Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«Владимирский государственный университет
имени Александра Григорьевича и Николая Григорьевича Столетовых»
(ВлГУ)**

Институт информационных технологий и радиоэлектроники

УТВЕРЖДАЮ:
Директор института
_____ Галкин А.А.
«01» июня 2023 г

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Методы анализа данных
(наименование дисциплины)

направление подготовки / специальность

10.05.04 «Информационно-аналитические системы безопасности»
(код и наименование направления подготовки (специальности)

направленность (профиль) подготовки

Автоматизация информационно-аналитической деятельности
(направленность (профиль) подготовки))

г. Владимир

2023 год

1. ЦЕЛИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Целями освоения дисциплины «Методы анализа данных» являются обеспечение подготовки студентов в соответствии с требованиями ФГОС ВО и учебного плана специальности 10.05.04 «Информационно-аналитические системы безопасности» ознакомление студентов с кругом задач в области автоматической обработки естественного языка (natural language processing) и компьютерной лингвистики (computational linguistics), а также с доступным программным инструментарием для решения прикладных задач обработки текста. В рамках курса рассматриваются основные понятия компьютерной лингвистики, а также существующее программное обеспечение для работы с текстами.

Задачами курса является изучение следующих вопросов:

- Изучение базовых алгоритмов анализа и интерпретации данных.
- Формирование практических навыков работы с современными пакетами прикладных программ для решения задач анализа и интерпретации данных.

При изучении курса студенты знакомятся с современными формальными методами, реализующими «восходящую» стратегию анализа: извлечение интерпретируемых зависимостей из эмпирических данных. Методы рассматриваются в рамках парадигмы интеллектуального анализа данных (data mining, knowledge discovery), являющейся важнейшим направлением современных исследований в области анализа гетерогенных данных с нечисловыми параметрами.

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОПОП

Дисциплина «Методы анализа данных» относится к обязательной части образовательной программы, код Б1.О.11 специальности 10.05.04 «Информационно-аналитические системы безопасности»

В учебном плане предусмотрены виды учебной деятельности, обеспечивающие синтез теоретических лекций, лабораторных работ и самостоятельной работы студентов. Курс тесно взаимосвязан с другими дисциплинами данного цикла.

3. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ

Планируемые результаты обучения по дисциплине, соотнесенные с планируемыми результатами освоения ОПОП (компетенциями и индикаторами достижения компетенций)

Формируемые компетенции (код, содержание компетенции)	Планируемые результаты обучения по дисциплине, в соответствии с индикатором достижения компетенции		Наименование оценочного средства
	Индикатор достижения компетенции *(код, содержание индикатор)а*	Результаты обучения по дисциплине	
ОПК-10 Способен разрабатывать и применять математические модели и методы анализа массивов данных и интерпретировать профессиональный смысл получаемых формальных результатов	ОПК-10.1.1.	Знать методы сбора и обработки больших данных	Тестовые вопросы, КР
	ОПК-10.1.2.	Знать методы и системы хранения больших данных	
	ОПК-10.1.3.	Знать типовые прикладные задачи анализа больших данных	
	ОПК-10.2.1.	Уметь выполнять автоматизацию операций в процессе сбора и обработки данных	
	ОПК-10.2.2.	Уметь решать типовые прикладные задачи анализа больших данных	

	ОПК-10.3.1.	Владеть навыками работы с программным обеспечением для автоматического анализа текстов: морфологическими и синтаксическими анализаторами, конкордансами, системами извлечения фактов и отношений, инструментами кластеризации, классификации и тематического моделирования коллекций документов	
ОПК-15 Способен осуществлять автоматизированную информационно-аналитическую поддержку процессов принятия решений на базе ситуационных центров	ОПК-15.1.1	Знать методологические основы анализа данных	Тестовые вопросы, КР
	ОПК-15.1.2	Знать методы снижения размерности многомерных данных	
	ОПК-15.2.1	Уметь применять методы анализа массивов данных при разработке алгоритмов анализа и обработки измерительной информации	
	ОПК-15.2.2	Уметь ставить и решать практические задачи анализа данных в условиях различной полноты исходной информации	
	ОПК-15.2.3	Уметь проводить комплексный анализ данных с использованием базовых параметрических и непараметрических моделей	
	ОПК-15.2.4	Уметь применять автоматизированные технологии семантической обработки текстов при решении прикладных информационно-аналитических задач, в том числе для автоматизированной информационно-аналитической поддержки процессов принятия решений	
	ОПК-15.3.1	Владеть- навыками работы с программным обеспечением для автоматического анализа текстов: морфологическими и синтаксическими анализаторами, конкордансами, системами извлечения фактов и отношений, инструментами кластеризации, классификации и тематического моделирования коллекций документов	
	ОПК-15.3.2	Владеть навыками решения формализованных математических задач анализа данных с помощью пакетов прикладных программ	
ОПК-1.1 Способен разрабатывать и применять автоматизированные технологии обработки естественно-языковых текстов	ОПК-1.1-1.1	Знать основные типы задач обработки и анализа естественно-языковых текстов, основные типы автоматизированной информационно-аналитической поддержки процессов принятия решений	Тестовые вопросы, КР
	ОПК-1.1-1.2	Знать основные виды автоматизированных систем обработки и анализа естественно-языковых текстов	
	ОПК-1.1-1.3	Знать основные математические модели, методы и алгоритмы решения типовых задач обработки и анализа естественно-языковых текстов в ИАС	

