БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

УТВЕРЖДАЮ
Ректор Белорусского
государственного университета
_____А.Д.Король
15 июля 2024 г.
Регистрационный №УД- 13461/уч.

АНАЛИЗ ДАННЫХ И ОСНОВЫ МАШИННОГО ОБУЧЕНИЯ

Учебная программа учреждения образования
по учебной дисциплине для специальностей:

1-31 03 08 Математика и информационные технологии

1-31 03 01 Математика (по направлениям).

Направление специальности:
1-31 03 01-04 Математика (научно-конструкторская деятельность)

2024 г.

Учебная программа составлена на основе ОСВО 1-31 03 08-2021, ОСВО 1-31 03 01-2021 и учебных планов №G31-1-011уч., №G31-1-017уч., №G31-1-018уч. от 25.05.2021, №G31-1-001/уч.ин., №G31-1-003/уч.ин. №G31-1-004/уч.з., №G31-1-003/уч.з. от 31.05.2021.

СОСТАВИТЕЛИ:

А.В. Кушнеров, старший преподаватель кафедры дифференциальных уравнений и системного анализа механико-математического факультета Белорусского государственного университета

РЕЦЕНЗЕНТЫ:

А.А. Перхунов, Начальник отдела разработки программного обеспечения управления цифровизации РУП «Производственное объединение Белоруснефть».

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:

Кафедрой дифференциальных уравнений и системного анализа БГУ
(протокол № 12 от 25.04.2024)

Научно-методическим советом БГУ
(протокол № 9 от 28.06.2024)

Зав. кафедрой дифференциальных уравнений
и системного анализа Л. Л. Голубева

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Цели и задачи учебной дисциплины

Цель учебной дисциплины – изучение математической базы для анализа данных и применения основных алгоритмов машинного обучения.

Образовательная цель: обучение студентов приёмам построения моделей машинного обучения (МО), а также приёмам применения этих моделей на реальных данных различных типов.

Развивающая цель: освоение практических построения моделей на основе алгоритмов МО без использования встроенных возможностей современных языков программирования.

Задачи учебной дисциплины:

1. Освоение теоретической информации по основным алгоритмам МО.
2. Построение моделей эстиматоров для различных задач.
3. Приобретение навыков практической реализации алгоритмов машинного обучения.
4. Оценка качества моделей МО и применение их в задачах реальной жизни.

Место учебной дисциплины в системе подготовки специалиста с высшим образованием.

Для специальности 1-31 03 08 «Математика и информационные технологии» учебная дисциплина является дисциплиной **модуля** «Информационные технологии 2» компонента учреждения высшего образования.

Для специальности 1-31 03 01 «Математика» учебная дисциплина является дисциплиной по выбору **модуля** «Программирование и информационные технологии» компонента учреждения высшего образования.

Связи с другими учебными дисциплинами, включая учебные дисциплины компонента учреждения высшего образования, дисциплины специализации и др.

В ходе изучения дисциплины студенты могут воспользоваться знаниями, полученными при прохождении дисциплин «Методы программирования» и «математический анализ»

Требования к компетенциям

Освоение учебной дисциплины «Анализ данных и основы машинного обучения» должно обеспечить формирование следующих компетенций:

Для специальности 1-31 03 08 «Математика и информационные технологии»:

специализированной компетенции:

Осуществлять анализ контекста и поставленной проблемы, аргументированно выбирать оптимальный способ ее решения, согласовывать частичные проекты решения в общую согласованную архитектуру, выполнять реализацию проекта с учетом оценки накопленных и поступающих данных;

универсальной компетенции:

Владеть основами исследовательской деятельности, осуществлять поиск, анализ и синтез информации;

базовой профессиональной компетенции:

Применять современные технологии и базовые конструкции языков программирования для реализации алгоритмических прикладных задач и разработки веб-проектов.

