Federal State Autonomous Educational Institution of Higher Education "Moscow Institute of Physics and Technology (National Research University)"

APPROVED
Head of the Phystech School of Biological and Medical Physics
D.V. Kuzmin

Work program of the course (training module)

course: Machine Learning in Biology/Машинное обучение в биологии
major: Applied Mathematics and Physics
specialization: Applied Bioinformatics/Прикладная биоинформатика
Phystech School of Biological and Medical Physics
Chair of Bioinformatics and Systems Biology
term: 1
qualification: Master

Semester, form of interim assessment: 1 (fall) - Grading test

Academic hours: 30 AH in total, including:
lectures: 15 AH.
seminars: 15 AH.
laboratory practical: 0 AH.

Independent work: 60 AH.

In total: 90 AH, credits in total: 2

Author of the program: Y.A. Medvedeva, candidate of biological sciences

The program was discussed at the Chair of Bioinformatics and Systems Biology 04.06.2020

Annotation

"Machine Learning in biology" give theoretical and practical skills in the application of methods of machine analysis and interpretation of biological data. Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. For example, subject will provide you information about GA. A genetic algorithm (GA) is a search algorithm and heuristic technique that mimics the process of natural selection, using methods such as mutation and crossover to generate new genotypes in the hope of finding good solutions to a given problem. In machine learning, genetic algorithms were used in the 1980s and 1990s. Conversely, machine learning techniques have been used to improve the performance of genetic and evolutionary algorithms

1. Study objective

Purpose of the course

give theoretical and practical skills in the application of methods of machine analysis and interpretation of biological data.

Tasks of the course

- give basic knowledge of Python programming language;;
- make students familiar with modern understanding of machine learning algorithms;
- teach how to use the main databases in the field;
- form students' understanding of the applicability of various machine learning algorithms to biological problems

2. List of the planned results of the course (training module), correlated with the planned results of the mastering the educational program

Mastering the discipline is aimed at the formation of the following competencies:

Code and the name of the competence	Competency indicators
Gen.Pro.C-3 Select and/or develop approaches to professional problem-solving with consideration to the limitations and specifics of different solution methods	Gen.Pro.C-3.1 Analyze problems, plan research strategy to achieve solution(s), propose, and combine solution approaches
	Gen.Pro.C-3.2 Employ research methods to solve new problems and apply knowledge from various fields of science (technology)
	Gen.Pro.C-3.3 Gain knowledge of analytical and computational methods of problem-solving, understand the limitations of the implementation of the obtained solutions in practice

3. List of the planned results of the course (training module)

As a result of studying the course the student should:

know:

- Main tasks and methods of machine learning
- Theory of the main methods of machine learning;
- Ways of validating a model built to solve a problem;
- Examples of the use of machine learning methods in biological tasks

be able to:

- Use Python packages for machine learning (sklearn, xgboost, lightgbm, catboost, pytorch);
- Train and validate machine learning models;
- Determine the applicability of a particular machine learning method to this task;
- Use ready-made models to solve the proposed problem;
- Implement custom algorithms for machine learning;
- Use models to analyze data and patterns in them.;

master:

- work with large volumes of biological data;
- plan and implementation of multi-stage bioinformatic analysis;
- work with various machine learning algorithms.

4. Content of the course (training module), structured by topics (sections), indicating the number of allocated academic hours and types of training sessions

4.1. The sections of the course (training module) and the complexity of the types of training sessions

№	Topic (section) of the course	Types of training sessions, including independent work			
		Lectures	Seminars	Laboratory practical	Independent work
1	Introduction to machine learning. Examples of real biological problems and their solutions.	1			2
2	Supervised learning. Regression, classification. Quality metrics. Bias-variance tradeoff. ROC curve and PR curve.	1			2
3	Train and test sample. Crossvalidation and its variants. Biological data train-test split examples.	1			2
4	Linear regression. Logistic regression. L1 and L2 regularisations. Elastic Net.	1	1		4
5	Linear regression in data analysis.		1		4
6	Decision tree. Oblivious decision tree. Tree with models in the leaves.	1	1		4
7	Bootstrap and its applications. Confidence intervals using bootstrap. Bagging. Random forest. Bias-variance trade-off for bagging.	1	1		4
8	Boosting, key ideas. AdaBoost. Gradient boosting. XGBoost. LightGBM. CatBoost. DART. BooBag and BagBoo.		1		2
9	Categorial features. One-hot encoding. Lookup tables. Mean encoding.		1		2
10	KNN. SVM. Kernel trick. K-mers. deltaSVM.	1	1		2
11	Unsupervised learning. Clustering. Gaussian mixture models. Expectation maximization. KMeans. DBScan. HDBScan. Dimension reduction problem. PCA. TSNE. UMAP.	1	1		4
12	Semisupervised learning. Key ideas. Semisupervised decision trees. SemiBoost.	1	1		4
13	Neural networks. Perceptron. DNN. Chain rule. Activation. Vanishing gradient problem. Dropout. Batch normalisation.	1	1		4
14	Convolutional neural networks. Pooling. Sequence problems. Modified convolutions.	1	1		4
15	Recurrent neural networks. LSTM. GRU. Attention.	1	1		4

