

Документ подписан простой электронной подписью

Информация о владельце:

ФИО: Горбунов Алексей Александрович

Должность: Заместитель Начальника университета по учебной работе

Дата подписания: 27.08.2024 15:56:48

Уникальный программный ключ:

286e49ee1471d400cc1f45539d51ed7bbf0e9cc7

**Федеральное государственное бюджетное образовательное
Учреждение высшего образования
«Санкт-Петербургский университет
Государственной противопожарной службы МЧС России»**

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

ТЕХНОЛОГИИ И СИСТЕМЫ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ

**Магистратура по направлению подготовки
27.04.03 Системный анализ и управление
направление (профиль) «Системный анализ и управление в
организационно-технических системах»**

Санкт-Петербург

1. Цели и задачи дисциплины

Цели освоения дисциплины:

- формирование у обучающихся необходимой теоретической базы и практических навыков, которые позволят всесторонне и системно понимать проблемы обработки и анализа информации;
- формирование навыков разработки и анализа концептуальных и теоретических моделей при решении научных и прикладных задач в области информационных технологий.

Перечень компетенций, формируемых в процессе изучения дисциплины

Компетенции	Содержание
ОПК-9	способен разрабатывать новые и модифицировать существующие методы системного анализа для адаптивного и робастного управления техническими объектами в условиях регулярной и хаотической динамики
ПК-5	способен применять методы математического и системного анализа и теории принятия решений для исследования функциональных задач управления организационно-техническими системами на основе отечественных и мировых тенденций развития методов, управления, информационных и интеллектуальных технологий

Задачи дисциплины

- сформировать целостное представление о современных проблемах анализа и обработки больших данных;
- владеть навыками разработки и анализа концептуальных и теоретических моделей прикладных задач анализа больших данных;
- уметь оценивать время и необходимые аппаратные ресурсы для решения задач анализа и обработки данных.

2. Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы

Индикаторы достижения компетенции	Планируемые результаты обучения по дисциплине
Осуществляет постановку и выполняет эксперименты по проверке корректности научно обоснованных решений в области системного анализа и автоматического управления ОПК-9.1	Знает
	Современные системно-аналитические комплексы, программное обеспечение для выполнения экспериментов по проверке корректности научно обоснованных решений в области системного анализа и автоматического управления ОПК-9.1
	Владеет
	Навыками постановки и выполнения экспериментов по проверке корректности научно обоснованных решений в области системного анализа и автоматического управления ПК-9.1

Тип задачи профессиональной деятельности: организационно-управленческий	
ПК-5.2. Умеет работать с автоматизированными средствами поддержки принятия решений для управления организационно-техническими системами	Знает
	основные требования, предъявляемые к автоматизированным средствам обработки больших данных при управлении сложными системами ПК 5.2
	Умеет ориентироваться и выбирать необходимое автоматизированное средство для обработки больших данных при управлении сложными системами ПК-5.2

3. Место дисциплины в структуре ОПОП

Дисциплина относится к обязательной части, формируемой участниками образовательных отношений основной профессиональной образовательной программы магистратуры по направлению подготовки 27.04.03 Системный анализ и управление направлением (профиль) «Системный анализ и управление в организационно-технических системах».

4. Структура и содержание дисциплины

Общая трудоемкость дисциплины составляет 4 зачетные единицы, 144 часа.

4.1 Распределение трудоемкости учебной дисциплины по заочной форме обучения и видам работ

Вид учебной работы	Трудоемкость			
	з.е.	час.	по курсам	
			1	2
Общая трудоемкость дисциплины по учебному плану	4	144	36	108
Контактная работа, в том числе:		14	4	10
Аудиторные занятия		12	4	8
Лекции (Л)		6	2	4
Практические занятия (ПЗ)		6	2	4
Консультация		2		2
Самостоятельная работа (СРС)		121	32	89
Экзамен		9		9

**4.2 Тематический план, структурированный по темам (разделам) с
указанием отведенного на них количества академических часов и видов
учебных занятий
для заочной формы обучения**

№ п.п.	Наименование разделов и тем	Всего часов	Количество часов по видам занятий			Консультация	Самостоятельная работа	Контроль
			Лекции	Практические занятия	Лабораторная работа			
1	Тема 1 Введение в анализ данных	36	2	2			32	
2	Итого за 1 курс	36	2	2			32	
3	Тема 2. Теория больших данных	50	2	4			44	
4	Тема 3. Технологии анализа больших данных	47	2				45	
5	Консультация	2				2		
6	Экзамен	9						9
7	Итого за 2 курс	108	4	4		2	89	9
8	Итого по дисциплине	144	6	6		2	121	9

**4.3. Содержание дисциплины для обучающихся
для заочной формы обучения**

Тема 1. Введение в анализ данных

Лекции. Данные. Подходы и определения. Жизненный цикл данных. Метаданные. Статистические пакеты для анализа данных.

