Белорусский государственный университет

УТВЕРЖДАЮ
Проректор по учебной работе
и образовательным инновациям
_____О.Г. Прохоренко
30 июня 2023 г.
Регистрационный № УД-12471/уч.

Машинное обучение и анализ данных

Учебная программа учреждения высшего образования по учебной дисциплине для специальности:

1-31 03 04 Информатика

2023 г.

Учебная программа составлена на основе ОСВО 1-31 03 04-2021, типового учебного плана №G 31-1-029/пр-тип от 30.06.2021, учебных планов №G 31-1-031/уч. от 30.06.2021, №G 31-1-021/уч.ин. от 23.07.2021.

СОСТАВИТЕЛЬ:

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

РЕЦЕНЗЕНТ:

В.А. Ковалев – ведущий научный сотрудник Объединенного института проблем информатики Национальной академии наук РБ, кандидат технических наук.

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:

Кафедрой дискретной математики и алгоритмики
(протокол № 16 от 05.05.2023 г.);

Научно-методическим Советом БГУ

(протокол № 9 от 29.06.2023 г.)

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 учебной 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 задачах на реальных данных при помощи алгоритмов обучения с учителей.

Место учебной дисциплины в системе подготовки специалиста с высшим образованием.

Учебная дисциплина относится **к циклу** дисциплин специализации компонента учреждения высшего образования.

Программа составлена с учетом межпредметных **связей** с учебными дисциплинами. Основой для изучения учебной дисциплины являются дисциплина «Теория вероятностей и математическая статистика» модуля «Теория вероятностей и математическая статистика» компонента учреждения высшего образования, дисциплина «Методы оптимизации» модуля «Математические методы принятия решений» компонента учреждения высшего образования, дисциплины «Основы и методологии программирования» и «Разработка кросс-платформенных приложений» модуля «Программирование» государственного компонента, дисциплина «Линейная алгебра» модуля «Геометрия и алгебра» государственного компонента, дисциплина «Дифференциальное и интегральное исчисление» модуля «Математический анализ» государственного компонента, дисциплина «Модели и алгоритмы задач дискретной оптимизации» модуля «Дискретные структуры и алгоритмы» государственного компонента. Знания, полученные в учебной дисциплине, используются при изучении дисциплин «Основы цифровой обработки изображений» и «Введение в биоинформатику» модуля

«Дисциплины специализации» компонента учреждения высшего образования.

Требования к компетенциям

Освоение учебной дисциплины «Машинное обучение и анализ данных» должно обеспечить формирование следующей **универсальной компетенции:**

УК-2. Решать стандартные задачи профессиональной деятельности на основе применения информационно-коммуникационных технологий.

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.

владеть:

– программными средствами для разработки алгоритмов машинного обучения,

– навыками практического решения задач интеллектуального анализа данных.

Структура учебной дисциплины

Дисциплина изучается в 5-м семестре. 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, из них: лекции – 34 часа, лабораторные занятия – 30 часов, управляемая самостоятельная работа – 4 часа.

Трудоемкость учебной дисциплины составляет 3 зачетные единицы.

Форма текущей 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. Знакомство с программными средствами для анализа данных.

Тема 1.2. Линейные методы машинного обучения.

Линейная регрессия. Градиентный спуск. Метрики качества регрессии. Переобучение. Методы регуляризации. Логистическая регрессия и оценки вероятности классов. Метрики качества классификации. Разложение ошибки на смещение и разброс.

Тема 1.3. Прикладной анализ данных.

Добыча знаний. Задачи аналитики. Разведочный анализ данных и визуализация данных. Стандарт CRISP-DM. Обучение, валидация и тестирование моделей. Перекрестная проверка.

Раздел 2. Алгоритмы обучения с учителем

Тема 2.1. Оценивание вероятностей.

Задача оценивания вероятностей. Идея калибровки вероятностей. Наивный байесовский классификатор. Многоклассовая и многометковая логистическая регрессия.

Тема 2.2. Метрические методы машинного обучения.

Метод ближайших соседей. Метод парзеновского окна. Проблема размерности.

Тема 2.3. Решающие деревья.

Дерево принятия решений. Критерии информативности. Жадный алгоритм построения. Учёт пропусков в данных. Работа с категориальными признаками.