16	Autoencoders. Sparse autoencoders. Generative neural network. VAE. CVAE. GAN. DCGAN.	1	1		4
17	Ensembles. Stacking. Blending.	1	1		4
18	Transfer learning. One-shot learning.	1	1		4
AH in total		15	15		60
Exam preparation		0 AH.			
Total complexity		90 AH., credits in total 2			

4.2. Content of the course (training module), structured by topics (sections)

Semester: 1 (Fall)

1. Introduction to machine learning. Examples of real biological problems and their solutions.

Introduction to machine learning. Examples of real biological problems and their solutions. The focus of this module is to introduce the concepts of machine learning with as little mathematics as possible. We will introduce basic concepts in machine learning, including logistic regression, a simple but widely employed machine learning (ML) method. Also covered is multilayered perceptron (MLP), a fundamental neural network. The concept of deep learning is discussed, and also related to simpler models.

2. Supervised learning. Regression, classification. Quality metrics. Bias-variance tradeoff. ROC curve and PR curve.

Supervised learning. Regression, classification. Quality metrics. Bias-variance tradeoff. ROC curve and PR curve. In this module we will be discussing the mathematical basis of learning deep networks. We'll first work through how we define the issue of learning deep networks as a minimization problem of a mathematical function. After defining our mathematical goal, we will introduce validation methods to estimate real-world performance of the learned deep networks. We will then discuss how gradient descent, a classical technique in optimization, can be used to achieve this mathematical goal. Finally, we will discuss both why and how stochastic gradient descent is used in practice to learn deep networks.

3. Train and test sample. Crossvalidation and its variants. Biological data train-test split examples.

Train and test sample. Crossvalidation and its variants. Biological data train-test split examples. will cover model training, as well as transfer learning and fine-tuning. In addition to learning the fundamentals of a CNN and how it is applied, careful discussion is provided on the intuition of the CNN, with the goal of providing a conceptual understanding.

4. Linear regression. Logistic regression. L1 and L2 regularisations. Elastic Net.

Linear regression. Logistic regression. L1 and L2 regularisations. Elastic Net. Logistic regression is a method for classifying data into discrete outcomes. For example, we might use logistic regression to classify an email as spam or not spam. In this module, we introduce the notion of classification, the cost function for logistic regression, and the application of logistic regression to multi-class classification.

5. Linear regression in data analysis.

Linear regression in data analysis. What if your input has more than one value? In this module, we show how linear regression can be extended to accommodate multiple input features. We also discuss best practices for implementing linear regression.

6. Decision tree. Oblivious decision tree. Tree with models in the leaves.

Decision tree. Oblivious decision tree. Tree with models in the leaves.

7. Bootstrap and its applications. Confidence intervals using bootstrap. Bagging. Random forest. Bias-variance trade-off for bagging.

Bootstrap and its applications. Confidence intervals using bootstrap. Bagging. Random forest. Bias-variance trade-off for bagging.

8. Boosting, key ideas. AdaBoost. Gradient boosting. XGBoost. LightGBM. CatBoost. DART. BooBag and BagBoo.

Boosting, key ideas. AdaBoost. Gradient boosting. XGBoost. LightGBM. CatBoost. DART. BooBag and BagBoo.

9. Categorial features. One-hot encoding. Lookup tables. Mean encoding.

Categorial features. One-hot encoding. Lookup tables. Mean encoding.

10. KNN. SVM. Kernel trick. K-mers. deltaSVM.

KNN. SVM. Kernel trick. K-mers. deltaSVM. Support vector machines, or SVMs, is a machine learning algorithm for classification. We introduce the idea and intuitions behind SVMs and discuss how to use it in practice.