Практическое занятие. Решение типовых задач статистической обработки данных.

Самостоятельная работа. Создание данных. Обслуживание данных. Синтез данных. Использование данных. Публикация данных. Архивация данных Постановка задач обучения по прецедентам. Объекты и признаки. Типы шкал: бинарные, номинальные, порядковые, количественные. Концептуальное обучение: пространство гипотез, поиск в пространстве гипотез, обучаемость, оценка качества решения задачи.

Рекомендуемая литература:

основная [1];

дополнительная [2].

Тема 2. Теория больших данных

Лекция. Большие данные. Системы управления большими данными.

Практическое занятие. Автоматизация процессов подготовки больших данных к анализу. Визуализация больших данных.

Самостоятельная работа. Архитектура системы обработки больших данных. Принятие решений на основе больших данных. Распределенные файловые системы. Распределенные фреймворки. Системы развертывания. Интеграция данных. Базы данных NoSQL и новые SQL базы данных. Прием данных. Сбор данных. Анализ данных. Представление результатов.

Рекомендуемая литература:

основная [1];

дополнительная [1].

Тема 3. Технологии анализа больших данных

Лекция. Параллельные алгоритмы для работы с данными. Программные платформы и системы для больших данных.

Самостоятельная работа. Оборудование для обработки больших данных. Языки программирования для статистической обработки данных и работы с графикой. Операторы Map и Reduce. Лямбда-архитектура. Системы управления потоками данных. Системы хранения больших данных. Обработка данных в реальном времени. Аналитические платформы. Оборудование для обработки больших данных.

Рекомендуемая литература:

основная [2];

дополнительная [1].

5. Методические рекомендации по организации изучения дисциплины

При реализации программы дисциплины используются используются лекционные и практические занятия.

Общими целями занятий являются:

- обобщение, систематизация, углубление, закрепление теоретических знаний по конкретным темам дисциплины;
- формирование умений применять полученные знания на практике, реализация единства интеллектуальной и практической деятельности;
- выработка при решении поставленных задач профессионально значимых качеств: самостоятельности, ответственности, точности, творческой инициативы.

Целями лекции являются:

- систематизированные основы научных знаний по дисциплине, раскрывать состояние и перспективы развития соответствующей области науки и техники;
- концентрировать внимание обучающихся на наиболее сложных и узловых вопросах;
- стимулировать их активную познавательную деятельность и способствовать формированию творческого мышления.

В ходе практического занятия обеспечивается процесс активного взаимодействия обучающихся с преподавателем; приобретаются практические

навыки и умения. Цели практического занятия: выработка практических умений и приобретение навыков, закрепление пройденного материала по соответствующей теме дисциплины.

Самостоятельная работа обучающихся направлена на углубление и закрепление знаний, полученных на лекциях и других занятиях, выработку навыков самостоятельного активного приобретения новых, дополнительных знаний, подготовку к предстоящим занятиям.

6. Оценочные материалы по дисциплине

Текущий контроль успеваемости обеспечивает оценивание хода освоения дисциплины, проводится в соответствии с содержанием дисциплины по видам занятий в форме опроса / тестирования.

Промежуточная аттестация обеспечивает оценивание промежуточных и окончательных результатов обучения по дисциплине, проводится в форме экзамена.

6.1. Примерные оценочные материалы:

6.1.1. Текущего контроля

Типовые вопросы для опроса:

1. В чем смысл феномена Big Data, и каково его влияние?
2. Иерархические алгоритмы. Иерархические образы. Представление результатов иерархического алгоритма.
3. Сложности и проблемы, которые могут возникнуть при применении кластерного анализа.
4. Анализ структурированной информации, хранящейся в базах данных.
5. Классификация и кластеризация текстовой информации.
6. Информационный поиск в текстах. Поиск по словарю.
7. Обработка запроса. Булева модель.
8. Новые алгоритмы и некоторые модификации алгоритмов кластерного анализа.
9. Каковы предпосылки возникновения Data Mining как отдельного направления в анализе данных?
10. В чем основная задача интеллектуального анализа данных?
11. Какие данные можно анализировать с использованием техник Data Mining?
12. Методы визуализации. Характеристика средств визуализации данных.
13. Визуализация инструментов метода анализа данных. Визуализация моделей.
14. Представление данных в одном, двух и трех измерениях. Представление данных в 4 + измерениях.
15. Представление пространственных характеристик. Основные тенденции в области визуализации.