Для специальности 1-31 03 01 «Математика»:

специализированной компетенции:

Осуществлять анализ контекста и поставленной проблемы, аргументированно выбирать оптимальный способ ее решения, согласовывать частичные проекты решения в общую согласованную архитектуру, выполнять реализацию проекта с учетом оценки накопленных и поступающих данных

базовых профессиональных компетенций:

– Применять современные компьютерные математические системы для проведения вычислительного (компьютерного) эксперимента;

– Применять основные понятия информатики, базовыми конструкциями языков программирования, технологиями объектно-ориентированного программирования для реализации алгоритмических прикладных задач и разработки веб-проектов;

– Применять инновационные информационные технологии и современные языки программирования.

В результате изучения учебной дисциплины студент должен:

знать:

– наиболее известные алгоритмы машинного обучения;

– критерии оценки качества моделей;

– инструменты для работы с алгоритмами МО в среде Python;

уметь:

– проектировать признаки, описывающие данные;

– создавать учебные программы моделирующие различные эстиматоры;

– осуществлять обоснованные подбор гиперпараметров моделей МО;

владеть:

– навыками прикладного анализа данных с помощью МО;

– методами постановки конкретных задач на основе технических требований.

Структура учебной дисциплины

Дисциплина изучается в 7 семестре очной формы и в 8 семестре заочной формы. Всего на изучение учебной дисциплины «Анализ данных и основы машинного обучения» отведено:

- для очной формы получения высшего образования: 108 часов, в том числе 72 аудиторных часа, из них: лекции – 36 часа, лабораторные занятия – 30 часов, управляемая самостоятельная работа – 6 часов;

- для заочной формы получения высшего образования: 108 часов, в том числе 16 аудиторных часов, из них: лекции – 8 часов, лабораторные занятия – 8 часов.

Трудоемкость учебной дисциплины составляет 3 зачетных единицы.

Для специальности 1-31 03 08 «Математика и информационные технологии»: форма промежуточной аттестации – экзамен.

Для специальности 1-31 03 01 «Математика»: форма промежуточной аттестации – зачёт.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Раздел 1. Основы машинного обучения.

Тема 1.1. Введение в машинное обучение.

Основные понятия МО. Обучение с учителем. Обучение без учителя. Тестовая и обучающая выборки. Понятие эстиматора. Метрики качества моделей. Простейшая линейная регрессия.

Тема 1.2. Предварительная обработка данных.

Работа с признаками. Нормализация данных и шкалирование. Поиск аномалий в данных. Числовые и категориальные признаки. Корреляция. Описательная статистика.

Раздел 2. Машинное обучение с учителем.

Тема 2.1. Алгоритмы регрессии.

Постановка задачи регрессии. Линейная регрессия. Полиномиальная регрессия. Графическая визуализация задачи регрессии. Проблема переобучения. Регуляризация. Lasso – регуляризация. Ridge – регуляризация.

Тема 2.2. Алгоритмы классификации.

Постановка задачи классификации. Простейшие алгоритмы классификации: дерево решений, метод ближайших соседей. Оценка качества модели классификации. Гиперпараметры моделей. Подбор гиперпараметров. Валидационные кривые и кривые обучения.

Тема 2.3. Линейный классификатор.

Линейная классификация. Разделяющая гиперплоскость. Логистическая регрессия. L2-регуляризация логистической функции потерь.

Тема 2.4. Бэггинг.

Ансамбли оценивателей. Бутстреп. Бэггинг. Случайные леса и подбор гиперпараметров для них. Сверхслучайные леса.

Тема 2.5. Бустинг и стэкинг.

Понятие стекинга и градиентного бустинга. Модели классификатора и регрессора.

Тема 2.6. Оценка качества модели.

Метрики качества модели. Кросс-валидация. Дисперсия и отклонение в задачах МО. Проблема переобучения. Выбор оптимальной сложности модели оценивателя. Преобразование признаков. Нормализация и изменения распределения.