11. Unsupervised learning. Clustering. Gaussian mixture models. Expectation maximization. KMeans. DBScan. HDBScan. Dimension reduction problem. PCA. TSNE. UMAP.

Unsupervised learning. Clustering. Gaussian mixture models. Expectation maximization. KMeans. DBScan. HDBScan. Dimension reduction problem. PCA. TSNE. UMAP. We use unsupervised learning to build models that help us understand our data better. We discuss the k-Means algorithm for clustering that enable us to learn groupings of unlabeled data points.

12. Semisupervised learning. Key ideas. Semisupervised decision trees. SemiBoost.

Semisupervised learning. Key ideas. Semisupervised decision trees. SemiBoost.

13. Neural networks. Perceptron. DNN. Chain rule. Activation. Vanishing gradient problem. Dropout. Batch normalisation.

Neural networks. Perceptron. DNN. Chain rule. Activation. Vanishing gradient problem. Dropout. Batch normalisation. will cover the application of neural networks to natural language processing (NLP), from simple neural models to the more complex. The fundamental concept of word embeddings is discussed, as well as how such methods are employed within model learning and usage for several NLP applications. A wide range of neural NLP models are also discussed, including recurrent neural networks, and specifically long short-term memory (LSTM) models.

14. Convolutional neural networks. Pooling. Sequence problems. Modified convolutions.

Convolutional neural networks. Pooling. Sequence problems. Modified convolutions. A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, ConvNets have the ability to learn these filters/characteristics.

15. Recurrent neural networks. LSTM. GRU. Attention.

(RNN) is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence. This allows it to exhibit temporal dynamic behavior. Unlike feedforward neural networks, RNNs can use their internal state (memory) to process sequences of inputs.

16. Autoencoders. Sparse autoencoders. Generative neural network. VAE. CVAE. GAN. DCGAN.

Autoencoders. Sparse autoencoders. Generative neural network. VAE. CVAE. GAN. DCGAN. The aim of an autoencoder is to learn a representation (encoding) for a set of data, typically for dimensionality reduction, by training the network to ignore signal “noise”.

17. Ensembles. Stacking. Blending.

Ensembles. Stacking. Blending. Ensemble learning technique that uses predictions from multiple models (for example decision tree, knn or svm) to build a new model.

18. Transfer learning. One-shot learning.

Transfer learning. One-shot learning. knowledge gained while learning to recognize cars could apply when trying to recognize trucks. This area of research bears some relation to the long history of psychological literature on transfer of learning, although formal ties between the two fields are limited.

5. Description of the material and technical facilities that are necessary for the implementation of the educational process of the course (training module)

Equipment needed for lectures and seminars: computer and multimedia equipment (projector, sound system).

6. List of the main and additional literature, that is necessary for the course (training module) mastering

Main literature

Provided at the department:

1. R. Durbin, S. Eddy, A. Krogh, G. Mitchison. Analysis of biological sequences. Regular and chaotic dynamics, 2006.
2. Ian Goodfellow, Joshua Bengio, Aaron Curville. Deep learning
3. Christopher M. Bishop. "Pattern Recognition and Machine Learning"

Additional literature

Provided at the department:

1. S. Nikolenko, A. Kadurin, E. Arkhangelskaya. “Deep learning. Immersion in the world of neural networks "

7. List of web resources that are necessary for the course (training module) mastering

Scientific bibliographic and patent databases in the field of physico-chemical biology, available on the Internet in free mode - Science Citation Index (Web of Science), Medline (PubMed), Scientific Electronic Library (NEB). Kaggle.

8. List of information technologies used for implementation of the educational process, including a list of software and information reference systems (if necessary)

Internet access. For some of the lessons, you need Zoom. Google Drive to access course materials. The presence of smartphones / laptops during classes is encouraged to participate in interactive exercises.

9. Guidelines for students to master the course

A student who studies discipline must, on the one hand, master a general conceptual apparatus, and on the other hand, must learn to apply theoretical knowledge in practice.

As a result of studying the discipline, the student should know the basic definitions of the discipline, be able to apply this knowledge to solve various problems.

Successful learning requires:

- visits to all classes provided by the curriculum for the discipline;
- conducting the abstract of occupations;
- intense independent work of the student.

Independent work includes:

- reading recommended literature;
- study of educational material, preparation of answers to questions intended for self-study;
- solving problems offered to students in the classroom;
- preparation for performance of tasks of the current and intermediate certification.

