БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

УТВЕРЖДАЮ
Ректор Белорусского
государственного университета
_____А.Д.Король
10 июня 2024
Регистрационный № УД- 13338/уч.

ОСНОВЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Учебная программа учреждения образования
по учебной дисциплине для специальностей:

1-31 03 04 Информатика
1-31 03 07 Прикладная информатика (по направлениям):
1-31 03 07-01 Прикладная информатика (программное обеспечение компьютерных систем)

2024 г.

Учебная программа составлена на основе ОСВО 1-31 03 04-2021, ОСВО 1-31 03 07-2021 и учебных планов № G31-1-031/уч. от 30.06.2021, № G31-1-034/уч. от 23.07.2021.

СОСТАВИТЕЛИ:

Т.В. Цеховая, доцент кафедры теории вероятностей и математической статистики факультета прикладной математики и информатики Белорусского государственного университета, кандидат физико-математических наук, доцент;

П.А. Пашук, старший преподаватель кафедры теории вероятностей и математической статистики факультета прикладной математики и информатики Белорусского государственного университета, магистр прикладной математики и информационных технологий;

И.Ю. Шевко, старший преподаватель кафедры теории вероятностей и математической статистики факультета прикладной математики и информатики Белорусского государственного университета, магистр;

Т.Д. Полузёров, ассистент кафедры теории вероятностей и математической статистики факультета прикладной математики и информатики Белорусского государственного университета.

РЕЦЕНЗЕНТЫ:

М.С. Абрамович, заведующий НИЛ статистического анализа и моделирования учреждения Белорусского государственного университета «НИИ прикладных проблем математики и информатики», кандидат физико-математических наук, доцент;

А.И. Кишкар, ведущий инженер-программист отдела информационных систем управления бизнес-приложений департамента производства ЗАО «Международный деловой альянс», магистр прикладной математики и информационных технологий.

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:

Кафедрой теории вероятностей и математической статистики БГУ
(протокол № 12 от 21.05.2024);

Научно-методическим советом БГУ
(протокол № 8 от 31.05.2024)

Заведующий кафедрой теории вероятностей
и математической статистики А.Ю.Харин

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Цели и задачи учебной дисциплины

Цель учебной дисциплины «Основы интеллектуального анализа данных» – формирование необходимых навыков владения методами и алгоритмами исследования зависимостей между различными типами данных и дальнейшего использования знаний об этих зависимостях в задачах визуализации, систематизации, анализа и прогнозирования.

Задачи учебной дисциплины:

1) изучение основных понятий, алгоритмов и методов работы для интеллектуального анализа данных;

2) формирование навыков практического применения изученных методов и интерпретации получаемых результатов.

Место учебной дисциплины в системе подготовки специалиста с высшим образованием.

Учебная дисциплина относится к компоненту учреждения высшего образования.

Учебная программа составлена с учетом межпредметных **связей** и программ по дисциплине «Теория вероятностей и математическая статистика».

Требования к компетенциям

Освоение учебной дисциплины «Основы интеллектуального анализа данных» должно обеспечить формирование следующих компетенций:

специализированные компетенции:

СК. Организовывать хранение больших данных и выполнять их анализ, определять подходящий инструмент анализа больших данных.

В результате освоения учебной дисциплины студент должен:

знать:

- основы математической и описательной статистики;
- методы и алгоритмы для анализа, обработки и систематизирования данных;

уметь:

- подбирать необходимую статистические модель или алгоритм для решения конкретной задачи;
- исследовать эффективность применения статистического метода для решения поставленной задачи;

владеть:

- основными методами предобработки и предварительного анализа данных;
- основными методами визуализации данных;
- основными статистическими методами принятия решений;
- навыками компьютерной реализации основных методов анализа данных.

Структура учебной дисциплины

Дисциплина изучается в 7 семестре. Всего на изучение учебной дисциплины «Основы интеллектуального анализа данных» отведено:

– в очной форме получения высшего образования: 100 часов, в том числе 64 аудиторных часов, из них: лекции – 32 часа, лабораторные занятия – 30 часов, управляемая самостоятельная работа (УСР) – 2 часа.

Трудоемкость учебной дисциплины составляет 3 зачетные единицы.

Форма промежуточной аттестации – зачет.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Раздел 1. Описательная статистика.

Тема 1.1. Введение в описательную статистику.

Предмет и основные категории дисциплины. Основные этапы статистического исследования. Формы организации и виды статистического наблюдения. Требования к собираемым данным. Способы отбора выборки. Точность статистического наблюдения.

