear one-dimensional feature transformations.	4	2	0	2
Topic 9. Logical classification algorithms Problem formulation for logical classification methods. Statistical definition of informativeness. Entropy definition of informativeness. Multi-class informativeness Weighted informativeness. Methods for searching for informative regularities: binarization of quantitative features, "greedy" zone merging algorithm, search for regularities in the form of conjunctions, "gradient" conjunction synthesis algorithm, "greedy" conjunction synthesis algorithm, stochastic local search, emergent conjunction synthesis algorithm junctions, forms of regularities.	3	2	0	1
Topic 10. Lists and decision trees Decision lists: a greedy decision list construction algorithm, examples of decision lists. Decision-making trees: synthesis of decision-making trees, ID3 decision-making tree construction algorithm, gap processing, probability estimation, complexity of the ID3 algorithm, advantages and disadvantages of the ID3 algorithm, reduction of decision-making trees. Converting a decision tree to a decision list. Looking ahead.	8	4	4	0
Topic 11. Weighted voting of rules	5	4	0	1

Names of sections, topics	Number of hours			
	In total	У тому числі		
		Lectures	Laboratory work	IWS
Voting principles: simple voting algorithm, weighted voting algorithm, adjustment of weights, diversification of rules, refusal of classification. The KORA algorithm: principles of construction of various modifications of the KORA algorithm, advantages and disadvantages of the KORA algorithm. The TEMP algorithm: principles of the algorithm, advantages and disadvantages of the TEMP algorithm. Boosting algorithm: exponential approximation of the marginal loss function, principles of the boosting algorithm, advantages and disadvantages of the boosting algorithm.				
Topic 12. Artificial neural networks The completeness problem and the "XOR" problem. Computing capabilities of neural networks. Multilayer neural networks and the method of error backpropagation. Advantages and disadvantages of the error backpropagation method. Heuristics for improving convergence: choice of initial approximation, choice of gradient optimization method. Optimization of the structure of the neural network: selection of the number of layers, selection of the number of neurons in the hidden layer, dynamic addition of neurons, removal of redundant connections.	4	2	0	2
Objectives of clustering. Types of cluster structures. Heuristic graph algorithms: the algorithm for selecting connected components, the shortest non-closed path algorithm, the FOREL algorithm. Clustering quality functionals. Statistical algorithms: hypothesis about the space of objects and the shape of clusters, the method of averages, clustering with partial learning. Hierarchical clustering: monotonicity property, Milligan's theorem, properties of stretching and compression, reducibility property, Didet and Moreau's theorem, determination of the number of clusters, advantages and disadvantages of clustering.	5	4	0	1

Names of sections, topics	Number of hours			
	Total	Including		
		Lectures	Laboratory work	IWS
Topic 14. Kohonen networks Models of competitive learning: the rule of hard competition WTA, the rule of fair competition CWTA, the rule of soft competition WTM. Kohonen self-organizing maps, the art of interpreting Kohonen maps. Disadvantages of Kohonen cards. Hybrid backpropagation networks: piecewise-continuous approximation, smooth approximation. Multidimensional scaling: placement of one object using the Newton-Raphson method, subquadratic multidimensional scaling algorithm, similarity map, Shepard diagram.	8	2	4	2
Control work	2		0	2
Exam	**36**		**0**	**36**
Total in semester:	**120**	**36**	**36**	**48**

The topics of the lectures, which are formed in accordance with the topics considered within the academic discipline, are listed in Table 2.