An indicator of possession of the material is the ability to answer questions on discipline topics without an outline.

It is important to achieve an understanding of the material being studied, and not its mechanical memorization. If it is difficult to study individual topics, questions, you should seek advice from the teacher.

SUPPLEMENT

Assessment funds for course (training module)

major:	Applied Mathematics and Physics
specialization:	Applied Bioinformatics/Прикладная биоинформатика Phystech School of Biological and Medical Physics Chair of Bioinformatics and Systems Biology
term:	1
qualification:	Master

Semester, form of interim assessment: 1 (fall) - Grading test

Author: Y.A. Medvedeva, candidate of biological sciences

1. Competencies formed during the process of studying the course

Code and the name of the competence	Competency indicators
Gen.Pro.C-3 Select and/or develop approaches to professional problem-solving with consideration to the limitations and specifics of different solution methods	Gen.Pro.C-3.1 Analyze problems, plan research strategy to achieve solution(s), propose, and combine solution approaches
	Gen.Pro.C-3.2 Employ research methods to solve new problems and apply knowledge from various fields of science (technology)
	Gen.Pro.C-3.3 Gain knowledge of analytical and computational methods of problem-solving, understand the limitations of the implementation of the obtained solutions in practice

2. Competency assessment indicators

As a result of studying the course the student should:

know:

- Main tasks and methods of machine learning
- Theory of the main methods of machine learning;
- Ways of validating a model built to solve a problem;
- Examples of the use of machine learning methods in biological tasks

be able to:

- Use Python packages for machine learning (sklearn, xgboost, lightgbm, catboost, pytorch);
- Train and validate machine learning models;
- Determine the applicability of a particular machine learning method to this task;
- Use ready-made models to solve the proposed problem;
- Implement custom algorithms for machine learning;
- Use models to analyze data and patterns in them.;

master:

- work with large volumes of biological data;
- plan and implementation of multi-stage bioinformatic analysis;
- work with various machine learning algorithms.

3. List of typical control tasks used to evaluate knowledge and skills

During the current control, the student should be able to answer the following questions:

1. Examples of real biological problems and their solutions.
2. Supervised learning. Regression, classification. Quality metrics. Bias-variance tradeoff. ROC curve and PR curve.
3. Train and test sample. Crossvalidation and its variants. Biological data train-test split examples.
4. Linear regression. Logistic regression. L1 and L2 regularisations. Elastic Net
5. Linear regression in data analysis.
6. Decision tree. Oblivious decision tree. Tree with models in the leaves.
7. Bootstrap and its applications. Confidence intervals using bootstrap. Bagging. Random forest. Bias-variance trade-off for bagging.
8. Boosting, key ideas. AdaBoost. Gradient boosting. XGBoost. LightGBM. CatBoost. DART. BooBag and BagBoo.
9. Categorial features. One-hot encoding. Lookup tables. Mean encoding.
10. KNN. SVM. Kernel trick. K-mers. deltaSVM.
11. Unsupervised learning. Clustering. Gaussian mixture models. Expectation maximization. KMeans. DBScan. HDBScan. Dimension reduction problem. PCA. TSNE. UMAP.
12. Semisupervised learning. Key ideas. Semisupervised decision trees. SemiBoost.
13. Neural networks. Perceptron. DNN. Chain rule. Activation. Vanishing gradient problem. Dropout. Batch normalisation.

14. Convolutional neural networks. Pooling. Sequence problems. Modified convolutions.
15. Recurrent neural networks. LSTM. GRU. Attention.
16. Autoencoders. Sparse autoencoders. Generative neural network. VAE. CVAE. GAN. DCGAN.
17. Ensembles. Stacking. Blending.
18. Transfer learning. One-shot learning.

During the class, interactive discussions can take place in the course chats, which will be homework. It is possible to perform patent search as an independent task. Successful completion of all tasks in the course and the completion of control slices of knowledge gives an advantage in the differential credit.

