

Белорусский государственный университет

УТВЕРЖДАЮ

Проректор по учебной работе и
образовательным инновациям

О.Н. Здрок

«02» июля 2021 г.



Регистрационный № УД- 9927/уч.

Машинное обучение и анализ данных

**Учебная программа учреждения высшего образования
по учебной дисциплине для специальности:**

1-31 03 04 Информатика

2021 г.

Учебная программа составлена на основе ОСВО 1-31 03 04-2013 и учебных планов УВО G31-169/уч. 30.05.2013 г., G31и-192/уч. 30.05.2013 г.

СОСТАВИТЕЛЬ:

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

РЕЦЕНЗЕНТЫ:

А.А. Толстик – старший преподаватель кафедры вычислительной математики ФПМИ БГУ;

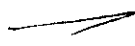
Д.В. Пекар – математик управления разработки программного обеспечения ООО «Бануба Девелопмент», кандидат технических наук.

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:

Кафедрой дискретной математики и алгоритмики
(протокол № 23 от 24.06.2021 г.);

Научно-методическим Советом БГУ
(протокол № 7 от 30.06.2021 г.)

Заведующий кафедрой

В 

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 задач на реальных данных при помощи алгоритмов обучения с учителями.

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 дисциплины «Теория вероятностей и математическая статистика», «Программирование», «Вычислительные методы алгебры» и «Математический анализ». Знания, полученные в учебной дисциплине, используются при изучении дисциплин специализации «Основы цифровой обработки изображений», «Алгоритмы в биоинформатике».

Требования к компетенциям

Освоение учебной дисциплины «Машинное обучение и анализ данных» должно обеспечить формирование следующих академических, социально-личностных и профессиональных компетенций:

академические компетенции:

АК-1. Уметь применять базовые научно-теоретические знания для решения теоретических и практических задач;

АК-2. Владеть системным и сравнительным анализом;

АК-3. Владеть исследовательскими навыками;

АК-4. Уметь работать самостоятельно;

АК-5. Быть способным вырабатывать новые идеи (обладать креативностью);

АК-6. Владеть междисциплинарным подходом при решении проблем;

АК-7. Иметь навыки, связанные с использованием технических устройств, управлением информацией и работой с компьютером;

АК-8. Иметь лингвистические навыки (устная и письменная коммуникация);

АК-9. Уметь учиться, повышать свою квалификацию в течение всей жизни.

социально-личностные компетенции:

СЛК-2. Быть способным к социальному взаимодействию;

СЛК-3. Обладать способностью к межличностным коммуникациям;

СЛК-5. Быть способным к критике и самокритике;

СЛК-6. Уметь работать в команде.

профессиональные компетенции:

ПК-8. Владеть современными средствами телекоммуникаций.

ПК-14. Работать с научной, нормативно-справочной и специальной литературой.

ПК-15. Заниматься аналитической и научно-исследовательской деятельностью в области информатики.

ПК-16. Владеть современными технологиями проектирования сложных систем и участвовать в разработке новых технологий.

ПК-17. Разрабатывать, анализировать и оптимизировать алгоритмы взаимодействия процессов в информационных средах.

ПК-18. Эксплуатировать и сопровождать программные системы.

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 (в том числе – 12 часов дистанционного обучения), управляемая самостоятельная работа – 4 часа дистанционного обучения.

Трудоемкость учебной дисциплины составляет 3 зачетные единицы.

Форма текущей аттестации – зачет.

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. Понижение размерности и метод главных компонент. 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.

Перцептрон. Функции активации. Многослойные сети прямого распространения. Метод обратного распространения ошибки. Стохастический градиентный спуск. Сверточные нейронные сети и глубокое обучение. Программные средства для обучения нейронных сетей.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

Дневная форма получения образования с применением электронных средств обучения (ДО)

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			4 4 (ДО)			
1.1	Введение	2			4 (ДО)			собеседование
1.2	Линейные методы машинного обучения	4			2			отчет
1.3	Прикладной анализ данных	2			2			контрольная работа
2	Алгоритмы обучения с учителем	12			8 4 (ДО)			
2.1	Оценивание вероятностей	4			2			собеседование
2.2	Метрические методы машинного обучения	2			2			отчет
2.3	Решающие деревья	2			4			коллоквиум
2.4	Композиции моделей	4			4 (ДО)			контрольная работа
3	Избранные главы машинного обучения	14			6 4 (ДО)		4 (ДО)	
3.1	Подготовка и предобработка данных	4			4 (ДО)			отчет
3.2	Рекомендательные системы	2			2			контрольная работа
3.3	Нейронные сети	8			4		4 (ДО)	проект

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Перечень основной 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. Николенко, С. Глубокое обучение. Погружение в мир нейронных сетей / С. Николенко, А. Кадури, Е. Архангельская. - Санкт-Петербург [и др.] : Питер, 2020. - 476 с.
3. Мэтиз, Эрик. Изучаем Python = Python Crash Course : программирование игр, визуализация данных, веб-приложения / Э. Мэтиз ; [перевел с англ. Е. Матвеев]. - 3-е изд. - Санкт-Петербург [и др.] : Питер, 2020. - 511 с.
4. Плас, Джейк Вандер. Python для сложных задач: наука о данных и машинное обучение / Дж. Вандер Плас ; [пер. с англ. И. Пальти]. - Санкт-Петербург [и др.]: Питер, 2018. - 573 с.