Тема 1.2. Одномерные статистические признаки.

Выборка. Вариационный ряд. Дискретные и интервальные распределения частот. Графическое изображение статистических данных: полигон частот, гистограмма, кумулятивная кривая, огива, диаграммы.

Числовые характеристики одномерных статистических признаков. Основное балансовое тождество. Степенные средние. Средние величины и их свойства. Теорема о фундаментальном неравенстве средних.

Числовые характеристики положения. Структурные характеристики вариационного ряда: мода; медиана; квантили. Примеры практического применения характеристик положения.

Числовые характеристики рассеяния вариационного ряда. Дисперсия и ее свойства. Виды дисперсий сгруппированной выборки. Примеры применения характеристик рассеяния в прикладных исследованиях.

Числовые характеристики формы частотного распределения и их практическое значение. Примеры.

Тема 1.3. Статистическое изучение взаимосвязи между признаками.

Методы выявления наличия зависимости между двумя признаками. Корреляционная таблица. Маргинальные частотные распределения количественных признаков. Условные частотные распределения количественных признаков. Показатели тесноты связи между двумя признаками. Линейный коэффициент корреляции Пирсона и его свойства. Коэффициент Фехнера. Коэффициенты корреляции рангов (Спирмена, Кендэла). Коэффициент конкордации. Коэффициент ассоциации. Коэффициент контингенции. Коэффициент взаимной сопряженности. Методы вычисления коэффициента корреляции признаков, измеренных в различных шкалах.

Уравнение регрессии. Понятие регрессии. Метод наименьших квадратов. Теоретическое корреляционное отношение. Оценка существенности коэффициента регрессии и уравнения связи.

Множественная корреляция. Многофакторные регрессионные модели. Совокупный коэффициент детерминации. Множественный коэффициент корреляции. Частный коэффициент корреляции. Проверка адекватности модели и правильности выбора формы связи.

Раздел 2. Интеллектуальный анализ данных.

Тема 2.1. Основные понятия интеллектуального анализа данных.

Цели анализа данных. Основные типы задач: «обучение с учителем» и «обучение без учителя». Общий алгоритм решения задач, CRISP-DM. Признаковое описание объекта, типы признаков. Понятия модели, метода обучения, функционала качества. Обучающая способность: переобучение и недообучение. Эмпирические меры обобщающей способности. Примеры.

Тема 2.2. Восстановление регрессии

Постановка задачи регрессии. Линейная регрессия. Нелинейная регрессия. Метод градиентного спуска, стохастический градиент. Регуляризация. Обобщение на неквадратичные функции потерь.

Тема 2.3. Байесовские методы классификации

Вероятностная постановка задачи классификации. Оптимальный байесовский классификатор. Наивный байесовский классификатор. Восстановление одномерных и многомерных плотностей распределений. Квадратичный дискриминант. Линейный дискриминант Фишера.

Тема 2.4. Линейные методы классификации.

Дискриминативная постановка задачи классификации. Понятие отступа. Метод опорных векторов. Оценивание вероятностей классов. Логистическая регрессия. Многоклассовая классификация.

Тема 2.5. Метрические методы классификации и регрессии.

Гипотезы непрерывности и компактности. Мера расстояния между объектами. Метод k-ближайших соседей. Метод окна Парзена. Ядерное сглаживание. Непараметрическая регрессия.

Тема 2.6. Деревья решений.

Определение дерева решений. Жадный алгоритм построения дерева. Критерии ветвления, остановки. Регуляризация деревьев. Методы работы с категориальными данными и пропусками.

Тема 2.7. Композиции алгоритмов.

Определение композиции алгоритмов. Разложение ошибки на смещение и разброс. Бутстрэп и метод случайных подпространств. Бэггинг. Случайный лес. Градиентный бустинг.

Тема 2.8. Кластеризация.

Постановка задачи кластеризации. Меры качества кластеризации. Алгоритм k-средних. Алгоритм DBSCAN. Меры расстояний между кластерами. Иерархическая кластеризация. Задачи частичного обучения. Обобщение методов кластеризации для задач частичного обучения.

Тема 2.9. Поиск ассоциативных правил.

Задача поиска ассоциативных правил. Понятие поддержки и значимости правила. Алгоритмы APRIORI, FP-Growth.