Table 2

№ In order	The name of the topic of the lecture and a list of main questions (a list of didactic tools, references to the literature and tasks on the IWS)
1	**Introduction to the theory of machine learning** The concept of machine learning. Fields of application of machine learning. Types of training. Deductive and inductive learning. Connection with other areas of science. Symbolic learning. Modern implementations of symbolic learning. **Tasks on SRS.** Tasks of learning by precedents.
2	**Statement of the problem and basic concepts of statistical learning** Objects and signs. Types of tasks. Algorithm model and learning method. Examples of models. Functionality of quality. Minimization of empirical risk. The problem of retraining. Applied learning problems: classification problems, regression estimation problems, ranking problems; clustering tasks; association search tasks. Methods of testing learning algorithms. Techniques for generating model data. **Tasks on SRS.** The concept of generalizing ability.
3	**Metric methods of classification** Nearest neighbor method and its generalization: generalized metric classifier, k-nearest neighbor method, Parsen window method, potential function method. Selection of reference objects: concept of object offset, STOLP algorithm for selection of reference objects. **Tasks on SRS.** Disadvantages of the simplest metric algorithms of the kNN type.
4	**Linear classification methods** An artificial neuron as a model of a nerve cell. Connectivism and neural networks. Approximation and regularization of empirical risk. Linear classification model.

№ In order	The name of the topic of the lecture and a list of main questions (a list of didactic tools, references to the literature and tasks on the IWS)
	Stochastic gradient method: classic cases, heuristics for improving gradient learning methods. **Tasks on SRS.** Independent parameters with unequal variances.
5	**Logistic regression** Justification of logistic regression. Stochastic gradient method for logistic regression: principle of maximum likelihood, gradient step, analogy with Hebb's rule, advantages and disadvantages of logistic regression. Scoring and evaluation of a priori probabilities, probabilistic output and risk assessment, probabilistic calibration. **Tasks on SRS.** Advantages and disadvantages of logistic regression.
6	**Method of support vectors.** Sampling with linear resolution: optimal separating hyperplane, normalization, width of the separating band. Linearly non-separate sampling: regularization of empirical risk, a dual problem. Cores and straightening spaces: constructive ways of building cores. Two-layer neural network for implementation of the support vector method. Advantages and disadvantages of the method of support vectors. **Tasks on SRS.** Method of relevant vectors (RVM).
7	**Bayesian theory of classification. Substantive provisions. Probabilistic formulation of the problem. Non-parametric classification.** Basic provisions of probability theory, necessary for use within the course: probability distribution function of a discrete random variable and its properties, continuous random variables, conditional probability, complete probability, Bayes' formula. Probabilistic formulation of the classification problem. Medium risk functional. Maximization and minimization arguments. Optimal Bayes classifier. "Naive" Bayesian classifier. Nonparametric classification: one-dimensional continuous case, multidimensional continuous case. The Parsen window method. **Tasks on SRS.** The principle of maximum posterior probability.
8	**Bayesian theory of classification. Normal discriminant analysis. Separation of a mixture of distributions.** Multivariate normal distribution. Quadratic discriminant. Fisher's linear discriminant. EM algorithm. Mixtures of multivariate normal distributions. **Tasks on SRS.** Mahalanobis distance.
9	**Regression recovery methods (regression estimates)** The method of least squares. Nonparametric regression: kernel smoothing, Nadaray-Watson formula, selection of kernel and window width, outliers problem (robust nonparametric regression), LOWESS algorithm (locally weighted smoothing), edge effects problem. Linear regression: singular decomposition, multicollinearity problem, ridge regression. Nonlinear regression recovery methods: nonlinear regression model, nonlinear one-dimensional feature transformations. **Tasks on SRS.** Lasso Tibshirani. Comparison of lasso and ridge regression.
10	**Логічні алгоритми класифікації** Постановка задачі для логічних методів класифікації. Статистичне визначення інформативності. Ентропійне визначення інформативності. Багатокласова інформативність Зважена інформативність. Методи пошуку інформативних закономірностей: бінаризація кількісних ознак, «жадібний» алгоритм злиття зон, пошук закономірностей у формі кон'юнкцій, «градієнтний» алгоритм синтезу кон'юнкції, «жадібний» алгоритм синтезу кон'юнкції, стохастичний локальний пошук, емерджентний алгоритм синтезу кон'юнкцій, форми закономірностей. **Завдання на СРС**. Співставлення двох критеріїв інформативності.
11	Logical classification algorithms