Тема 2.4. Композиции моделей.

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

Раздел 3. Избранные главы машинного обучения

Тема 3.1. Подготовка и предобработка данных.

Конструирование признаков. Отбор признаков. Подготовка текстовых данных и изображений.

Тема 3.2. Рекомендательные системы.

Постановка задачи и метрики качества построения рекомендаций. Коллаборативная фильтрация. Матричные разложения.

Тема 3.3. Нейронные сети.

Человеческий мозг и история возникновения искусственных нейронных сетей. Перцептрон. Функции активации. Многослойные сети прямого распространения. Метод обратного распространения ошибки. Стохастический градиентный спуск. Современные сверточные нейронные сети и их применение. Адаптация сверточных нейронных сетей к новым наборам данных. Программные средства для обучения нейронных сетей.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

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

Номер раздела, темы	Название раздела, темы	Количество аудиторных часов					Количество часов УСР	Форма контроля знаний
		Лекции	Практические занятия	Семинарские занятия	Лабораторные занятия	Иное		
1	2	3	4	5	6	7	8	9
1	**Основы машинного обучения и анализа данных**	**8**			**8**			
1.1	Введение	2			4			Собеседование.
1.2	Линейные методы машинного обучения	4			2			Отчеты по домашним упражнениям с их устной защитой.
1.3	Прикладной анализ данных	2			2			Контрольная работа № 1.
2	**Алгоритмы обучения с учителем**	**12**			**12**			
2.1	Оценивание вероятностей	4			2			Собеседование.
2.2	Метрические методы машинного обучения	2			2			Отчеты по домашним упражнениям с их устной защитой.

2.3	Решающие деревья	2			4			Коллоквиум.
2.4	Композиции моделей	4			4			Контрольная работа № 2.
3	**Избранные главы машинного обучения**	**14**			**10**		**4**	
3.1	Подготовка и предобработка данных	2			2			Отчеты по домашним упражнениям с их устной защитой.
3.2	Рекомендательные системы	2			2			Контрольная работа № 3.
3.3	Нейронные сети	10			6		4	Оценивание на основе проектного метода.

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Перечень основной литературы

1. Лимановская, О. В. Основы машинного обучения : учебное пособие / О. В. Лимановская, Т. И. Алферьева. — 2-е изд. — Москва : ФЛИНТА, 2022. — 88 с. — ISBN 978-5-9765-5006-3. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/231677 (дата обращения: 31.03.2023). — Режим доступа: для авториз. пользователей.
2. Митина, О. А. Технологии и инструментарий машинного обучения : учебное пособие / О. А. Митина, В. В. Жаров. — Москва : РТУ МИРЭА, 2023. — 203 с. — ISBN 978-5-7339-1758-0. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/368633 (дата обращения: 31.03.2023). — Режим доступа: для авториз. пользователей.
3. Николенко С. Глубокое обучение. — (Серия «Библиотека программиста») / С. Николенко, А. Кадурин, Е. Архангельская. - Санкт-Петербург : Питер, 2020. - 480 с. - ISBN 978-5-4461-1537-2. - URL: https://ibooks.ru/bookshelf/377026/reading (дата обращения: 31.03.2023). - Текст: электронный.
4. Мэрфи, К. П. Вероятностное машинное обучение. Введение / К. П. Мэрфи ; перевод с английского А. А. Слинкина. — Москва : ДМК Пресс, 2022. — 940 с. — ISBN 978-5-93700-119-1. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/314891 (дата обращения: 31.03.2023). — Режим доступа: для авториз. пользователей.
5. 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. - ISBN 978-5-89818-300-4. - URL: https://ibooks.ru/bookshelf/392001/reading (дата обращения: 10.03.2024). - Текст: электронный.
6. Шолле Франсуа. Глубокое обучение на Python. 2-е межд. издание. — (Серия «Библиотека программиста»). - Санкт-Петербург : Питер, 2023. - 576 с. - ISBN 978-5-4461-1909-7. - URL: https://ibooks.ru/bookshelf/386793/reading (дата обращения: 31.03.2023). - Текст: электронный.

Перечень дополнительной литературы

1. 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.