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	Формы контроля знаний
		Лекции	Практические занятия	Семинарские занятия	Лабораторные занятия	Иное		
I	**Описательная статистика**	**14**			**14**			
1.1	**Введение в описательную статистику**	**1**			**1**			Устный опрос
1.2	**Одномерные статистические признаки**	**7**			**7**			
1.2.1	Вариационный ряд. Графическое изображение статистических данных.	1			1			Отчет по лабораторной работе
1.2.2	Числовые характеристики одномерных статистических признаков. Числовые характеристики положения.	2			2			Индивидуальные задания
1.2.3	Числовые характеристики рассеяния вариационного ряда.	2			2			Устный опрос, отчет по лабораторной работе
1.2.4	Числовые характеристики формы частотного распределения.	2			2			Контрольная работа.
1.3	**Статистическое изучение взаимосвязи между признаками**	**6**			**6**			
1.3.1	Показатели тесноты связи между двумя признаками.	2			2			Индивидуальные задания
1.3.2	Уравнение регрессии.	2			2			Отчет по лабораторной работе

1.3.3	Множественная корреляция.	2			2			Контрольная работа.
II	**Интеллектуальный анализ данных**	**18**			**16**		**2**	
2.1	Основные понятия интеллектуального анализа данных	2			2			Устный опрос
2.2	Восстановление регрессии	2					2	Отчет по лабораторной работе
2.3	Байесовские методы классификации	2			2			Индивидуальные задания
2.4	Линейные методы классификации	2			2			Контрольная работа.
2.5	Метрические методы классификации и регрессии	2			2			Отчет по лабораторной работе
2.6	Деревья решений.	2			2			Отчет по лабораторной работе
2.7	Композиции алгоритмов	2			2			Устный опрос, индивидуальные задания
2.8	Кластеризация	2			2			Контрольная работа.
2.9	Поиск ассоциативных правил	2			2			Отчет по лабораторной работе.
ИТОГО		**32**			**30**		**2**	

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Основная литература

1. Жукова, А. А. Биометрия : пособие для студ. учрежд. высш. образ. : в 3 ч. / А. А. Жукова, М. Л. Минец ; БГУ. - Минск : БГУ, 2019 - Ч. 2 : Основные техники анализа данных. - 2020. - 151 с.
2. Жукова, А. А. Биометрия : пособие для студ. учрежд. высш. образ. : в 3 ч. / А. А. Жукова, М. Л. Минец ; БГУ. - Минск : БГУ, 2019 - Ч. 3 : Корреляция и регрессия. - 2021. - 103 с.
3. Труш, Н. Н. Введение в компьютерный и интеллектуальный анализ данных: учеб. материалы / Н.Н. Труш. - Минск: БГУ, 2022 - 69 с. - https://elib.bsu.by/handle/123456789/277034.
4. Яцков, Н. Н. Интеллектуальный анализ данных : электронный учебно-методический комплекс / Н. Н. Яцков ; БГУ, Фак. радиофизики и компьютерных технологий , Каф. системного анализа и компьютерного моделирования. - Минск : БГУ, 2023. - URL: https://elib.bsu.by/handle/123456789/306842.

Дополнительная литература

1. Лагутин, Б.М. Наглядная математическая статистика: учеб. пособие / Б.М. Лагутин. – М.: БИНОМ. Лаборатория знаний, 2023. – 475 с.
2. Ефимова, М.Р. Общая теория статистики. Практический курс : учебное пособие для вузов / М. Р. Ефимова, Е. В. Петрова, О. И. Ганченко, М. А. Михайлов ; под редакцией М. Р. Ефимовой. – М. : Издательство Юрайт, 2024. – 323 с.
3. Харин, Ю.С. Теория вероятностей, математическая и прикладная статистика: ученик / Ю.С. Харин, Н.М. Зуев, Е.Е. Жук. – Минск: БГУ, 2011 – 464 с.
4. Харин, Ю.С. Математическая и прикладная статистика: учеб. пособие / Ю.С. Харин, Е.Е. Жук – Минск: БГУ, 2005. – 279 с.
5. Волкова, П.А. Статистическая обработка данных в учебно-исследовательских работах: учебное пособие / П.А. Волкова, А.Б. Шипунов. – М.: Форум, 2012. – 96 с.
6. Елисеева И.И. Общая теория статистики / И.И. Елисеева, М.М. Юзбашев – М.: Финансы и статистика, 2004.– 656 с.
7. Шмойлова Р.А. Теория статистики / Р.А. Шмойлова, В.Г. Минашкин, Н.А. Садовникова и др. М.: Финансы и статистика, 2007. – 656 с.
8. Ефимова М.Р. Общая теория статистики / М.Р. Ефимова, Е.В. Петрова, В.Н. Румянцев – М: НИЦ ИНФРА-М, 2011. – 416 с.
9. Айвазян С.А. Прикладная статистика и основы эконометрики / С.А. Айвазян, В.С. Мхитарян - М.: Изд. объедин. «ЮНИТИ», 1998. – 1022 с.
10. Жукова, А.А. Биометрия. В 3 ч. Ч. 1. Описательная статистика: пособие / А.А. Жукова, М.Л. Минец. – Минск: БГУ, 2019. – 100 с.

