

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

УТВЕРЖДАЮ

Ректор Белорусского
государственного университета

А.Д.Король

24 марта 2025 г.

Регистрационный № 2785/б.



МАШИННОЕ ОБУЧЕНИЕ И АНАЛИЗ ДАННЫХ

Учебная программа учреждения образования по учебной дисциплине
для специальности:

6-05-0533-10 Информатика

Профилизация: Дискретная математика и биоинформатика

2025 г.

Учебная программа составлена на основе ОСВО 6-05-0533-10-2023; учебных планов БГУ: № 6-5.3-58/03 от 15.05.2023.

СОСТАВИТЕЛИ:

Д.И.Пирштук – старший преподаватель кафедры дискретной математики и алгоритмики факультета прикладной математики и информатики Белорусского государственного университета.

РЕЦЕНЗЕНТЫ:

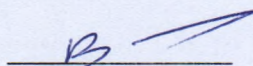
В.А. Ковалев – ведущий научный сотрудник ГНУ «Объединенный институт проблем информатики Национальной академии наук Беларуси», кандидат технических наук.

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:

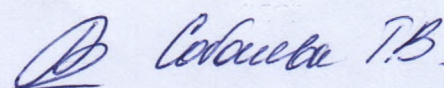
Кафедрой дискретной математики и алгоритмики БГУ
(протокол № 14 от 20.03.2025);

Научно-методическим Советом БГУ
(протокол № 8 от 20.03.2025).

Заведующий кафедрой



В.М.Котов



ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Цели и задачи учебной дисциплины

Учебная дисциплина «Машинное обучение и анализ данных» знакомит студентов с основами машинного обучения и закладывает необходимую теоретическую базу для применения алгоритмов анализа данных в прикладных задачах.

Цель дисциплины – сформировать теоретические и практические знания в области машинного обучения, современных методов восстановления зависимостей по эмпирическим данным.

Задачи учебной дисциплины:

1. Изучить теоретические основы машинного обучения и алгоритмы, обучаемые на данных.
2. Научить выполнять исследовательские циклы анализа данных: выдвижение гипотез, визуализацию, выбор подходящих алгоритмов, обучение моделей, оценку качества и интерпретацию полученных результатов.
3. Сформировать навыки решения прикладных задач на реальных данных при помощи алгоритмов обучения с учителей.

Место учебной дисциплины. В системе подготовки специалиста с высшим образованием учебная дисциплина **относится к** модулю дисциплин профилизации «Дискретная математика и биоинформатика» компонента учреждения образования.

Учебная программа составлена с учетом межпредметных связей и программ по дисциплинам. Основой для изучения учебной дисциплины являются дисциплина «Теория вероятностей и математическая статистика» модуля «Теория вероятностей и математическая статистика» компонента учреждения высшего образования, дисциплина «Основы и методологии программирования» модуля «Программирование» государственного компонента, дисциплина «Линейная алгебра» модуля «Геометрия и алгебра» государственного компонента, дисциплина Математический анализ» модуля «Математический анализ» государственного компонента, дисциплина «Модели и алгоритмы задач дискретной оптимизации» модуля «Дискретные структуры и алгоритмы» государственного компонента. Знания, полученные в учебной дисциплине, применяются при написании студентами курсовых и дипломных работ.

Требования к компетенциям

Освоение учебной дисциплины «Машинное обучение и анализ данных» должно обеспечить формирование следующих компетенций.

универсальные компетенции:

Решать стандартные задачи профессиональной деятельности на основе применения информационно-коммуникационных технологий.

специализированные компетенции:

Использовать современные методы машинного обучения для интеллектуального анализа данных и решения прикладных задач

В результате освоения учебной дисциплины студент должен:

знать:

- математические основы теории машинного обучения,
- основные алгоритмы классификации объектов.