Раздел 3. Машинное обучение без учителя. Временные ряды.

Тема 3.1. Анализ временных рядов.

Временной ряд в МО. Rolling window estimations. Экспоненциальное сглаживание, модель Хольта-Винтерса. Кросс-валидация на временных рядах, подбор параметров. Извлечение признаков (Feature extraction). Линейная регрессия и XGBoost.

Тема 3.2. Клатеризация.

Понятие кластеризации. Алгоритм к-средних. Агломеративная кластеризация. Алгоритм DBSCAN. Оценка качества кластеризации.

Тема 3.3. Понижение размерности.

Метод главных компонент. Оценка количества компонент с учётом объяснимой дисперсии. Графическая интерпретация метода главных компонент. Алгоритм TSNE.

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**							
1.1	Введение в машинное обучение	4			4			Отчет по лабораторной работе с устной защитой
1.2	Предварительная обработка данных.	2			2			Отчет по лабораторной работе с устной защитой.
2	**Машинное обучение с учителем**							
2.1	Алгоритмы регрессии.	4			4			Отчет по лабораторной работе с устной защитой.
2.2	Алгоритмы классификации.	4			2		2	Отчет по лабораторной работе с устной защитой. Контрольная работа по темам 1.1-2.2.
2.3	Линейный классификатор	2			2			Отчет по лабораторной работе с устной защитой

2.4	Бэггинг	4			4			Отчет по лабораторной работе с устной защитой.
2.5	Бустинг и стэкинг.	4			2			Отчет по лабораторной работе с устной защитой.
2.6	Оценка качества модели.	2			2		2	Отчет по лабораторной работе с устной защитой. Контрольная работа.
3	**Машинное обучение без учителя. Временные ряды.**							
3.1	Анализ временных рядов	4			4			Отчет по лабораторной работе с устной защитой
3.2	Клатеризация	4			2			Отчет по лабораторной работе с устной защитой.
3.3	Понижение размерности	2			2		2	Отчет по лабораторной работе с устной защитой. Контрольная работа.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

Заочная форма получения высшего образования

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**							
1.1	Введение в машинное обучение	2			2			Отчет по лабораторной работе с устной защитой
1.2	Предварительная обработка данных.	2			2			Отчет по лабораторной работе с устной защитой.
2	**Машинное обучение с учителем**							
2.1	Алгоритмы регрессии.	2			2			Отчет по лабораторной работе с устной защитой.
2.2	Алгоритмы классификации.	2			2			Отчет по лабораторной работе с устной защитой.

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Основная литература

1. Плас, Д. В. Python для сложных задач: наука о данных и машинное обучение. - Санкт-Петербург ; Москва ; Минск : Питер, 2023. - 573 с. - URL: https://ibooks.ru/bookshelf/376830/reading.
2. Хливненко, Л. В. Практика нейросетевого моделирования : учебное пособие [для вузов] / Л. В. Хливненко, Ф. А. Пятакович. - Изд. 3-е, стер. - Санкт-Петербург ; Москва ; Краснодар : Лань, 2023. - 196 с. - URL: https://e.lanbook.com/book/310190.
3. Ростовцев, В. С. Искусственные нейронные сети : учебник / В. С. Ростовцев. - Изд. 2-е, стер. - Санкт-Петербург ; Москва ; Краснодар : Лань, 2021. - 213 с. - URL: https://reader.lanbook.com/book/310184#1.
4. 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.