4. Evaluation criteria

1. Examples of real biological problems and their solutions.
2. Supervised learning. Regression, classification. Quality metrics. Bias-variance tradeoff. ROC curve and PR curve.
3. Train and test sample. Crossvalidation and its variants. Biological data train-test split examples.
4. Linear regression. Logistic regression. L1 and L2 regularisations. Elastic Net
5. Linear regression in data analysis.
6. Decision tree. Oblivious decision tree. Tree with models in the leaves.
7. Bootstrap and its applications. Confidence intervals using bootstrap. Bagging. Random forest. Bias-variance trade-off for bagging.
8. Boosting, key ideas. AdaBoost. Gradient boosting. XGBoost. LightGBM. CatBoost. DART. BooBag and BagBoo.
9. Categorial features. One-hot encoding. Lookup tables. Mean encoding.
10. KNN. SVM. Kernel trick. K-mers. deltaSVM.
11. Unsupervised learning. Clustering. Gaussian mixture models. Expectation maximization. KMeans. DBScan. HDBScan. Dimension reduction problem. PCA. TSNE. UMAP.
12. Semisupervised learning. Key ideas. Semisupervised decision trees. SemiBoost.
13. Neural networks. Perceptron. DNN. Chain rule. Activation. Vanishing gradient problem. Dropout. Batch normalisation.
14. Convolutional neural networks. Pooling. Sequence problems. Modified convolutions.
15. Recurrent neural networks. LSTM. GRU. Attention.
16. Autoencoders. Sparse autoencoders. Generative neural network. VAE. CVAE. GAN. DCGAN.
17. Ensembles. Stacking. Blending.
18. Transfer learning. One-shot learning.

Exemples of exam tasks:

1. Unsupervised learning. Clustering. Gaussian mixture models. Expectation maximization. KMeans. DBScan. HDBScan. Dimension reduction problem. PCA. TSNE. UMAP.
2. Semisupervised learning. Key ideas. Semisupervised decision trees. SemiBoost.
3. Neural networks. Perceptron. DNN. Chain rule. Activation. Vanishing gradient problem. Dropout. Batch normalisation.
4. Convolutional neural networks. Pooling. Sequence problems. Modified convolutions.
5. Recurrent neural networks. LSTM. GRU. Attention.

The mark is excellent (10 points) - it is given to a student who has shown comprehensive, systematic, deep knowledge of the curriculum of the discipline, who has an interest in this subject area, has demonstrated the ability to confidently and creatively put them into practice in solving specific problems, and a free and proper substantiation of decisions.

The mark is excellent (9 points) - it is given to a student who has shown comprehensive, systematic, in-depth knowledge of the curriculum of the discipline and the ability to confidently put them into practice in solving specific problems, free and proper substantiation of the decisions made.

The mark is excellent (8 points) - given to a student who has shown comprehensive, systematic, in-depth knowledge of the curriculum of the discipline and the ability to confidently apply them in practice in solving specific problems, correct justification of decisions made, with some shortcomings.

A mark is good (7 points) - it is put up for a student, if he knows the material firmly, sets it up competently and in essence, knows how to apply the knowledge gained in practice, but does not competently substantiate the results obtained.

Evaluation is good (6 points) - it is put up to a student, if he knows the material firmly, sets it up correctly and in essence, knows how to apply this knowledge in practice, but admits some inaccuracies in the answer or in solving problems.

A mark is good (5 points) - it is given to a student, if he basically knows the material, correctly and essentially sets it out, knows how to apply this knowledge in practice, but allows a sufficiently large number of inaccuracies to answer or solve problems.

Grade satisfactorily (4 points) is given to a student who has shown the fragmented, fragmented nature of knowledge, insufficiently correct formulations of basic concepts, violations of the logical sequence in the presentation of program material, but at the same time he has mastered the main sections of the curriculum necessary for further education and can apply knowledge is modeled in a standard situation.

Grade satisfactorily (3 points) - given to a student who showed the fragmented, scattered nature of knowledge, making mistakes in formulating basic concepts, disrupting the logical sequence in presenting program material, poorly masters the main sections of the curriculum required for further education and even applies the knowledge gained in a standard situation.

The rating is unsatisfactory (2 points) - is given to a student who does not know most of the main content of the curriculum of the discipline, makes gross mistakes in the wording of the basic principles and does not know how to use this knowledge when solving typical tasks.

Unsatisfactory mark (1 point) - is given to a student who does not know the main content of the discipline's curriculum, makes gross errors in the wording of the basic concepts of the discipline and does not have any skills to solve typical practical problems.

5. Methodological materials defining the procedures for the assessment of knowledge, skills, abilities and/or experience

When conducting an oral differential test, the student is given 60 minutes to prepare. Interrogation of a student on a ticket on an oral differential test should not exceed one astronomical hour.