	ОПК-1.1-2.1	Уметь проводить оценку качества и осуществлять выбор автоматизированной технологии семантической обработки текстов в конкретных условиях решения прикладных информационно-аналитических задач	
	ОПК-1.1-2.2	Уметь применять автоматизированные технологии семантической обработки текстов при решении прикладных информационно-аналитических задач, в том числе для автоматизированной информационно-аналитической поддержки процессов принятия решений	
	ОПК-1.1-3.1	Владеть навыками работы с программным обеспечением для автоматического анализа текстов: морфологическими и синтаксическими анализаторами, конкордансами, системами извлечения фактов и отношений, инструментами кластеризации, классификации и тематического моделирования коллекций документов	
	ОПК-1.1-3.2	Владеть навыками решения формализованных математических задач анализа данных с помощью пакетов прикладных программ	

4. ОБЪЕМ И СТРУКТУРА ДИСЦИПЛИНЫ

Трудоемкость дисциплины составляет 12 зачетных единиц, 432 часа

Тематический план
форма обучения – очная

№ п/п	Наименование тем и/или разделов/тем дисциплины	Семестр	Неделя семестра	Контактная работа обучающихся с педагогическим работником				Самостоятельная работа	Формы текущего контроля успеваемости, форма промежуточной аттестации *(по семестрам)*
				Лекции	Практические занятия	Лабораторные работы	*в форме практической подготовки*		
1	Введение в анализ данных. Проблема обработки данных. Матрица данных.	7	1-2	2		4		14	
2	Классификация данных с использованием детерминированных моделей. Решающие поверхности и дискриминантные функции.	7	3-4	2		4		14	
3	Процедуры обучения с коррекцией ошибок: правило с фиксированным приращением, правило абсолютной коррекции, частично корректирующее правило.	7	5-6	2		4		14	Рейтинг-контроль №1

4	Классификация данных на основе статистических моделей. Функция потерь. Байесовская дискриминантная функция.	7	7-8	2		4		14	
5	Примеры построения статистических дискриминантных функций для различных статистических нескольких моделей данных.	7	9-10	2		4		14	
6	Кластер-анализ. Основные типы задач кластер-анализа. Меры подобия и функции расстояния. Выбор критерия кластеризации.	7	11-12	2		4		14	Рейтинг-контроль №2
7	Методы снижения размерностей данных. Анализ матриц исходных данных.	7	13-14	2		4		14	
8	Методы прогнозирования временных рядов.	7	15-16	2		4		14	
9	Системы DATA MINING. в задачах анализа и интерпретации данных	7	17-18	2		4		14	Рейтинг-контроль №3
Всего за 7 семестр:		**180**		**18**		**36**		**126**	**Зачет**
1	Этапы развития систем искусственного интеллекта (СИИ). Основные направления развития исследований в области систем искусственного интеллекта. Нейробионический подход.	8	1	2				7	
2	Системы, основанные на знаниях. Извлечение знаний. Интеграция знаний. Базы знаний.	8	2	2		4		8	
3	Структура систем искусственного интеллекта. Архитектура СИИ. Методология построения СИИ.	8	3	2				7	
4	Экспертные системы (ЭС) как вид СИИ. Общая структура и схема функционирования ЭС.	8	4	2		4		8	
5	Представление знаний. Основные понятия. Состав знаний СИИ. Организация знаний СИИ.	8	5	2				8	
6	Модели представления знаний. Представление знаний с помощью системы продукций. Суб-технологии искусственного интеллекта.	8	6	2		4		7	Рейтинг-контроль №1
7	Стандарт для решения задач анализа данных. Роли участников в проектах по анализу данных.	8	7	2				7	
8	Внедрение систем машинного обучения в «отрасли»: ключевые примеры использования ИИ в отрасли (кейсы)	8	8	2		4		8	
9	Системы продукций. Управление выводом в продукционной системе. Представление знаний с помощью логики предикатов.	8	9	2				7	
10	Логические модели. Логика предикатов как форма представления знаний. Синтаксис и семантика логики предикатов.	8	10	2		4		8	

11	Технологии манипулирования знаниями СИИ. Программные комплексы решения интеллектуальных задач. Естественно-языковые программы.	8	11	2				8	
12	Представление знаний фреймами и вывод на фреймах. Теория фреймов. Модели представления знаний фреймами.	8	12	2		4		7	Рейтинг-контроль №2
13	Основные положения нечеткой логики. Представление знаний и вывод в моделях нечеткой логики.	8	13	2				7	
14	Программные комплексы. Основы программирования для задач анализа данных. Изучение отдельных направлений анализа данных.	8	14	2		4		8	
15	Задача классификации. Ансамбли моделей машинного обучения для задачи классификации.	8	15	2				7	
16	Нейронные сети. Глубокие нейронные сети (компьютерное зрение, разбор естественного языка, анализ табличных данных).	8	16	2		4		8	
17	Кластеризация и другие задачи обучения. Задачи работы с последовательным данным, обработка естественного языка.	8	17	2				8	
18	Рекомендательные системы. Определение важности признаков и снижение размерности	8	18	2		4		7	Рейтинг-контроль №3
Всего за 8 семестр:		**180**		**36**		**36**		**135**	**Экзамен (45)**
Наличие в дисциплине КП/КР		**Есть (8)**							
Итого по дисциплине		**432**		**54**		**72**		**261**	**Зачет Экзамен (45) Курсовая работа**

Содержание лекционных занятий по дисциплине

7 семестр

Раздел 1. Введение в дисциплину

Тема 1. Введение в анализ данных. **Содержание темы.**

Проблема обработки данных. Матрица данных. Гипотезы компактности и скрытых факторов. Структура матрицы данных и задачи обработки. Матрица объект-объект и признак-признак. Расстояние и близость. Измерение признаков. Отношения и их представление. Основные проблемы измерений. Основные типы шкал. Проблема адекватности. Основные задачи анализа и интерпретации данных.