уметь:

- выбирать метод машинного обучения, соответствующий исследуемой задаче,
- визуализировать результаты работы алгоритмов машинного обучения,
- интерпретировать полученные результаты.

иметь навык:

- владения программными средствами для разработки алгоритмов машинного обучения,
- практического решения задач интеллектуального анализа данных.

Структура учебной дисциплины

Дисциплина изучается в 5-м семестре. В соответствии с учебным планом всего на изучение учебной дисциплины «Машинное обучение и анализ данных» отведено для очной формы получения высшего образования – 108 часов, в том числе 68 аудиторных часов, лекции – 34 часа, лабораторные занятия – 34 часа.

Из них:

Лекции – 34 часа, лабораторные занятия – 30 часов, управляемая самостоятельная работа (УСР) – 4 часа.

Трудоемкость учебной дисциплины составляет 3 зачетные единицы.

Форма промежуточной аттестации – экзамен.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Раздел 1. Основы машинного обучения и анализа данных

Тема 1.1 Введение.

Постановки задач машинного обучения. Примеры задач. Виды данных. Признаки. Разведочный анализ данных и визуализация данных. Стандарт CRISP-DM. Добыча знаний. Задачи аналитики. Знакомство с программными средствами для анализа данных.

Тема 1.2. Линейные методы машинного обучения.

Линейная регрессия. Градиентный спуск. Метрики качества регрессии. Переобучение. Методы регуляризации. Многоклассовая и многометковая логистическая регрессия. Метрики качества классификации.

Тема 1.3. Оценка качества моделей.

Обучение, валидация и тестирование моделей. Перекрестная проверка. Метрики качества классификации и регрессии.

Тема 1.4. Основы глубокого обучения.

Человеческий мозг и история возникновения искусственных нейронных сетей. Перцептрон. Функции активации. Многослойные сети прямого распространения. Понятие глубокого обучения. Метод обратного распространения ошибки. Стохастический градиентный спуск. Способы борьбы с переобучением. Программные средства для обучения нейронных сетей.

Тема 1.5. Решающие деревья.

Дерево принятия решений. Критерии информативности для задач классификации и регрессии. Критерии остановки. Жадный алгоритм построения. Учёт пропусков в данных. Работа с категориальными признаками.

Тема 1.6. Композиции моделей.

Бутстрап, бэггинг и случайный лес. Разложение ошибки на смещение и разброс. Метод случайных подпространств. Ансамбли моделей. Бустинг. Градиентный бустинг над решающими деревьями и его реализации.

Тема 1.7. Байесовский подход в машинном обучении.

Задача оценивания вероятностей. Идея калибровки вероятностей. Наивный байесовский классификатор. Основы байесовского вывода.

Раздел 2. Избранные главы машинного обучения и анализа данных

Тема 2.1. Современные сверточные нейронные сети.

Сверточные нейронные сети. Современные архитектуры нейронных сетей для классификаций изображений. Подготовка изображений для обучения

сверточных нейронных сетей. Адаптация сверточных нейронных сетей к новым наборам данных. Примеры применения сверточных нейронных сетей.

Тема 2.2. Понижение размерности.

Метод главных компонент. Отбор признаков.

Тема 2.3. Алгоритмы обучения без учителя.

Постановка задачи кластеризации. Примеры задач. Обзор алгоритмов кластеризации. Метод k -средних. DBScan. EM-алгоритм. Оценка качества.

Тема 2.4. Анализ текстовых данных.