11. Джеймс, Г. Введение в статистическое обучение с примерами на языке R. Пер. с англ. С.Э. Мастицкого / Г. Джеймс, Д. Уиттон, Т. Хасти, Р. Тибширани. – М.: ДМК Пресс, 2016. – 450 с.

12. Буяльская, Ю.В. Введение в компьютерный и интеллектуальный анализ данных: метод. указания / Ю.В. Буяльская, В.В. Казаченок – Минск: БГУ, 2016. – 46 с.

13. Харин, А. Статистическое моделирование / А. Харин, К. Кроукс, Ш. ван Аэльст, П. Фильцмозер, М. Хьюберт – Минск: БГУ, 2017 – 151 с.

14. Степанов, Р.Г. Технология DATA MINING: Интеллектуальный анализ данных / Р.Г. Степанов – Казань, Издательство Казанского госуниверситета, 2008 – 58 с.

15. Мусаев, А.А. Интеллектуальный анализ данных: учебное пособие / А.А. Мусаев. – СПб.: СПбГТИ(ТУ), 2018. – 176 с.

16. Рафалович, В. Data mining, или Интеллектуальный анализ данных для занятых. Практический курс / В. Рафалович – Москва: Литагент И-Трейд, 2014. – 96 с.

17. Маунт, Дж. Погружение в аналитику данных: пер. с англ. / Дж. Маунт – СПб.: БХБ-Петербург, 2023. – 224 с.: ил.

18. Петрунин, Ю.Ю Информационные технологии анализа данных. Data analysis. Изд.4 / Ю.Ю. Петрунин – М.: Издательство МГУ, 2023. – 296 с.

19. Зарова, Е.В. Методы Data mining в обработке и анализе статистических данных (решения в R) / Е.В. Зарова – М.: Инфра-М, 2021. – 232 с.

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 текущего контроля и промежуточной аттестации.

Для диагностики компетенций могут использоваться следующие средства текущего контроля: устный опрос, контрольные работы, отчеты по лабораторным работам, индивидуальные задания.

Формой промежуточной аттестации по дисциплине «Основы интеллектуального анализа данных» учебным планом предусмотрен **зачет**.

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):

- контрольные работы – 40%;
- защита отчетов по лабораторным работам – 60%.

Зачет по дисциплине проходит в устной и(или) письменной форме. В случае успешного прохождения всех форм текущего контроля и получения положительной (4 и выше) отметки текущей аттестации, допускается получение зачета без проведения дополнительного опроса. При этом явка обучающегося на зачет является обязательной.

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

Примерный перечень заданий для управляемой самостоятельной работы студентов

Тема 2.2. Восстановление регрессии. (2 ч.)

Линейная регрессия. Нелинейная регрессия. Метод градиентного спуска, стохастический градиент. Регуляризация. Обобщение на неквадратичные функции потерь.

Форма контроля – защита отчета по лабораторной работе.

Примерная тематика лабораторных занятий

Занятие № 1. Введение в описательную статистику. Графическое изображение статистических данных.

Занятие № 2. Числовые характеристики одномерных статистических признаков. Числовые характеристики положения.

Занятие № 3. Числовые характеристики рассеяния вариационного ряда.

Занятие № 4. Числовые характеристики формы частотного распределения.

Занятие № 5. Показатели тесноты связи между двумя признаками.

Занятие № 6. Уравнение регрессии.

Занятие № 7. Множественная корреляция.

Занятие № 8. Основные понятия интеллектуального анализа данных.

Занятие № 9. Байесовские методы классификации.

Занятие № 10. Линейные методы классификации.

Занятие № 11. Метрические методы классификации и регрессии.

Занятие № 12. Деревья решений.

Занятие № 13. Композиции алгоритмов.

Занятие № 14. Кластеризация.

Занятие № 15. Поиск ассоциативных правил.