№ In order	The name of the topic of the lecture and a list of main questions (a list of didactic tools, references to the literature and tasks on the IWS)
	Problem formulation for logical classification methods. Statistical definition of informativeness. Entropy definition of informativeness. Multi-class informativeness Weighted informativeness. Methods for searching for informative regularities: binarization of quantitative features, "greedy" zone merging algorithm, search for regularities in the form of conjunctions, "gradient" conjunction synthesis algorithm, "greedy" conjunction synthesis algorithm, stochastic local search, emergent conjunction synthesis algorithm junctions, forms of regularities. **Tasks on SRS.** Comparison of two informativeness criteria.
12	**Lists and decision trees (continued)** Algorithm for building the ID3 decision-making tree, processing gaps, estimating probabilities, complexity of the ID3 algorithm, advantages and disadvantages of the ID3 algorithm, reduction of decision-making trees. Converting a decision tree to a decision list. Looking ahead. **Tasks on SRS.** Labor intensity of the task of building a decision-making tree in general.
13	Weighted Rules Voting Voting principles: simple voting algorithm, weighted voting algorithm, adjustment of weights, diversification of rules, refusal of classification. The KORA algorithm: principles of construction of various modifications of the KORA algorithm, advantages and disadvantages of the KORA algorithm. **Tasks on SRS.** Details of the implementation of the KORA algorithm.
14	Weighted Rules Voting The TEMP algorithm: principles of the algorithm, advantages and disadvantages of the TEMP algorithm. Boosting algorithm: exponential approximation of the marginal loss function, principles of the boosting algorithm, advantages and disadvantages of the boosting algorithm. **Tasks on SRS.** Theorem on the convergence of the boosting algorithm. different types of experimental data. - Novosibirsk: Nauka, 1981.
15	Artificial neural networks The completeness problem and the "XOR" problem. Computing capabilities of neural networks. Multilayer neural networks and the method of error backpropagation. Advantages and disadvantages of the error backpropagation method. Heuristics for improving convergence: choice of initial approximation, choice of gradient optimization method. Optimization of the structure of the neural network: selection of the number of layers, selection of the number of neurons in the hidden layer, dynamic addition of neurons, removal of redundant connections. **Tasks on SRS.** Kolmogorov's theorem on the representation of any continuous function of n arguments by a superposition of functions of one argument.
16	Clustering. Clustering algorithms Objectives of clustering. Types of cluster structures. Heuristic graph algorithms: the algorithm for selecting connected components, the shortest non-closed path algorithm, the FOREL algorithm. Clustering quality functionals. **Tasks on SRS.** Disadvantages of the algorithm for selecting connected components.
17	Clustering. Clustering algorithms (continued)

№ In order	The name of the topic of the lecture and a list of main questions (a list of didactic tools, references to the literature and tasks on the IWS)
	Statistical algorithms: hypothesis about the space of objects and the shape of clusters, the method of averages, clustering with partial learning. Hierarchical clustering: monotonicity property, Milligan's theorem, properties of stretching and compression, reducibility property, Didet and Moreau's theorem, determination of the number of clusters, advantages and disadvantages of clustering. **Tasks on SRS.** The concept of chain effect.
18	Kohonen networks Models of competitive learning: the rule of hard competition WTA, the rule of fair competition CWTA, the rule of soft competition WTM. Kohonen self-organizing maps, the art of interpreting Kohonen maps. Disadvantages of Kohonen cards. Hybrid backpropagation networks: piecewise-continuous approximation, smooth approximation. Multidimensional scaling: placement of one object using the Newton-Raphson method, subquadratic multidimensional scaling algorithm, similarity map, Shepard diagram. **Tasks on SRS**. An ideal Shepard chart.

1. Independent work of the student

A student's independent work consists of a theoretical and a practical component. The theoretical component involves the study of additional material that deepens the knowledge obtained at the lecture. The practical component of the student's independent work consists in performing laboratory work, the list of which is given in Table 3.