Конструирование признаков из текстовых данных. Рекуррентные нейронные сети для обработки последовательностей. Обратное распространение ошибки во времени. SimpleRNN, LSTM, GRU. Задача seq2seq. Механизм внимания. Генерация текстов с помощью нейронных сетей.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

очная (дневная) форма получения высшего образования с применением дистанционных образовательных технологий
(ДОТ)

Номер раздела, темы	Название раздела, темы	Количество аудиторных часов					Количество УСР	Форма контроля знаний
		Лекции	Практические занятия	Семинарские занятия	Лабораторные занятия	Иное		
1	2	3	4	5	6	7	8	9
1	О с н о в ы м а ш и н н о г о о б у ч е н и я и а н а л и з а д а н н ы х	18			18			
1.1	В в е д е н и е	2			4			О т ч е т ы п о д о м а ш н и м у п р а ж н е н и я м с и х у с т н о й з а щ и т о й .
1.2	Л и н е й н ы е м е т о д ы м а ш и н н о г о о б у ч е н и я	4			2			К о н т р о л ь н а я р а б о т а № 1.
1.3	Оценка качества моделей	2			2			О т ч е т ы п о д о м а ш н и м

								у п р а ж н е н и я м с и х у с т н о й з а щ и т о й .
1.4	О с н о в ы г л у б о к о г о о б у ч е н и я	4			4			К о н т р о л ь н а я р а б о т а № 2.
1.5	Р е ш а ю щ и е д е р е в ь я	2			2			О т ч е т ы п о д о м а ш н и м у п р а ж н е н и я м с и х у с т н о й з а щ и т о й .
1.6	К о м п о з и ц и и м о д е л е й	2			2			К о н т р о л ь н а я р а б о т а №3.
1.7	Б а й е с о в с к и й п о д х о д в м а ш и н н о м о б у ч е н и и	2			2			С о б е с е д о в а н и е
2	Избранные главы машинного обучения и анализа данных	16			12		4	
2.1	С о в р е м е н н ы е с в е р т о ч н ы е н е й р о н н ы е с е т и	4			2		4	з а щ и т а индивидуального или группового студенческого проекта, о ц е н и в а н и е

								на основе проектного метода.
2.2	Понижение размерности	2			2			Собеседова ние
2.3	Алгоритмы обучения без учителя	4			4			Собеседова ние.
2.4	Анализ текстовых данных	6			4			Контрольная работа №4.

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Основная литература

1. Николенко, С. И. Глубокое обучение : погружение в мир нейронных сетей / С. Николенко, А. Кудрин, Е. Архангельская. – Санкт-Петербург ; Москва ; Минск : Питер, 2022. – 476 с. – URL: <https://ibooks.ru/reading.php?short=1&productid=377026>.
2. Вейдман, С. Глубокое обучение. Легкая разработка проектов на Python / Сет Вейдман ; [пер. с англ. И. Рузмайкиной, А. Павлова]. – Санкт-Петербург [и др.] : Питер, 2021. – 272 с. – URL: <https://ibooks.ru/bookshelf/374461>.
3. Труш, Н. Н. Введение в компьютерный и интеллектуальный анализ данных : учебные материалы / Н. Н. Труш ; БГУ, Факультет прикладной математики и информатики, Кафедра теории вероятностей и математической статистики. – Минск : БГУ, 2022. – 69 с. – URL: <https://elib.bsu.by/handle/123456789/277034>.