Описание инновационных подходов и методов к преподаванию учебной дисциплины

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

– освоение содержание образования через решения практических задач;

– приобретение навыков эффективного выполнения разных видов профессиональной деятельности;

– ориентацию на генерирование идей, реализацию групповых студенческих проектов, развитие предпринимательской культуры;

– использованию процедур, способов оценивания, фиксирующих сформированность профессиональных компетенций.

Методические рекомендации по организации самостоятельной работы обучающихся

Самостоятельная работа с целью изучения материала учебной дисциплины предполагает работу с рекомендованной учебной литературой и Интернет-ресурсами. Теоретические сведения закрепляются выполнением лабораторных заданий, при выполнении которых следует руководствоваться методическими разработками, размещенными в электронной библиотеке университета и на образовательном портале. Также могут быть предложены дополнительные задания (тесты, задания для самостоятельного выполнения) для самооценивания и более глубокого усвоения полученного материала.

Примерный перечень вопросов к зачету

1. Основные этапы статистического исследования. Формы организации и виды статистического наблюдения. Требования к собираемым данным. Способы отбора выборки. Точность статистического наблюдения.
2. Выборка. Вариационный ряд. Дискретные и интервальные распределения частот. Графическое изображение статистических данных: полигон частот, гистограмма, кумулятивная кривая, огива. Диаграммы.
3. Числовые характеристики одномерных статистических признаков. Основное балансовое тождество. Степенные средние. Средние величины и их свойства. Теорема о фундаментальном неравенстве средних.
4. Числовые характеристики положения. Структурные характеристики вариационного ряда: мода; медиана; квантили. Примеры практического применения характеристик положения.
5. Числовые характеристики рассеяния вариационного ряда. Дисперсия и ее свойства. Виды дисперсий сгруппированной выборки. Примеры применения характеристик рассеяния в прикладных исследованиях.
6. Числовые характеристики формы частотного распределения и их практическое значение. Примеры.

7. Методы выявления наличия зависимости между двумя признаками. Корреляционное поле (диаграмма рассеяния). Корреляционная таблица. Маргинальные частотные распределения количественных признаков. Показатели тесноты связи между двумя признаками. Шкалы измерений. Примеры.

8. Измерение взаимной зависимости, когда оба признака заданы в количественной шкале. Свойства парного линейного коэффициента корреляции Пирсона. Коэффициент корреляции Фехнера.

9. Измерение взаимной зависимости, когда оба признака заданы в порядковой шкале.

10. Измерение взаимной зависимости, когда оба признака заданы в номинальной шкале.

11. Уравнение регрессии. Понятие регрессии. Корреляционное отношение. Оценка существенности коэффициента регрессии и уравнения связи. Коэффициент детерминации. Процесс линеаризации.

12. Множественная корреляция. Многофакторные регрессионные модели. Совокупный коэффициент детерминации. Множественный коэффициент корреляции. Частный коэффициент корреляции. Проверка адекватности модели.

13. Метод градиентного спуска, стохастический градиент.

14. Регуляризация регрессии. Обобщение на неквадратичные функции потерь.

15. Вероятностная постановка задачи классификации. Оптимальный байесовский классификатор.

16. Наивный байесовский классификатор. Восстановление одномерных и многомерных плотностей распределений.

17. Квадратичный дискриминант. Линейный дискриминант Фишера.

18. Метод опорных векторов.

19. Оценивание вероятностей классов. Логистическая регрессия.

20. Гипотезы непрерывности и компактности. Мера расстояния между объектами.

21. Метод k-ближайших соседей. Метод окна Парзена.

22. Ядерное сглаживание. Непараметрическая регрессия.

23. Определение дерева решений. Жадный алгоритм построения дерева.

24. Критерии ветвления, остановки. Регуляризация деревьев.

25. Определение композиции алгоритмов. Разложение ошибки на смещение и разброс.

26. Бутстрэп и метод случайных подпространств. Бэггинг.

27. Случайный лес. Градиентный бустинг.

28. Постановка задачи кластеризации. Меры качества кластеризации. Алгоритм k-средних.

29. Алгоритм DBSCAN.

30. Меры расстояний между кластерами. Иерархическая кластеризация.

31. Задача поиска ассоциативных правил. Понятие поддержки и значимости правила.

32. Алгоритмы APRIORI, FP-Growth.

ПРОТОКОЛ 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 «Основы интеллектуального анализа данных» не требует согласования			

Заведующий кафедрой теории
вероятностей и математической статистики,
доктор физ.-мат. наук, профессор А.Ю. Харин

21. мая 2024 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
_____ _____ _____