Тема 2. Классификация данных с использованием детерминированных моделей. **Содержание темы.**

Решающие поверхности и дискриминантные функции. Линейные дискриминантные функции классификатор по минимуму расстояния. Линейная разделимость. Кусочно-линейные дискриминантные функции. Нелинейные дискриминантные функции. Фи-машины. Потенциальные функции как дискриминантные функции. Пространство весов.

Тема 3. Процедуры обучения с коррекцией ошибок: правило с фиксированным приращением, правило абсолютной коррекции, частично корректирующее правило. **Содержание темы.**

Обобщенные градиентные методы. Персептронный критерий. Процедуры обучения на основе минимальной среднеквадратичной ошибки: псевдоинверсный метод, метод Хо-Кашьпа.

Тема 4. Классификация данных на основе статистических моделей. **Содержание темы.**

Функция потерь. Байесовская дискриминантная функция. Принятие решение по максимуму правдоподобия. Оптимальная дискриминантная функция для нормально распределенных образов. Дискриминантная функция Фишера. Множественный дискриминантный анализ. Пошаговый дискриминантный анализ. Ошибки классификации.

Тема 5. Примеры построения статистических дискриминантных функций для различных статистических нескольких моделей данных. **Содержание темы.**

Обучение для статистических дискриминантных функций. Оценки максимального правдоподобия, байесовские оценки. Непараметрическое оценивание. Парзеновские окна, метод непараметрического оценивания на основе К-ближайшего соседства.

Раздел 2. Методы прогнозирование, анализ матриц.

Тема 6. Кластер-анализ. Основные типы задач кластер-анализа. **Содержание темы.**

Меры подобия и функции расстояния. Выбор критерия кластеризации. Кластерные методы, основанные на евклидовой метрике. Иерархическая кластеризация. Метод К-внутригрупповых средних. Использование методов теории графов в задачах кластеризации. Кластеризация на основе анализа плотностей вероятностей.

Тема 7. Методы снижения размерностей данных. **Содержание темы.**

Анализ матриц исходных данных. Метод главных компонент. Корреляционная матрица и ее основные свойства. Собственные векторы и собственные числа корреляционной матрицы. Приведение корреляционной матрицы к диагональной форме. Геометрическая интерпретация главных компонент на плоскости. Модели факторного анализа. Оценка факторных нагрузок методом максимального правдоподобия и центроидным методом. Вращение факторов и их интерпретация. Использование кластеризации признаков для снижения размерности. Многомерное шкалирование (МИ). Метрический и неметрический подход к МИ. Методы ортогонального проектирования. Нелинейные методы МИ. Многомерное шкалирование неметрических данных. Многомерные развертки.

Тема 8. Методы прогнозирования временных рядов. **Содержание темы.**

Классификация методов прогнозирования. Оценивание трендов. Методы скользящего среднего. Экспоненциальное сглаживание. Регрессионный анализ и прогнозирование. Линейные параметрические модели временных рядов. Методы оценивания моделей авторегрессии, скользящего среднего и смешанных моделей. Сезонные модели. Прогнозирование на основе параметрических моделей. Прогнозирование с использованием нейронных сетей.

Тема 9. Системы DATA MINING. В задачах анализа и интерпретации данных. **Содержание темы.**

Понятие об интеллектуальных системах анализа и интерпретации данных. DATA MINING – системы извлечения новых знаний из данных. Типы систем DATA MINING - предметно-ориентированные аналитические системы, статистические пакеты, нейронные сети, деревья решений, обнаружение логических закономерностей, генетические алгоритмы, системы визуализации многомерных данных

8 семестр

Раздел 1.

Тема 1. Этапы развития систем искусственного интеллекта (СИИ). Основные направления развития исследований в области систем искусственного интеллекта. Нейробионический подход.

Тема 2. Системы, основанные на знаниях. Извлечение знаний. Интеграция знаний. Базы знаний.

Тема 3. Структура систем искусственного интеллекта. Архитектура СИИ. Методология построения СИИ.

Тема 4. Экспертные системы (ЭС) как вид СИИ. Общая структура и схема функционирования ЭС.

Тема 5. Представление знаний. Основные понятия. Состав знаний СИИ. Организация знаний СИИ.

Тема 6. Модели представления знаний. Представление знаний с помощью системы продукций. Суб-технологии искусственного интеллекта.

Тема 7. Стандарт для решения задач анализа данных. Роли участников в проектах по анализу данных.

Тема 8. Внедрение систем машинного обучения в «отрасли»: ключевые примеры использования ИИ в отрасли (кейсы)

Тема 9. Управление выводом в продукционной системе. Представление знаний с помощью логики предикатов.

Раздел 2

Тема 10. Логические модели. Логика предикатов как форма представления знаний. Синтаксис и семантика логики предикатов.

Тема 11. Технологии манипулирования знаниями СИИ. Программные комплексы решения интеллектуальных задач. Естественно-языковые программы.

Тема 12. Представление знаний фреймами и вывод на фреймах. Теория фреймов. Модели представления знаний фреймами.