Дополнительная литература

1. Андрей Бурков. Машинное обучение без лишних слов. - Санкт-Петербург : Питер, 2020. –192 с.
2. Вьюгин, В. В. Математические основы машинного обучения и прогнозирования : учебное пособие / В. В. Вьюгин. – Москва : МЦНМО, 2014. – 304 с.
3. Головкин, В. А. Нейросетевые технологии обработки данных : учебное пособие / В. А. Головкин, В. В. Краснопрошин. – Минск : БГУ, 2017. – 263 с. – URL: <https://elib.bsu.by/handle/123456789/193558>.
4. Гудфеллоу Я. Глубокое обучение / пер. с англ. А. А. Слинкина. – 2-е изд., испр. / Я. Гудфеллоу, И. Бенджио, А. Курвилль. - Москва : ДМК Пресс, 2018. – 652 с.
5. Жерон О. Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow: концепции, инструменты и техники для создания интеллектуальных систем / пер. с англ.– 2-е изд., доп. и перераб. – М.: Диалектика, 2020. – 1040 с.
6. Джеймс Г. Введение в статистическое обучение с примерами на языке R. Изд. второе, испр. Пер. с англ. С. Э. Мастицкого / Г. Джеймс, Д. Уиттон, Т. Хастис, Р. Тибишани. – Москва : ДМК Пресс, 2017. – 456 с.
7. Лобановский, Л. С. Теория распознавания: пособие для студентов учреждений высшего образования / Л. С. Лобановский, С. В. Леончик ; БГУ. – Минск : БГУ, 2020. – 79 с. – URL: <https://elib.bsu.by/handle/123456789/258076>.
8. Мэрфи, К. П. Вероятностное машинное обучение. Введение / К. П. Мэрфи ; перевод с английского А. А. Слинкина. – Москва : ДМК Пресс, 2022. – 940 с.
9. Николенко С., Тулупьев А. Самообучающиеся системы. – М.: МЦНМО, 2009. – 288 с.

10. Паклин, Н. Б. Бизнес-аналитика: от данных к знаниям : учеб. пособие / Н. Паклин, В. Орешков. – 2-е изд., доп. и перераб. – Санкт-Петербург [и др.] : Питер, 2010. – 701 с.
11. Плас Дж. Вандер. Python для сложных задач: наука о данных и машинное обучение. – (Серия «Бестселлеры O'Reilly»). – Санкт-Петербург : Питер, 2021. – 576 с.
12. Флах Петер. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / пер. с англ. А. А. Слинкина. – 2-е изд., эл. - Москва : ДМК Пресс, 2023. – 401 с.
13. Шолле, Ф. Глубокое обучение на Python = Deep Learning with Python / Франсуа Шолле ; [пер. с англ. А. Киселева]. – Санкт-Петербург [и др.] : Питер, 2018. – 397 с.
14. Яцков, Н. Н. Интеллектуальный анализ данных : пособие / Н. Н. Яцков. – Минск : БГУ, 2014. – 151 с. – URL: <https://elib.bsu.by/handle/123456789/114127>.

Перечень рекомендуемых средств диагностики и методика формирования итоговой отметки

Объектом диагностики компетенций студентов являются знания, умения, полученные ими в результате изучения учебной дисциплины. Выявление учебных достижений студентов осуществляется с помощью мероприятий текущей и промежуточной аттестации.

Для диагностики компетенций могут использоваться следующие средства текущей аттестации: контрольная работа, отчеты по домашним упражнениям с их устной защитой, оценивание на основе проектного метода, собеседования.

Формой промежуточной аттестации по дисциплине «Машинное обучение и анализ данных» учебным планом предусмотрен **экзамен**.

Для формирования итоговой отметки по учебной дисциплине используется модульно-рейтинговая система оценки знаний студента, дающая возможность проследить и оценить динамику процесса достижения целей обучения. Рейтинговая система предусматривает использование весовых коэффициентов для текущей и промежуточной аттестации студентов по учебной дисциплине.

Формирование итоговой отметки в ходе проведения контрольных мероприятий текущей аттестации (примерные весовые коэффициенты, определяющие вклад текущей аттестации в отметку при прохождении промежуточной аттестации):

- о т ч е т ы п о д о м а ш н и м у п р а ж н е н и я м с
и х у с т н о й з а щ и т о й – 50 %;
- контрольные работы – 50 %.

Итоговая отметка по дисциплине рассчитывается на основе итоговой отметки текущей аттестации (модульно-рейтинговой системы оценки знаний) – 60% и экзаменационной отметки – 40%.

Примерный перечень заданий для управляемой самостоятельной работы

Управляемая самостоятельная работа предлагается в виде заданий проектного типа.

Тема 2.1. Современные нейронные сети. (4 ч)

Участие в соревновании по анализу данных на платформе <https://kaggle.com> или решение других исследовательских задач.

