Laboratory work in the courses: “Data Mining”, “Machine Learning Methods”, “Mathematical Methods of Data Analysis”, “Theory of Statistical Decisions”

To be admitted to the exam (test), the minimum requirement (for a satisfactory grade) is to complete three laboratory works. It is not possible to receive an automatic grade for the exam. It is possible to take the exam early. To do this, you need to complete coursework and from three (for a “satisfactory” grade) to five or six (for an “excellent” grade) laboratory work.

Each work is assessed differentially, so the final assessment depends on both the number of works and the quality of their implementation. The early exam is accepted simultaneously with the defense of the course work and consists of answering questions on the lecture program. The final grade is determined at least from the marks for practice and for theoretical questions.

Coursework is not required to take the regular exam. Deadlines for submission (defense):

01.10, 22.10, 12.11, 02.12, 16.12. If the number of passed work is less than the number of past deadlines, then the grade for the (next) work decreases.

Laboratory work 1.

(Quasi)linear classification methods.

Exercise.

1. Download the Iris table from the UCI repository (can be from any available source: UCI, sklearn, kaggle). Place the data in a pandas DataFrame object. Calculate correlations between features on the entire table and separately by class (use pd.groupby). Visualize class distributions on all pairs of variables.

2. Select two variables. Construct and visualize (show the dividing curve) solutions using linear and quadratic discriminant, logistic regression, SVM (linear and quadratic kernel) methods. An example of a suitable visualization is available in sklearn.

In the following tasks, you can leave only two (least separable) classes.

3. Construct a linear discriminant on all variables. Visualize algorithm responses (highlighted, for example, with color) and object classes (highlighted, for example, with the shape of markers) in all two-dimensional subspaces.

4. On two variables from step 2, calculate the quadratic separating function directly from estimates of covariance matrices and averages (independently

implement a method without using a ready-made one). Visualize it and compare it with the solution from step 2.

Laboratory work 2.

"Naive Bayes" classifier.

Exercise.

1. Load the Mushroom table from the UCI repository (any source) into the object DataFrame.
Calculate the distribution of categorical feature values across classes (use pd.groupby). Visualize
distributions.

2. Construct a decision function for each variable based on frequencies. Calculate the accuracy of each solution (on the same training set). Find the most informative variable (with the minimum number of errors).

3. Build a “naive” Bayes classifier from sklearn. Assess accuracy.

4. Implement the method yourself without using a ready-made one. Compare the resulting solution with the library one. Add a regularizer to frequency estimates.

5. Apply the logistic regression method, using probability estimates subjected to inverse logistic transformation as variables.

Laboratory work 3.

Decision trees. Ensembles of decision trees.

1. Select a suitable data table. Construct and visualize a decision tree.

2. Apply the gradient boosting method. Calculate the significance of variables. Display a list of constructed trees.

3. Construct the dependence of the quality of the solution (on training and sliding control) on the number of vertices of the tree.

4. For the gradient boosting method, plot the dependence of the quality of the solution (on training and sliding control) on the number of trees.
For different tree depths, it is necessary to construct several dependences of quality on the number of trees in order to find the optimal combination of these parameters.

5. Follow the previous step for a random forest.

Laboratory work 4.

Dependency recovery task. Manipulation of signs. Dimensionality reduction.

1. Select a suitable data table (must contain numeric and categorical variables). Temporarily remove categorical features. Construct linear regression.

2. Construct a solution using the boosting method. Compare with linear regression.

3. Apply one hot and target encoding for categorical features (take data where such features exist). Compare accuracy.

4. Visualize objects (not necessarily for the same table) in space
main components.

Laboratory work 5.

Quality criteria. Error curve. Quality assessment.

1. Select a data table with unbalanced classes. Solve the classification problem using any suitable method.

2. Calculate accuracy, recall, specificity.

3. Construct an error curve and find the area under it. Plot a precision-recall curve for comparison.

4. Divide the data into training and testing samples. Construct an ROC curve for each of the subsamples. Construct an ROC curve based on cross-validation.

5. Investigate the impact of the choice of training criterion on AUC. One of the criteria is log loss, another one or two are optional. Alternatively: compare AdaBoost with gradient boosting.

Laboratory work 6.

Investigating the effectiveness of classification methods using statistical modeling.

Goal: to explore the statistical properties of empirical risk assessment rolling examination and the probability of misclassification on synthesized data.

Progress.

1. Come up with a two-dimensional probabilistic model for two classes. It may be a mixture of normal distributions (with different parameters) with a number of components greater than the number of classes.

2. Set parameters: training sample size (about 100 objects), number of cross-validation partitions. Select a classification method and set its parameters.

3. Repeat steps 4–7 a specified number of times (50–100). 4. Generate a training sample of a given size. 5. Construct a decision function. Calculate the empirical risk (the number of errors during training). 6. Generate a test sample of a sufficiently large size (more than 10,000 objects). Calculate an estimate of the probability of misclassification.

7. Calculate an estimate of the probability of misclassification using the sliding examination method (on the original training sample).

8. Tabulate the results

sample number	empirical moving risk	exam	control select

9. Calculate the means and standard deviations for each column.

10. Carry out a similar simulation by changing the probabilistic model, or classification method, or method parameters. 11. Draw conclusions about how the chosen classification method and its the parameters correspond to the complexity of the model and the sample size.

12. Optional: calculate bias-variance decomposition.

Laboratory work 7.

Ensemble methods: stacking, blending. Out-of-fold assessment.

1. Select a suitable data table for the classification task. Study the article by A.G. Dyakonova about staking.

2. Choose three different classification methods. Apply them to the problem individually. Rate the quality.

3. Implement and apply averaging ***(blending),*** incl. weighted. Estimate quality.

4. Implement and apply stacking. Rate the quality.

5. Investigate the influence of response bias of the basic methods on the training sample (if the basic methods and the top-level method are trained on the same sample).

Laboratory work 8.

Neural networks on tabular data.

1. Select a suitable classification or regression problem (data table).

2. Build a solution based on a fully connected neural network with several layers.

3. Select the architecture parameters (number and size of layers), activation functions and number of epochs.

4. Conduct an experiment with changing the loss function (use MSE for classification or logloss for regression).

5. Build a solution using the boosting method. Compare the quality with the neural network solution.

Course work.

Solving a real data analysis problem (kaggle, UCI). Alternative: extended work option 6.

The work is performed in teams (from 1 to 3 people), each team has an individual task, agreed upon with the teacher.

1. Select source data (kaggle, UCI or any other real data), formulate the problem.

2. Solve the problem using a suitable method. Pay attention to the selection of parameters (there must be an experimental justification for the choice of parameters).

3. Justify the choice of method. If there is more than 1 person in the team, then solve the problem using different methods.

4. Assess the quality of the solution. Assess the accuracy of quality assessment.

5. Submit a report (can be in the form of a notebook with comments).