The main task of the cycle of laboratory classes is for students to acquire the necessary practical skills in developing algorithms and software for solving data processing problems on a PC.

The purpose of conducting a cycle of laboratory work is for students to acquire the necessary practical skills in the use of methods.

Laboratory work includes creating a block diagram of the algorithm, developing a program in the Python scripting language using the Theano library, printing out the results of the program execution, and analyzing the results.

The basic environment for developing programs is the development environment for Python - PySharm Edu.

Table 3

List of laboratory works

№ In order	Name of laboratory work (computer workshop)	Number of aud. hours
1	Laboratory work No. 1. Symbolic computation and symbolic learning (Topic 1).	7
2	Laboratory work No. 2. Metric and linear methods of learning (Topic 2,3,4,5).	7
3	Laboratory work No. 3. Bayesian methods of learning and regression recovery. (Topic 6,7).	8
4	Laboratory work No. 4. Logical teaching methods. (Topic 9, 10, 11).	7
5	Laboratory work No. 5. Neural network learning methods. (Topic 12,13,14).	7

Policy and control

1. Policy of academic discipline (educational component)

During classes on the subject "Statistical methods of machine learning", students must follow certain disciplinary rules:

• it is forbidden to be late for classes;

• at the entrance of the teacher, as a sign of greeting, persons studying at KPI named after Igor Sikorsky, must stand up;

• extraneous conversations or other noise that interferes with classes are not allowed;

• leaving the classroom during the lesson is allowed only with the teacher's permission;

• the use of mobile phones and other technical means is not allowed without the teacher's permission.

Laboratory works are submitted in person with a preliminary check of theoretical knowledge, which is necessary for the performance of laboratory work. Validation of practical results includes code review and execution of test tasks.

In the course of training, the teacher has the right to award up to 5 incentive points for early completion of laboratory work, for a demonstrated creative approach when completing an individual task, or for active participation in the discussion of issues related to the subject of a lecture or practical session.

The teacher may assign up to 5 penalty points for completing and submitting laboratory work after the specified deadline, for a significant number of missed classes, or for violating the rules of behavior in classes.

When conducting control measures and performing laboratory work, students must adhere to the rules of academic integrity. If a significant percentage of plagiarism or plagiarism is detected, the teacher may refuse to accept the given work and demand the honest implementation of the curriculum.

Types of control and rating system for evaluating learning outcomes (RSO)

Types of control from the academic discipline "Algorithms and calculation methods" include:

Laboratory works:

Independent performance of 5 laboratory works is planned.

The topics of laboratory works are coordinated in time and content with the topics of lectures. Performing laboratory work in its entirety allows you to acquire practical skills in applying numerical methods to solving various mathematical problems and to master modern algorithm programming technologies that are built on the basis of these methods.

Current control:

There are 5 current tests with closed tests in the TCEXAM system, which fully cover the subject of this academic discipline. Each current closed test contains 10 questions and lasts 10 minutes. The total time is 2 hours and includes testing time and time to solve organizational issues. In the case of distance learning, a closed current test is conducted at the beginning of the lecture that follows the lecture that concludes the next topic. In the case of face-to-face training, the time of the next current test is set by the teacher in agreement with the students.

Semester control:

The semester closed test is held at the end of the semester, lasts 30 minutes and consists of 25 questions.

Test:

conducted in the form of an interview with the student to objectively determine the level of knowledge, skills and practical skills acquired during the semester

Since the credit module has a semester certification in the form of credit, the rating evaluation system is built according to the RSO type - 1. The rating of the student from the credit module consists of the points he receives for the types of work according to table 4.