Перечень дополнительной литературы

1. Вьюгин В.В. Математические основы теории машинного обучения и прогнозирования. – М.: 2013. – 387 с.
2. Джеймс Г., Уиттон Д., Хасты Т., Тибширани Р. Введение в статистическое обучение с примерами на языке R. Изд. второе, испр. Пер. с англ. С. Э. Мастицкого – М.: ДМК Пресс, 2017. – 456 с.
3. Николенко С., Тулупьев А. Самообучающиеся системы. – М.: МЦНМО, 2009. – 288 с.
4. Жерон О. Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow: концепции, инструменты и техники для создания интеллектуальных систем / пер. с англ.– М.: Диалектика, 2018. – 688 с.
5. Силен, Дэви. Основы Data Science и Big Data. Python и наука о данных / Дэви Силен, Арно Мейсман, Мохамед Али ; [пер. с англ. Е. Матвеева]. - Санкт-Петербург [и др.] : Питер, 2017. - 334 с.
6. Christopher M. Bishop. Pattern Recognition and Machine Learning. – Springer, Information Science and Statistics series, 2006. – 738 pp.

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, отчеты по лабораторным работам.

3. Устно-письменная форма: отчеты по домашним упражнениям с их устной защитой, оценивание на основе проектного метода.

Формой текущей аттестации по дисциплине «Машинное обучение и анализ данных» учебным планом предусмотрен зачет.

При формировании итоговой оценки используется рейтинговая оценка знаний студента, дающая возможность проследить и оценить динамику процесса достижения целей обучения. Рейтинговая оценка предусматривает использование весовых коэффициентов для текущего контроля знаний студентов по дисциплине.

Примерные весовые коэффициенты, определяющие вклад текущего контроля знаний в рейтинговую оценку:

- отчет по лабораторной работе – 50 %;
- контрольные работы – 30 %;
- коллоквиум – 20 %.

Примерный перечень заданий для управляемой самостоятельной работы студентов

Управляемая самостоятельная работа предлагается в виде заданий проектного типа.

Тема 3.3. Нейронные сети. (4 ч/ДО)

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. Понижение размерности на примере набора данных MNIST.

Занятие 13. Рекомендательные системы.

Занятие 14. Знакомство с библиотеками Tensorflow и Keras.

Занятие 15-16. Обучение нейронных сетей для классификации данных из набора MNIST.

Занятие 17. Защита индивидуальных и групповых проектов.

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 зачёту, задания, вопросы для самоконтроля, список рекомендуемой литературы, информационных ресурсов и др.).

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

В отдельных случаях управляемая самостоятельная работа проводится в форме аудиторных занятий, согласно утвержденному графику.

Объем часов на составление и размещение заданий, консультации и контроль, осуществляемые с использованием технологий дистанционного обучения, планируется в пределах учебных часов, отведенных на УСР.

Примерный перечень вопросов к зачету

1. Постановки задач машинного обучения. Примеры задач. Виды данных. Признаки.
2. Метрики качества классификации и регрессии.
3. Линейная регрессия.
4. Градиентный спуск.
5. Методы регуляризации.
6. Добыча знаний. Задачи аналитики. Стандарт CRISP-DM.
7. Разведочный анализ данных и визуализация данных.
8. Наивный байесовский классификатор.
9. Логистическая регрессия и оценки вероятности классов.
10. Разложение ошибки на смещение и разброс.

- 11.Метод ближайших соседей. Метод парзеновского окна. Проблема размерности.
- 12.Дерево принятия решений. Критерии информативности. Жадный алгоритм построения.
- 13.Учёт пропусков в данных. Работа с категориальными признаками.
- 14.Бутстрап, бэггинг и случайный лес.
- 15.Градиентный бустинг над решающими деревьями и его реализации.
- 16.Отбор признаков. Цель и алгоритмы.
- 17.Понижение размерности. Метод главных компонент.
- 18.Постановка задачи и метрики качества построения рекомендаций.
- 19.Основные виды рекомендательных систем.
- 20.Матричные разложения для построения рекомендаций.
- 21.Перцептрон.
- 22.Функции активации. Многослойные сети прямого распространения.
- 23.Метод обратного распространения ошибки.
- 24.Стохастический градиентный спуск.
25. Сверточные нейронные сети.

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 № 23 от 24.06.2021 г.)
2. Алгоритмы в биоинформатике	Биомедицинской информатики	нет	Изменений не требуется (протокол № 23 от 24.06.2021 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 _____ 201_ г.)

Заведующий кафедрой

УТВЕРЖДАЮ
Декан факультета