Тема 13. Основные положения нечеткой логики. Представление знаний и вывод в моделях нечеткой логики.

Тема 14. Программные комплексы. Основы программирования для задач анализа данных. Изучение отдельных направлений анализа данных.

Тема 15. Задача классификации. Ансамбли моделей машинного обучения для задачи классификации.

Тема 16. Нейронные сети. Глубокие нейронные сети (компьютерное зрение, разбор естественного языка, анализ табличных данных).

Тема 17. Кластеризация и другие задачи обучения. Задачи работы с последовательным данным, обработка естественного языка.

Тема 18. Рекомендательные системы. Определение важности признаков и снижение размерности

Содержание лабораторных занятий по дисциплине
Перечень лабораторных работ 7 семестр:

Лабораторная работа № 1. Предварительный анализ данных.

Содержание лабораторной работы.

Предварительный анализ данных с использованием специализированного программного обеспечения (по выбору преподавателя).

Лабораторная работа № 2. Изучение методов дискриминантного анализа с использованием специализированного программного обеспечения (по выбору преподавателя).

Содержание лабораторной работы. Дискриминантный анализ используется для принятия решения о том, какие переменные различают (дискриминируют) две или более возникающие совокупности (группы). Например, некий исследователь в области образования может захотеть исследовать, какие переменные относят выпускника средней школы к одной из трех категорий: (1) поступающий в колледж, (2) поступающий в профессиональную школу или (3) отказывающийся от дальнейшего образования или профессиональной подготовки. Для этой цели исследователь может собрать данные о различных переменных, связанных с

учащимися школы. После выпуска большинство учащихся естественно должно попасть в одну из названных категорий. Затем можно использовать *Дискриминантный анализ* для определения того, какие переменные дают наилучшее предсказание выбора учащимися дальнейшего пути.

Лабораторная работа № 3. Изучение методов кластер-анализа с использованием специализированного программного обеспечения (по выбору преподавателя).

Содержание лабораторной работы. Кластерный анализ (англ. cluster analysis) — многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы.

Лабораторная работа № 4. Изучение методов факторного-анализа с использованием специализированного программного обеспечения (по выбору преподавателя). **Содержание лабораторной работы. Фáкторный анализ** — многомерный **метод**, применяемый для изучения взаимосвязей между значениями переменных. Предполагается, что известные переменные зависят от меньшего количества неизвестных переменных и случайной ошибки.

Лабораторная работа № 5. Классификация данных и изучение методов снижения размерности данных с использованием специализированного программного обеспечения (по выбору преподавателя). **Содержание лабораторной работы.** В статистике, машинном обучении и теории информации **снижение размерности** — это преобразование данных, состоящее в уменьшении числа переменных путём получения главных переменных. Преобразование может быть разделено на отбор признаков и выделение признаков

Перечень лабораторных работ 8 семестр:

Лабораторная работа №1. Состав знаний и способы их представления. Управляющий механизм. Объяснительные способности

Лабораторная работа №2. Нейроподобные структуры. Системы типа персептронов. Нейрокомпьютеры и их программное обеспечение.

Лабораторная работа №3. Системы когнитивной графики. Интеллектуальные системы. Обучающие системы

Лабораторная работа №4. Интеллектуальный интерфейс: лингвистический процессор, анализ и синтез речи

Лабораторная работа №5. Онтологии и онтологические системы. Системы и средства представления онтологических знаний

Лабораторная работа №6. Онтологии как аппарат моделирования системы знаний. Методы представления онтологий

Лабораторная работа №7. Программные реализации моделей нечеткой логики

Лабораторная работа №8. Программные реализации алгоритмов Мамдани, Суджсно

Лабораторная работа №9. Программные реализации алгоритмов Цукамото, Ларсена

5. ОЦЕНОЧНЫЕ СРЕДСТВА ДЛЯ ТЕКУЩЕГО КОНТРОЛЯ УСПЕВАЕМОСТИ, ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ПО ИТОГАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ И УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ САМОСТОЯТЕЛЬНОЙ РАБОТЫ СТУДЕНТОВ

5.1. Текущий контроль успеваемости

7 семестр.

Вопросы рейтинг-контроля №1

- Введение в анализ данных. Проблема обработки данных. Матрица данных. Гипотезы компактности и скрытых факторов.

- Структура матрицы данных и задачи обработки. Матрица объект-объект и признак-признак.
- Расстояние и близость. Измерение признаков. Отношения и их представление.
- Основные проблемы измерений. Основные типы шкал. Проблема адекватности. Основные задачи анализа и интерпретации данных.
- Классификация данных с использованием детерминированных моделей.
- Решающие поверхности и дискриминантные функции. Линейные дискриминантные функции классификатор по минимуму расстояния.
- Линейная разделимость. Кусочно-линейные дискриминантные функции.
- Нелинейные дискриминантные функции.
- Фи-машины. Потенциальные функции как дискриминантные функции. Пространство весов.
- Процедуры обучения с коррекцией ошибок: правило с фиксированным приращением, правило абсолютной коррекции, частично корректирующее правило.
- Обобщенные градиентные методы. Персептронный критерий. Процедуры обучения на основе минимальной среднеквадратичной ошибки: псевдоинверсный метод, метод Хо-Кашьпа.
- Классификация данных на основе статистических моделей.
- Функция потерь. Байесовская дискриминантная функция.
- Принятие решение по максимуму правдоподобия. Оптимальная дискриминантная функция для нормально распределенных образов.
- Дискриминантная функция Фишера. Множественный дискриминантный анализ.
- Пошаговый дискриминантный анализ. Ошибки классификции.
- Примеры построения статистических дискриминантных функций для различных статистических нескольких моделей данных.