Обеспечение на Образовательном портале БГУ – инструкция по выполнению проектов.

(Форма контроля – защита индивидуального или группового студенческого проекта).

Примерный перечень лабораторных занятий

Занятие 1. Знакомство с языком программирования Python и средой Google Colab. Матричные вычисления с помощью библиотеки NumPy.

Занятие 2. Работа с табулированными данными. Знакомство с библиотекой pandas. Разведочный анализ данных. Визуализация данных с помощью matplotlib и seaborn.

Занятие 3. Библиотека scikit-learn. Линейная регрессия.

Занятие 4. Обучение линейных классификаторов и оценка качества полученных моделей.

Занятие 5. Основы тензорных вычислений с помощью библиотеки PyTorch. Матрично-векторное дифференцирование.

Занятие 6. Набор данных MNIST. Основы обучения полносвязных нейронных сетей с помощью библиотеки PyTorch.

Занятие 7. Деревья принятия решений. Визуализация деревьев. Оценка важности признаков.

Занятие 8. Библиотека CatBoost.

Занятие 9. Байесовский подход.

Занятие 10. Обучение сверточных нейронных сетей с помощью библиотеки PyTorch. Аугментация изображений.

Занятие 11. Метод главных компонент для изображений.

Занятие 12. Кластеризация данных.

Занятие 13. EM-алгоритм.

Занятие 14. Предобработка текстовых данных. Конструирование признаков из текстовых данных. Наивный байесовский классификатор в задачах обработки текстов.

Занятие 15. Генерация поэзии с помощью рекуррентных нейронных сетей.

Рекомендуемая тематика контрольных работ:

1. Контрольная работа № 1 «Линейные модели».
2. Контрольная работа № 2 «Обучение нейронных сетей».
3. Контрольная работа № 3 «Композиции моделей».

4. Контрольная работа № 4 «Избранные главы машинного обучения и анализа данных».

Описание инновационных подходов и методов к преподаванию учебной дисциплины

При организации образовательного процесса большинства практических занятий используется **практико-ориентированный подход**, который предполагает:

- освоение содержания образования через решения практических задач;
- приобретение навыков эффективного выполнения разных видов профессиональной деятельности.

Также при организации образовательного процесса используются **методы группового обучения, проектного обучения и учебной дискуссии**. Студентам предлагается выполнить часть домашних заданий в форме проекта в группах до 4 человек. Задания предполагают предварительное обсуждение в форме мозгового штурма.

Выполнение проекта предусматривает самостоятельную работу с научными и техническими источниками по теме курса, самостоятельный поиск и выбор способа решения задачи, составление плана и разделение задач между участниками группы.

В конце курса предусмотрена устная защита проекта с критическим анализом идей, сгенерированных в ходе мозгового штурма, и ретроспективой выполненной работы.

Комбинация методов предполагает

- ориентацию на генерирование идей, реализацию групповых студенческих проектов, развитие предпринимательской культуры;
- способ организации учебной деятельности студентов, развивающий актуальные для учебной и профессиональной деятельности навыки планирования, самоорганизации, сотрудничества и предполагающий создание собственного продукта;
- приобретение навыков для решения исследовательских, творческих, социальных, предпринимательских и коммуникационных задач;
- появление нового уровня понимания изучаемой темы, применение знаний (теорий, концепций) при решении проблем, определение способов их решения.

Методические рекомендации по организации самостоятельной работы

Для организации самостоятельной работы студентов по учебной дисциплине следует использовать современные информационные технологии: разместить в сетевом доступе комплекс учебных и учебно-методических материалов (учебно-программные материалы, учебное издание для теоретического изучения дисциплины, презентации лекций, методические указания к практическим занятиям, электронные версии домашних заданий,

материалы текущего контроля и текущей аттестации, позволяющие определить соответствие учебной деятельности обучающихся требованиям образовательных стандартов высшего образования и учебно-программной документации, в том числе вопросы для подготовки к экзамену, задания, вопросы для самоконтроля, список рекомендуемой литературы, информационных ресурсов и др.).