2. Вьюгин, В. В. Математические основы машинного обучения и прогнозирования : учебное пособие / В. В. Вьюгин. — Москва : МЦНМО, 2014. — 304 с. — ISBN 978-5-4439-2014-6. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/56397 (дата обращения: 31.03.2023). — Режим доступа: для авториз. пользователей.
3. 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.
4. Андрей Бурков. Машинное обучение без лишних слов. - Санкт-Петербург : Питер, 2020. - 192 с. - ISBN 978-5-4461-1560-0. - URL: https://ibooks.ru/bookshelf/367991/reading (дата обращения: 31.03.2023). - Текст: электронный.
5. 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. Хасти, Р. Тибширани. - Москва : ДМК Пресс, 2017. - 456 с. - ISBN 978-5-97060-495-3. - URL: https://ibooks.ru/bookshelf/364394/reading (дата обращения: 31.03.2023). - Текст: электронный.
6. Николенко С., Тулупьев А. Самообучающиеся системы. – М.: МЦНМО, 2009. – 288 с.
7. Шолле, Ф. Глубокое обучение с R и Keras : практическое руководство / Ф. Шолле ; пер. с англ. В.С. Яценкова. - Москва : ДМК Пресс, 2023. - 646 с. - ISBN 978-5-93700-189-4. - Текст : электронный. - URL: https://znanium.com/catalog/product/2109573 (дата обращения: 31.03.2023). – Режим доступа: по подписке.
8. Гудфеллоу Я. Глубокое обучение / пер. с анг. А. А. Слинкина. - 2-е изд., испр. / Я. Гудфеллоу, И. Бенджио, А. Курвилль. - Москва : ДМК Пресс, 2018. - 652 с. - ISBN 978-5-97060-618-6. - URL: https://ibooks.ru/bookshelf/363710/reading (дата обращения: 31.03.2023). - Текст: электронный.
9. Плас Дж. Вандер. Python для сложных задач: наука о данных и машинное обучение. — (Серия «Бестселлеры O’Reilly»). - Санкт-Петербург : Питер, 2021. - 576 с. - ISBN 978-5-4461-0914-2. - URL: https://ibooks.ru/bookshelf/376830/reading (дата обращения: 31.03.2023). - Текст: электронный.
10. Лобановский, Л. С. Теория распознавания: пособие для студентов учреждений высшего образования, обучающихся по специальности 1-31 03 07 "Прикладная информатика (по направлениям)", направление специальности 1-31 03 07-02 "Прикладная информатика (информационные технологии телекоммуникационных систем)" / Л. С. Лобановский, С. В. Леончик ; БГУ. - Минск : БГУ, 2020. - 79 с. : ил. ; 20х14 см. - (Пособие). - Библиогр.: с. 76–78. - ISBN 978-985-566-990-7.

Перечень рекомендуемых средств диагностики и методика формирования итоговой оценки

Для диагностики компетенций в рамках учебной дисциплины рекомендуется использовать следующие формы:

1. Устная форма: собеседование, коллоквиум.
2. Письменная форма: контрольные работы.
3. Устно-письменная форма: отчеты по домашним упражнениям с их устной защитой, оценивание на основе проектного метода.

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

При формировании итоговой отметки используется рейтинговая система оценки знаний студента, дающая возможность проследить и оценить динамику процесса достижения целей обучения.

Рейтинговая система предусматривает использование весовых коэффициентов в ходе проведения контрольных мероприятий текущей аттестации.

Примерные весовые коэффициенты, определяющие вклад текущей аттестации в отметку при прохождении промежуточной аттестации:

Формирование отметки за текущую аттестацию:

- отчет по домашним упражнениям с их устной защитой - 50 %;
- контрольные работы - 30 %;
- коллоквиум - 20 %.

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

Примерный перечень заданий для управляемой самостоятельной работы студентов

Управляемая самостоятельная работа предлагается в виде заданий проектного типа.

Тема 3.3. Нейронные сети (4 ч)

Примером такого задания может быть участие в международных соревнованиях по анализу данных на платформе https://kaggle.com или

решение других исследовательских задач. Обеспечение на образовательном портале – инструкция по выполнению проектов.

Форма контроля – защита индивидуального или группового студенческого проекта.

Примерная тематика лабораторных занятий

Занятия 1-2. Знакомство с языком программирования Python и средой Google Colab.

Занятие 3. Построение линейных классификаторов.