Вопросы рейтинг-контроля №2

- Вращение факторов и их интерпретация.
- Использование кластеризации признаков для снижения размерности.
- Многомерное шкалирование (МИ). Метрический и неметрический подход к МИ.
- Методы ортогонального проектирования.
- Нелинейные методы МИ. Многомерное шкалирование неметрических данных. Многомерные развертки.
- Методы прогнозирования временных рядов.
- Классификация методов прогнозирования. Оценивание трендов.
- Методы скользящего среднего. Экспоненциальное сглаживание.
- Регрессионный анализ и прогнозирование. Линейные параметрические модели временных рядов.
- Методы оценивания моделей авторегрессии, скользящего среднего и смешанных моделей.
- Сезонные модели. Прогнозирование на основе параметрических моделей. Прогнозирование с использованием нейронных сетей.
- Системы DATA MINING. в задачах анализа и интерпретации данных.
- Понятие об интеллектуальных системах анализа и интерпретации данных. DATA MINING - системы извлечения новых знаний из данных. Типы систем DATA MINING -предметно-ориентированные аналитические системы, статистические пакеты, нейронные сети, деревья решений, обнаружение логических закономерностей, генетические алгоритмы, системы визуализации многомерных данных
- Автоматическая обработка языка и компьютерная лингвистика. Задачи автоматической обработки текста в научных исследованиях.
- Основные задачи компьютерной лингвистики и история развития автоматической обработки языка.

Вопросы рейтинг-контроля №3

- Буквенные и словарные n-граммы. Контекстное окно. Применения N-грамм в автоматической обработке языка. Роль биграмм и триграмм.
- Определение языка по письменности. Языковые модели. Цепь Маркова. Коллокации.
- Формальные определения и лингвистический смысл коллокаций. Меры ассоциации.
- Коэффициент взаимной информации (MI). T-score. Отношение правдоподобия (log-likelihood).
- Статистические тесты ассоциации: хи-квадрат и Fisher exact test.
- Выделение коллокаций по синтаксическому шаблону. Разрывные коллокации.
- Автоматическое определение тематики. Векторное представление текста для задач информационного поиска.
- Открытые и закрытые классы слов. Стоп-слова. Динамические списки стоп слов. Порог отсечения по частотности и DF. Дистрибутивная семантика. Совместная встречаемость и семантическая близость.
- Кластеризация текстов. Задачи и область применения кластерных методов.
- Виды кластеризации: плоские, аггломеративные, нечеткие. Меры близости: евклидово расстояние, косинусная мера. Популярные алгоритмы кластеризации: k-средних, DBCLUST, спектральные алгоритмы.
- ПО для кластеризации текстов. Пакеты кластеризации для R. gCLUTO.
- Классификация текстов. Машинное обучение с учителем и без учителя в задачах классификации текстов.
- Популярные алгоритмы классификации: наивный байесовский метод, метод опорных векторов, деревья принятия решений.
- ПО для классификации текстов. SVMLight. Пространственное моделирование семантических отношений (word space). Латентный семантический анализ.
- Вероятностный латентно семантический анализ. Тематическое моделирование. Метод латентного размещения Дирихле. ПО для латентного семантического анализа и тематического моделирования. Mallet. sTMT.
- Извлечение мнений и оценок (Sentiment analysis). Область применения методов извлечения мнений и оценок.

8 семестр.

Вопросы рейтинг-контроля № 1

- Этапы развития систем искусственного интеллекта (СИИ). Основные направления развития исследований в области систем искусственного интеллекта.
- Нейробионический подход.
- Системы, основанные на знаниях. Извлечение знаний.
- Интеграция знаний. Базы знаний.
- Структура систем искусственного интеллекта.
- Архитектура СИИ. Методология построения СИИ.
- Экспертные системы (ЭС) как вид СИИ.
- Общая структура и схема функционирования ЭС.
- Представление знаний. Основные понятия.
- Состав знаний СИИ. Организация знаний СИИ.

Вопросы рейтинг-контроля №2

- Модели представления знаний.
- Представление знаний с помощью системы продукций.
- Суб-технологии искусственного интеллекта.
- Стандарт для решения задач анализа данных. Роли участников в проектах по анализу данных.
- Системы продукций. Управление выводом в продукционной системе.

- Представление знаний с помощью логики предикатов.
- Логические модели. Логика предикатов как форма представления знаний.
- Синтаксис и семантика логики предикатов.
- Технологии манипулирования знаниями СИИ.
- Программные комплексы решения интеллектуальных задач.
- Естественно-языковые программы.
- Представление знаний фреймами и вывод на фреймах.
- Теория фреймов. Модели представления знаний фреймами.

Вопросы рейтинг-контроля №3

- Основные положения нечеткой логики.
- Представление знаний и вывод в моделях нечеткой логики.
- Программные комплексы. Основы программирования для задач анализа данных. Изучение отдельных направлений анализа данных.
- Задача классификации. Ансамбли моделей машинного обучения для задачи классификации.
- Нейронные сети. Глубокие нейронные сети (компьютерное зрение, разбор естественного языка, анализ табличных данных).
- Кластеризация и другие задачи обучения.
- Задачи работы с последовательным данным, обработка естественного языка.
- Рекомендательные системы. Определение важности признаков и снижение размерности

5.2. Промежуточная аттестация по итогам освоения дисциплины.