Управляемая самостоятельная работа (консультационно-методическая поддержка и контроль) дисциплины обеспечивается средствами образовательного портала БГУ LMS Moodle.

Примерный перечень вопросов к экзамену

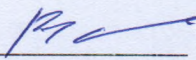
1. Постановка задач машинного обучения.
2. Примеры задач машинного обучения. Виды данных. Признаки.
3. Добыча знаний. Задачи аналитики. Стандарт CRISP-DM.
4. Линейная регрессия. Векторная форма и аналитическое решение для задачи линейной регрессии.
5. Градиентный спуск для задачи линейной регрессии. Стохастический градиентный спуск.
6. Линейная классификация. Логистическая регрессия.
7. Способы оценки качества моделей.
8. Метрики качества классификации и регрессии.
9. Проблема переобучения. Методы регуляризации для линейных моделей.
10. Человеческий мозг и история возникновения искусственных нейронных сетей.
11. Перцептрон. Функции активации.
12. Метод обратного распространения ошибки.
13. Глубокое и поверхностное обучение. Многослойные сети прямого распространения. Проблема внутреннего сдвига переменных.
14. Дерево принятия решений. Жадный алгоритм построения.
15. Критерии информативности для задач классификации при построении деревьев принятия решений.
16. Критерии информативности для регрессии и критерии остановки в задачах построения деревьев принятия решений. Стрижка деревьев.
17. Подготовка категориальных признаков для решающих деревьев. Учёт пропусков в данных при использовании решающих деревьев.
18. Важность признаков в решающих деревьях. Отбор признаков с помощью линейной регрессии.
19. Бутстрап, бэггинг и случайный лес. Метод случайных подпространств.
20. Разложение ошибки на смещение и разброс.
21. Градиентный бустинг. Идея, алгоритм, применение в задачах регрессии.
22. Реализации градиентного бустинга над решающими деревьями. Настройка гиперпараметров. Способы борьбы с переобучением.

23. Калибровка вероятностей.
24. Сверточные нейронные сети. Идея. Операции свертки и субдискретизации. Архитектура VGG.
25. Современные сверточные нейронные сети. Остаточное обучение на примере архитектуры ResNet151. Аксиальные и сепарабельные свертки. MobileNetV2. Адаптация сверточных нейронных сетей к новым наборам данных.
26. Проблема затухающего градиента и способы ее решения.
27. Подготовка данных для обучения сверточных нейронных сетей.
28. Стилизация изображений с помощью алгоритма Гатиса.
29. Метод главных компонент.
30. Алгоритмы кластеризации.
31. Конструирование текстовых признаков. Наивный байесовский классификатор в задачах обработки текстов.
32. Рекуррентные нейронные сети. Проблема обучения SimpleRNN и ее решение в LSTM и GRU.
33. Механизм внимания. Squeeze-and-Excitation блок в сверточных нейронных сетях.
34. Генерация текстов с помощью алгоритма CharRNN.

СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ УО

Название учебной дисциплины, с которой требуется согласование	Название кафедры	Предложения об изменениях в содержании учебной программы учреждения высшего образования по учебной дисциплине	Решение, принятое кафедрой, разработавшей учебную программу (с указанием даты и номера протокола)
Учебная дисциплина не требует согласования			

Заведующий кафедрой дискретной математики и алгоритмики

доктор физико-математических наук, профессор  В.М.Котов

20 марта 2025 г.

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ УО

на ____ / ____ учебный год

№ п/п	Дополнения и изменения	Основание

Учебная программа пересмотрена и одобрена на заседании кафедры
_____ (протокол № ____ от _____ 202_ г.)

Заведующий кафедрой

УТВЕРЖДАЮ
Декан факультета