Занятие 4. Работа с табулированными данными. Знакомство с библиотекой Pandas. Разведочный анализ данных. Визуализация данных.

Занятие 5. Задача кредитного скоринга. Логистическая регрессия.

Занятие 6. Набор данных MNIST. Метод ближайших соседей.

Занятие 7-8. Задача кредитного скоринга. Решающие деревья и их ансамбли. Работа с пропусками в данных и с категориальными признаками.

Занятие 9-10. Задача кредитного скоринга. Ансамбли решающих деревьев. Интерпретация результатов.

Занятие 11. Задача кредитного скоринга. Построение дополнительных признаков.

Занятие 12. Рекомендательные системы.

Занятие 13. Знакомство с библиотеками Tensorflow и Keras.

Занятие 14. Обучение нейронных сетей для классификации данных из набора MNIST.

Занятие 15. Адаптация сверточных нейронных сетей к новым наборам данных.

Рекомендуемая тематика контрольных работ и коллоквиума:

1. Контрольная работа № 1 «Основы машинного обучения».
2. Контрольная работа № 2 «Решающие деревья и их ансамбли».
3. Контрольная работа № 3. «Избранные главы машинного обучения».
4. Коллоквиум «Основы машинного обучения. Алгоритмы обучения с учителем».

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 обучающихся

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 задачи обучения по прецедентам.
2. Примеры задач машинного обучения. Виды данных. Признаки.
3. Добыча знаний. Задачи аналитики. Стандарт CRISP-DM.
4. Метод *k* ближайших соседей.
5. Линейная регрессия. Векторная форма и аналитическое решение для задачи линейной регрессии.
6. Градиентный спуск для задачи линейной регрессии. Стохастический градиентный спуск.
7. Линейная классификация. Логистическая регрессия.
8. Способы оценки качества моделей.
9. Метрики качества классификации и регрессии.
10. Проблема переобучения. Методы регуляризации для линейных моделей.
11. Конструирование текстовых признаков. Наивный байесовский классификатор в задачах обработки текстов.
12. Дерево принятия решений. Жадный алгоритм построения.
13. Критерии информативности для задач классификации при построении деревьев принятий решений.
14. Критерии информативности для регрессии и критерии остановки в задачах построения деревьев принятий решений. Стрижка деревьев.
15. Подготовка категориальных признаков для решающих деревьев. Учёт пропусков в данных при использовании решающих деревьев.
16. Бутстрап, бэггинг и случайный лес. Метод случайных подпространств.
17. Разложение ошибки на смещение и разброс.
18. Градиентный бустинг. Идея, алгоритм, применение в задачах регрессии.
19. Реализации градиентного бустинга над решающими деревьями. Настройка гиперпараметров. Способы борьбы с переобучением.
20. Алгоритмы коллаборативной фильтрации для построения рекомендаций.
21. Построение рекомендаций с помощью матричного разложения со скрытыми признаками.
22. Факторизационные машины. Машины факторизации с учетом полей.
23. Важность признаков в решающих деревьев. Отбор признаков с помощью линейной регрессии.

24. Человеческий мозг и история возникновения искусственных нейронных сетей.

25. Перцептрон. Функции активации.

26. Метод обратного распространения ошибки.

27. Глубокое и поверхностное обучение. Многослойные сети прямого распространения. Проблема внутреннего сдвига переменных.

28. Проблема затухающего градиента и способы ее решения.

29. Сверточные нейронные сети. Идея. Операции свертки и субдискретизации. Архитектура VGG.

30. Современные сверточные архитектуры: ResNet. Сепарабельные свертки. Squeeze-and-Excitation блок.

31. Подготовка данных для обучения сверточных нейронных сетей.

32. Стилизация изображений с помощью сверточных нейронных сетей.

33. Адаптация сверточных нейронных сетей к новым наборам данных.

ПРОТОКОЛ СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ УВО

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)
1. Основы цифровой обработки изображений	Биомедицинской информатики	нет	Изменений не требуется (протокол № 16 от 05.05.2023 г.)
2. Введение в биоинформатику	Биомедицинской информатики	нет	Изменений не требуется (протокол № 16 от 05.05.2023 г.)

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ ПО ИЗУЧАЕМОЙ УЧЕБНОЙ ДИСЦИПЛИНЕ

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
_____ _____ _____