Дополнительная литература

1. Прикладная статистика. Классификация и снижение размерности : Справ. изд. / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин; под ред. С. А. Айвазяна. - Москва : Финансы и статистика, 1989. - 606, [1] с.
2. Айвазян, С. А. Прикладная статистика [Текст] = Applied statistics : исследование зависимостей : справочное издание / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин ; под ред. С. А. Айвазяна. - Москва : Финансы и статистика, 1985. - 487 с.
3. Воронцов К. В. Математические методы обучения по прецедентам. 2005–2010. - URL: http://www.machinelearning.ru/wiki, страница «Машинное обучение (курс лекций, К.В.Воронцов).
4. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. Springer. 2008.
5. Себастьян, Р. Python и машинное обучение / Рашка Себастьян. - М.: ДМК Пресс, 2017. - 614 с.
6. Домингос, П. Верховный алгоритм: как машинное обучение изменит наш мир / Педро Домингос. - Москва: РГГУ, 2015. - 447 с.
7. Андреас, М. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными / Мюллер Андреас. - М.: Альфа-книга, 2017. - 487 с.

Рекомендуемое учебно-лабораторное оборудование

Для проведения занятий требуется следующее программное обеспечение: пакет *Mathematica*, Python.

Перечень рекомендуемых средств диагностики и методика формирования итоговой отметки

Объектом диагностики компетенций студентов являются знания, умения, полученные ими в результате изучения учебной дисциплины.

Для диагностики компетенций могут использоваться следующие средства текущего контроля: отчет по лабораторной работе с устной защитой и контрольная работа.

Отметка текущего контроля знаний студента по дисциплине «Анализ данных и основы машинного обучения» формируется в результате регулярной и систематической проверки знаний студентов во время лабораторных занятий и по итогам их самостоятельной работы. Текущий контроль знаний проходит во время устной защиты отчёта по лабораторным работам, выполняемым в учебной лаборатории и самостоятельно вне аудитории.

При защите лабораторных работ оценивается полнота ответа, аргументация выбранных решений, последовательность и оригинальность изложения материала, оригинальность кода, корректность оформления, самостоятельность выполнения заданий. Предполагается постановка дополнительных практических и теоретических задач во время отчёта по лабораторной работе.

Формой промежуточной аттестации по дисциплине «Анализ данных и основы машинного обучения» учебным планом предусмотрен **зачёт** для специальности 1-31 03 01 «Математика» и **экзамен** для специальности 1-31 03 08 «Математика и информационные технологии».

Экзамен и зачёт по дисциплине проходит в форме контрольного опроса в устной или письменной форме.

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):

- отчет по лабораторной работе с устной защитой – 60 %;
- контрольная работа – 40 %.

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

Примерный перечень заданий для управляемой самостоятельной работы студентов

Темы 2.2. Алгоритмы классификации. (2 ч)

Примерный перечень заданий:

1. Постройте модель эстиматора для классификации ос и пчёл. https://www.kaggle.com/datasets/stealthtechnologies/classification-between-a-bee-and-a-wasp
2. Постройте модель эстиматора для классификации ос и пчёл, предварительно понизив размерность данных. https://www.kaggle.com/datasets/stealthtechnologies/classification-between-a-bee-and-a-wasp

Форма контроля – **контрольная работа.**

Тема 2.6. Оценка качества модели. (2 ч.)

Примерный перечень заданий:

1. Проведите предварительную обработку данных.
2. Постройте модели классификации на основе различных методов, изученных вами из встроеной библиотеки.
3. Подберите оптимальные гиперпараметры моделей используя различные оценки, кросс-валидацию и валидационные кривые.
4. Сделайте выводы о точности моделей. Выберите самую оптимальную. Тщательно поясните свой выбор!
5. Выберите самую оптимальную на ваш взгляд модель (Использовать встроенную и собственную реализацию моделей и сравнить результат)

Форма контроля – **контрольная работа.**

Тема 3.3. Понижение размерности. (2 ч.)

Примерный перечень заданий:

Реализуйте функцию для визуализации обучения модели DBSCAN на двумерных данных полученных с помощью метода главных компонент со следующими условиями:

1. Подсветить корневые точки.
2. Показать итеративное формирование каждого кластера, подсвечивая текущую корневую точку и её окрестность.
3. Текстом вывести на экран количество точек в каждом кластере.