Примерный перечень вопросов к зачету за 7 семестр

1. Введение в анализ данных. Проблема обработки данных. Матрица данных. Гипотезы компактности и скрытых факторов.
2. Структура матрицы данных и задачи обработки. Матрица объект-объект и признак-признак.
3. Расстояние и близость. Измерение признаков. Отношения и их представление.
4. Основные проблемы измерений. Основные типы шкал. Проблема адекватности. Основные задачи анализа и интерпретации данных.
5. Классификация данных с использованием детерминированных моделей.
6. Решающие поверхности и дискриминантные функции. Линейные дискриминантные функции классификатор по минимуму расстояния.
7. Линейная разделимость. Кусочно-линейные дискриминантные функции.
8. Нелинейные дискриминантные функции.
9. Фи-машины. Потенциальные функции как дискриминантные функции. Пространство весов.
10. Процедуры обучения с коррекцией ошибок: правило с фиксированным приращением, правило абсолютной коррекции, частично корректирующее правило.
11. Обобщенные градиентные методы. Персептронный критерий. Процедуры обучения на основе минимальной среднеквадратичной ошибки: псевдоинверсный метод, метод Хо-Кашьпа.
12. Классификация данных на основе статистических моделей.
13. Функция потерь. Байесовская дискриминантная функция.
14. Принятие решение по максимуму правдоподобия. Оптимальная дискриминантная функция для нормально распределенных образов.
15. Дискриминантная функция Фишера. Множественный дискриминантный анализ.
16. Пошаговый дискриминантный анализ. Ошибки классификции.
17. Примеры построения статистических дискриминантных функций для различных статистических нескольких моделей данных.
18. Обучение для статистических дискриминантных функций.
19. Оценки максимального правдоподобия, байесовские оценки.

20. Непараметрическое оценивание. Парзеновские окна, метод непараметрического оценивания на основе К-ближайшего соседства.
21. Кластер-анализ. Основные типы задач кластер-анализа.
22. Меры подобия и функции расстояния. Выбор критерия кластеризации. Кластерные методы, основанные на евклидовой метрике.
23. Иерархическая кластеризация. Метод К-внутригрупповых средних.
24. Использование методов теории графов в задачах кластеризации.
25. Кластеризация на основе анализа плотностей вероятностей.
26. Методы снижения размерностей данных. Анализ матриц исходных данных.
27. Метод главных компонент. Корреляционная матрица и ее основные свойства.
28. Собственные векторы и собственные числа корреляционной матрицы. Приведение корреляционной матрицы к диагональной форме.
29. Геометрическая интерпретация главных компонент на плоскости.

Примерный перечень вопросов к экзамену за 8 семестр

1. Этапы развития систем искусственного интеллекта (СИИ). Основные направления развития исследований в области систем искусственного интеллекта.
3. Нейробионический подход.
4. Системы, основанные на знаниях. Извлечение знаний.
5. Интеграция знаний. Базы знаний.
6. Структура систем искусственного интеллекта.
7. Архитектура СИИ. Методология построения СИИ.
8. Экспертные системы (ЭС) как вид СИИ.
9. Общая структура и схема функционирования ЭС.
10. Представление знаний. Основные понятия.
11. Состав знаний СИИ. Организация знаний СИИ.
12. Модели представления знаний.
13. Представление знаний с помощью системы продукций.
14. Суб-технологии искусственного интеллекта.
15. Стандарт для решения задач анализа данных. Роли участников в проектах по анализу данных.
16. Системы продукций. Управление выводом в продукционной системе.
17. Представление знаний с помощью логики предикатов.
18. Логические модели. Логика предикатов как форма представления знаний.
19. Синтаксис и семантика логики предикатов.
20. Технологии манипулирования знаниями СИИ.
21. Программные комплексы решения интеллектуальных задач.
22. Естественно-языковые программы.
23. Представление знаний фреймами и вывод на фреймах.
24. Теория фреймов. Модели представления знаний фреймами.
25. Основные положения нечеткой логики.
26. Представление знаний и вывод в моделях нечеткой логики.
27. Программные комплексы. Основы программирования для задач анализа данных. Изучение отдельных направлений анализа данных.
28. Задача классификации. Ансамбли моделей машинного обучения для задачи классификации.
29. Нейронные сети. Глубокие нейронные сети (компьютерное зрение, разбор естественного языка, анализ табличных данных).
30. Кластеризация и другие задачи обучения.
31. Задачи работы с последовательным данным, обработка естественного языка.
32. Рекомендательные системы. Определение важности признаков и снижение размерности

5.3. Самостоятельная работа обучающегося.

Примерные темы курсовой работы 8 семестр.

Анализ данных с использованием специализированного программного обеспечения по выбору преподавателя по следующим основным направлениям:

- Частотный анализ лексики и ключевые слова. Частотное распределение лексики в языке.
- Локальные модели контекста. Вероятностные языковые модели
- Автоматическое определение тематики.
- Векторное представление текста для задач информационного поиска.
- ПО для кластеризации текстов. Пакеты кластеризации для R. gCLUTO. Классификация текстов
- Извлечение мнений и оценок (Sentiment analysis).
- Извлечение фактов и отношений. Синтаксис и формальные языки
- Автоматический анализ стиля. Стилометрия.

Примерные вопросы и задания для самостоятельной работы студентов

7 семестр.