Table 4

Assessment of individual types of student's academic work

4 курс	
Type of educational work	Maximum number of points
Performance and protection of laboratory work #1	10
Performance and protection of laboratory work #2	10
Performance and protection of laboratory work #3	10
Performance and protection of laboratory work #4	10
Performance and protection of laboratory work #5	10
Current control	20
Semester control with closed tests	30
Total in semester	**100**

The individual semester rating (**RD**) of a student in an academic discipline consists of the points he receives for:
1) Current control: thematic tests with closed tests in the TCEXAM system (10 minutes each) – 6x5=30 points.
2) Semester control, which includes testing with closed tests in the TCEXAM system (60 minutes) - 30 points,
Total for semester control and current control - 60 points.
3) Performing laboratory work.
During the semester, students perform 4 laboratory works.
The maximum number of points for each laboratory work is 10.
Points are awarded for:
- timeliness of work submission for defense 0 – 1 point,
- preparation of the protocol of laboratory work 0 – 1 point,
- performance of meaningful task for work 0 - 5 points,
- verification of theoretical knowledge necessary for performing laboratory work 0 - 5 points.
Total for laboratory works (maximum number of points) – 40.
Performance of semester control testing.
The semester control test is performed on the cathedral server in the "TCEXAM" testing system. The same number of questions are randomly selected from each topic, a total of 30 questions, 4 answer options for each question, the time of the test is 60 minutes.
Execution of current control tests. Current control tests are performed on the cathedral server in the "TCEXAM" testing system.
Each current control test consists of 10 questions.
A full thorough answer to the test question is 0.5 points. An answer that includes the flaw of incompleteness - 0.375 points.
An answer containing two selected incorrect answers - 0.25 points.
An answer containing two selected incorrect answers - 0.125 points.
An answer that does not contain correct answer options - 0 points.

Calculation of the scale size (R) of the rating:
The sum of the weighted points of control measures during the semester is:
$R = \sum_k r_k$, where r_k is the maximum rating score of each of the control activities (control testing, laboratory work). The size of the rating scale from the credit module is:
R = 60 + 40 = 100 points.
If a student misses classes without a valid reason, fines are charged in the form of 1 point from the total points for 1 hour of absence (but not more than 0.1 R).

The student's individual semester rating (final semester rating grade **RD**) is the sum of points received by the student during the semester by participating in the prescribed control measures (control and laboratory works).

A necessary condition for a student's admission to credit is his individual semester rating (**RD**) of not less than 30 points, the absence of full arrears from laboratory work and not less than one positive certification. If at least one of the mentioned requirements is not fulfilled, the student will not be admitted to the credit.
The sum of the final semester (**RD**) and credit rating grades in points constitutes the final semester rating grade, which is converted into grades according to the national scale and the ECTS scale (Table 5).

Table 5

Correspondence of rating points to grades on the university scale

Scores	*Rating*
100-95	Perfectly
94-85	Very good
84-75	Good
74-65	Satisfactorily
64-60	Enough
Below 60	Unsatisfactorily
Admission conditions not met	Not allowed

1. Additional information on the discipline (educational component)

Teaching the discipline "Statistical methods of machine learning" for the specialty "Computer engineering" has its own specificity, which is related to the fact that the development and operation of computer equipment requires knowledge of the rules for building and analyzing algorithms. Familiarity with universal models of algorithms allows you to determine the complexity of the problem and the possibility of solving it with the help of a computer. Considerable attention should be paid to approaches that contribute to the assimilation of theoretical material and practical methods of algorithmization, the calculation of mathematical problems, the study of the peculiarities of the application of methods, the creation of highly effective algorithms for the implementation of numerical methods. Numerical methods are the basis of all calculations that can be performed using computers.

Working program of the academic discipline (syllabus):

Compiled by Doctor of Technical Sciences, Prof. Mykhailo Anatoliyovych Novotarskyi

Approved by the OT department (protocol No. 10 dated 05/25/2022)

Agreed: by the Methodical Commission of FIOT (protocol No. 10 dated 06/09/2022)

Minutes of the meeting of the FIOT Scientific Council No. 10 dated 06/13/2022