Выполнение заданий на основе методических указаний к лабораторным занятиям.

Форма контроля – **контрольная работа.**

Описание инновационных подходов и методов к преподаванию учебной дисциплины

При организации образовательного процесса используется ***практико-ориентированный подход,*** который предполагает освоение содержания через решения практических задач.

При организации образовательного процесса ***используются методы и приемы развития критического мышления,*** которые представляют собой систему, формирующую навыки работы с информацией в процессе чтения и письма; понимании информации как отправного, а не конечного пункта критического мышления.

Методические рекомендации по организации самостоятельной работы обучающихся

Для организации самостоятельной работы студентов по учебной дисциплине рекомендовано разместить на образовательном портале БГУ курсы лекций и лабораторные практикумы. Также следует разместить перечень вопросов к зачёту.

Самостоятельная работа студента включает в себя работу с учебной литературой по заданным разделам дисциплины, поиск в Интернете новейшей учебной и научной информации в указанных областях знаний и знакомство с ней, а также выполнение задач, поставленных на занятиях.

Возможна организация текущих консультаций в формате видеоконференции.

Примерный перечень вопросов к экзамену и зачёту

1. Введение в машинное обучение. Основные понятия МО. Обучение с учителем. Обучение без учителя.
2. Тестовая и обучающая выборки. Понятие эстиматора. Метрики качества моделей.
3. Python библиотеки numpy, matpotlib, sklearn, seaborn. Краткий обзор функционала, применение в контексте МО.
4. Работа с признаками. Нормализация данных и шкалирование. Поиск аномалий в данных.
5. Числовые и категориальные признаки. Корреляция. Описательная статистика.
6. Алгоритмы регрессии. Постановка задачи регрессии. Линейная регрессия.
7. Полиномиальная регрессия. Регрессия с использованием гауссовских базисных функций.
8. Линейная регрессия по произвольному базису. Графическая визуализация задачи регрессии.
9. Проблема переобучения. Регуляризация. Ridge – регуляризация.

10. Проблема переобучения. Регуляризация. Lasso – регуляризация.
11. Алгоритмы классификации. Постановка задачи классификации.
12. Простейшие алгоритмы классификации: дерево решений.
13. Простейшие алгоритмы классификации: метод ближайших соседей.
14. Гиперпараметры моделей. Подбор гиперпараметров. Валидационные кривые и кривые обучения.
15. Линейный классификатор. Линейная классификация. Разделяющая гиперплоскость. Логистическая регрессия.
16. L2-регуляризация логистической функции потерь.
17. Бэггинг. Ансамбли оценивателей. Бутстреп.
18. Случайные леса и подбор гиперпараметров для них. Сверхслучайные леса.
19. Оценка качества модели. Метрики качества модели. Кросс-валидация. Дисперсия и отклонение в задачах МО.
20. Проблема переобучения. Выбор оптимальной сложности модели оценивателя.
21. Преобразование признаков. Удаление аномальных данных.
22. Обучение без учителя. Понижение размерности. Метод главных компонент.
23. Обучение без учителя. Понижение размерности. Метод TSNE.
24. Обучение без учителя. Кластеризация. Алгоритм k-средних.
25. Метрики оценки качества кластеризации.
26. Обучение без учителя. Кластеризация. Алгоритм DBSACN. Вычислительная сложность.
27. Визуализация в задача МО. Библиотека matplotlib. Библиотека seaborn.

ПРОТОКОЛ СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ УО

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

Зав. кафедрой дифференциальных уравнений
и системного анализа Л. Л. Голубева

25.04.2024

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.)
(название кафедры)

Заведующий кафедрой

_____ _____
(ученая степень, ученое звание) (И.О.Фамилия)

УТВЕРЖДАЮ
Декан факультета

_____ _____
(ученая степень, ученое звание) (И.О.Фамилия)