- Классификация данных с использованием детерминированных моделей.
- Классификация данных на основе статистических моделей.
- Кластер-анализ данных.
- Методы снижения размерностей данных.
- Методы прогнозирования временных рядов.
- Системы DATA MINING. в задачах анализа и интерпретации данных.
- Современные пакеты прикладных программ для решения задач обработки экспериментальных данных.
- Частотный анализ лексики и ключевые слова.
- Локальные модели контекста. Вероятностные языковые модели.
- Автоматическое определение тематики при исследовании текстов.
- Извлечение мнений и оценок при исследовании текстов.
- Извлечение фактов и отношений при исследовании текстов.
- Автоматический анализ стиля при исследовании текстов.
- Основы обработки неструктурированных (текстовых) данных в корпоративных информационных системах (ERP, АСУП и др.) и современных веб-приложениях.

8 семестр.

Тема 1. Структура систем искусственного интеллекта.
Тема 2. Архитектура СИИ. Методология построения СИИ.
Тема 3. Общая структура и схема функционирования ЭС.
Тема 4. Представление знаний. Основные понятия.
Тема 5. Состав знаний СИИ. Организация знаний СИИ.
Тема 6. Представление знаний с помощью системы продукций.
Тема 7. Суб-технологии искусственного интеллекта.
Тема 8. Стандарт для решения задач анализа данных. Роли участников в проектах по анализу данных.

Фонд оценочных материалов (ФОМ) для проведения аттестации уровня сформированности компетенций обучающихся по дисциплине оформляется отдельным документом.

6. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

6.1. Книгообеспеченность

Наименование литературы: автор, название, вид издания, издательство	Год издания	КНИГООБЕСПЕЧЕННОСТЬ
		Наличие в электронном каталоге ЭБС
Основная литература*		
1. Монахова, Г. Е. ИНФОРМАЦИОННЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ. Визуализация многомерных пространственных данных средствами геоинформационных систем : учеб. пособие [Электронный ресурс] / Г. Е. Монахова, М. М. Монахова ; под ред. проф. М. Ю. Монахова ; Владим. гос. ун-т им. А. Г. и Н. Г. Столетовых. – Владимир : Изд-во ВлГУ, 2019. – 392 с.	2021	http://dspace.www1.vlsu.ru/handle/123456789/8324
2. Никитин, О. Р. Методы статистической обработки экспериментальных исследований : учеб. пособие / О. Р. Никитин Н. Н. Корнеева ; Владим. гос. ун-т им. А. Г. и Н. Г. Столетовых. – Владимир : Изд-во ВлГУ, 2019. – 191 с. ISBN 978-5-9984-0982-0	2019	http://dspace.www1.vlsu.ru/bitstream/123456789/7805/1/01830.pdf
3. Интеллектуальный анализ данных : учеб. пособие / Д. В. Виноградов ; Владим. гос. ун-т им. А. Г. и Н. Г. Столетовых. –Владимир : Изд-во ВлГУ, 2021. – 260 с. ISBN 978-5-9984-1452-7	2021	http://dspace.www1.vlsu.ru/handle/123456789/4626
Дополнительная литература		
1. Статистический анализ данных, моделирование и исследование вероятностных закономерностей. Компьютерный подход / Б.Ю. Лемешко, С.Б. Лемешко, С.Н. Постовалов и др. - М.: НИЦ ИНФРА-М- 890 с. ежим доступа:	2015	http://znanium.com/catalog.php?bookinfo=515227
2. Численный вероятностный анализ неопределенных данных/ДобронецБ.С., ПоповаО.А. - Краснояр.: СФУ. - 168 с.: ISBN 978-5-7638-3093-4	2014	http://znanium.com/catalog.php?bookinfo=549444

6.2. Периодические издания

1. Журнал «Вопросы защиты информации». Режим доступа: http://i-vimi.ru/editions/detail.php?SECTION_ID=155/;
2. Журнал "Information Security/Информационная безопасность". Режим доступа: http://www.itsec.ru/insec-about.php.
3. Ежемесячный теоретический и прикладной научно-технический журнал «Информационные технологии». Режим доступа http://novtex.ru/IT/.

6.3. Интернет-ресурсы

http://www.dialog-21.ru/— Диалог.Международная конференция по компьютерной лингвистике.

http://nlpub.ru— Каталог лингвистических ресурсов для обработки русского языка.

http://www.regular-expressions.info— The Premier website about Regular Expressions.

http://sentiment.christopherpotts.net/— Sentiment symposium tutorial.

http://www.aclweb.org/anthology/— ACL Anthology

A Digital Archive of Research Papers in Computational Linguistics.

Программные средства. Для успешного освоения дисциплины, студент использует следующие программные средства: - Программа построения частотных словарей. http://alingva.ru/index.php/lingvosoft/12-ngramfrequency; - mystem. Морфологический анализатор для русского языка. http://company.yandex.ru/technologies/mystem/-

LSA. Латентно-семантический анализ текстовых данных. http://alingva.ru/index.php/lingvosoft/17--lsa; - Tomita-пасрер. Инструмент для извлечения структурированных данных из текста на естественном языке. http://api.yandex.ru/tomita/; - Модуль Perl Text::NSP. N-gram statistics and association measures. http://search.cpan.org/dist/Text-NSP/lib/Text/NSP/Measures.pm; - Stanford Topic Modeling Toolbox; - http://nlp.stanford.edu/software/tmt/tmt-0.4/

Тестовые массивы текстов для обработки публикуются на сайте: http://maslinsky.spb.ru/courses/cmta2013/

7. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

ауд. 408-2, Лекционная аудитория, количество студенческих мест – 50, площадь 60 м2, оснащение: мультимедийное оборудование (интерактивная доска Hitachi FX-77WD, проектор BenQ MX 503 DLP 2700ANSI XGA), ноутбук Lenovo Idea Pad B5045

ауд. 427а-2, лаборатория сетевых технологий, количество студенческих мест – 14, площадь 36 м2, оснащение: компьютерный класс с 8 рабочими станциями Core 2 Duo E8400 с выходом в Internet, 3 маршрутизатора Cisco 2800 Series, 6 маршрутизаторов Cisco 2621, 6 коммутаторов Cisco Catalyst 2960 Series, 3 коммутатора Cisco Catalyst 2950 Series, коммутатор Cisco Catalyst Express 500 Series, проектор BenQ MP 620 P, экран настенный рулонный. Лицензионное программное обеспечение: операционная система Windows 7 Профессиональная, офисный пакет приложений Microsoft Office Профессиональный плюс 2007, бесплатно распространяемое программное обеспечение: линейка интегрированных сред разработки Visual Studio Express 2012, программный продукт виртуализации Oracle VM VirtualBox 5.0.4, симулятор сети передачи данных Cisco Packet Tracer 7.0, интегрированная среда разработки программного обеспечения IntelliJ IDEA Community Edition 15.0.3.

ауд. 427б-2, УНЦ «Комплексная защита объектов информатизации», количество студенческих мест – 15, площадь 52 м2, оснащение: компьютерный класс с 7 рабочими станциями Alliance Optima P4 с выходом в Internet, коммутатор D-Link DGS-1100-16 мультимедийный комплект (проектор Toshiba TLP X200, экран настенный рулонный), прибор ST-031P «Пиранья-Р» многофункциональный поисковый, прибор «Улан-2» поисковый, виброакустический генератор шума «Соната АВ 1М», имитатор работы средств нелегального съема информации, работающих по радиоканалу «Шиповник», анализатор спектра «GoodWill GSP-827», индикатор поля «SEL SP-75 Black Hunter», устройство блокирования работы систем мобильной связи «Мозайка-3», устройство защиты телефонных переговоров от прослушивания «Прокруст 2000», диктофон Edic MINI Hunter, локатор «Родник-2К» нелинейный, комплекс проведения акустических и виброакустических измерений «Спрут мини-А», видеорегистратор цифровой Best DVR-405, генератор Шума «Гном-3»,учебно-исследовательский комплекс «Сверхширокополосные беспроводные сенсорные сети» (Nano Xaos), сканирующий приемник «Icom IC-R1500», анализатор сетей Wi-Fi Fluke AirCheck с активной антенной. Лицензионное программное обеспечение: Windows 8 Профессиональная, офисный пакет приложений Microsoft Office Профессиональный плюс 2010, бесплатно распространяемое программное обеспечение: линейка интегрированных сред разработки Visual Studio Express 2012, инструмент имитационного моделирования AnyLogic 7.2.0 Personal Learning Edition, интегрированная среда разработки программного обеспечения IntelliJ IDEA Community Edition 14.1.4.

Рабочую программу составил: к.т.н, доцент кафедры ИЗИ Монахов Ю. М.______________

Рецензент: Руководитель направления по информационной безопасности акционерного общества «ОМК» г. Владимир, к.т.н. Абрамов К. Г.______________________________

Программа рассмотрена и одобрена на заседании кафедры ИЗИ

Протокол № 13 от 12.05.23 года
Заведующий кафедрой д.т.н., профессор______________________________/М.Ю. Монахов/

Рабочая программа рассмотрена и одобрена на заседании учебно-методической комиссии специальности 10.05.04 «Информационно-аналитические системы безопасности»

Протокол № 13 от 12.05.23 года
Председатель комиссии д.т.н., профессор______________________________/М.Ю. Монахов/

ЛИСТ ПЕРЕУТВЕРЖДЕНИЯ
РАБОЧЕЙ ПРОГРАММЫ ДИСЦИПЛИНЫ

Рабочая программа одобрена на 20____ / 20____ учебный год

Протокол заседания кафедры № ______ от _________ года

Заведующий кафедрой _д.т.н., профессор_____________________________/М.Ю. Монахов/__

(ФИО, подпись)

Рабочая программа одобрена на 20____ / 20____ учебный год

Протокол заседания кафедры № ______ от _________ года

Заведующий кафедрой _д.т.н., профессор_____________________________/М.Ю. Монахов/__

(ФИО, подпись)

Рабочая программа одобрена на 20____ / 20____ учебный год

Протокол заседания кафедры № ______ от _________ года

Заведующий кафедрой _д.т.н., профессор_____________________________/М.Ю. Монахов/__

(ФИО, подпись)

Рабочая программа одобрена на 20____ / 20____ учебный год

Протокол заседания кафедры № ______ от _________ года

Заведующий кафедрой _д.т.н., профессор_____________________________/М.Ю. Монахов/__

(ФИО, подпись)

ЛИСТ РЕГИСТРАЦИИ ИЗМЕНЕНИЙ

в рабочую программу дисциплины

Методы анализа данных

образовательной программы специальности

10.05.04 «Информационно-аналитические системы безопасности»

Номер изменения	Внесены изменения в части/разделы рабочей программы	Исполнитель ФИО	Основание (номер и дата протокола заседания кафедры)
1			
2			

Заведующий кафедрой _____/М.Ю. Монахов/__

Подпись *ФИО*