16. Модули текстового анализа.
17. Классификация инструментов анализа данных.
18. Программное обеспечение анализа данных для поиска ассоциативных правил.
19. Практическое применение интеллектуального анализа данных.
20. Информационное хранилище(витрины данных, информационное хранилище двухуровневой и трехуровневой архитектуры.
21. Модели данных (реляционная, сетевая, иерархическая модели данных).
22. Концепция многомерного представления данных.
23. Методы извлечения знаний и области их применения в экономике.
24. Методы геометрических преобразований.
25. Концептуальное моделирование информационных потребностей в технологии Хранилищ данных.
26. Обзор архитектуры систем поддержки принятия решений.
27. Принципы построения и использования систем на основе технологии OLAP.
28. Методы анализа и обработки данных.

Типовые задания для тестирования:

1. Принятый способ представления данных: показатели должны быть:

- 1) по строкам;
- 2) по ячейкам;
- 3) по столбцам;
- 4) по диагонали.

2. Интервальные данные – это:

- 1) данные с интервалом;
- 2) количество измерений в каждом интервале;
- 3) данные об интервалах;
- 4) количество интервалов в каждом измерении.

2. Среди ниже приведённых нечисловые данные следующие:

- 1) баллы;
- 2) ранги;
- 3) дихотомические;
- 4) рейтинги.

3. Простейшие статистические характеристики – это:

- 1) среднее;
- 2) с.к.о.;
- 3) математическое ожидание;
- 4) дисперсия.

4. Следующие программы являются специализированными статистическими пакетами:

- 1) EXCEL;
- 2) GRAPHER;
- 3) SPSS;
- 4) STATISTICA.

5. Проверка статистической гипотезы включает в себя:

- 1) ранжирование;
- 2) вычисление эмпирического значения;
- 3) принятие уровня значимости;
- 4) вычисление критического значения.

6. Кластерный анализ предназначен для:

- 1) группировки объектов;
- 2) группировки показателей;
- 3) ранжирования объектов;
- 4) ранжирования показателей.

7. Опции кластерного анализа:

- 1) расстояние между группами;
- 2) расстояние между показателями;
- 3) расстояние между телами;
- 4) расстояние между объектами;

8. Кластерный анализ реализован в программах:

- 1) EXCEL;
- 2) SPSS;
- 3) AGRAPHER;
- 4) STATISTICA.

9. Снижение размерности это:

- 1) уменьшение числа измерений;
- 2) уменьшение числа объектов;
- 3) уменьшение числа показателей;
- 4) уменьшение числа знаков;

10. Компонентный анализ реализован в программах:

- 1) EXCEL;
- 2) SPSS;
- 3) AGRAPHER;
- 4) STATISTICA.

11. Методы, относящиеся к снижению размерности:

- 1) факторный анализ;

- 3) регрессия;
- 2) компонентный анализ;
- 4) корреляция.

12. Компонентный анализ позволяет:

- 1) сортировать;
- 2) группировать;
- 3) ранжировать;
- 4) упорядочивать.

13. Дихотомическая шкала это:

- 1) состоящая из “да” и “нет”;
- 2) состоящая из “истина” и “ложь”;
- 3) состоящая из двух чисел;
- 4) состоящая из двух рангов.

14. К нечисловым шкалам относятся:

- 1) номинальная;
- 2) интервалов;
- 3) абсолютная;
- 4) ранговая.

15. Существует шкал для описания данных:

- 1) 4;
- 2) 6;
- 3) 5;
- 4) 7.

16. Количество наблюдений - это:

- 1) размерность;
- 2) ширина;
- 3) объём выборки;
- 4) поверхность выборки.

17. Элементы таблицы сопряжённости называются:

- 1) координаты;
- 3) скорости;
- 2) длины;
- 4) частоты.

18. Методы анализа таблиц сопряжённости:

- 1) Критерий Розенбаума;
- 2) хи-квадрат;
- 3) Критерий Колмогорова-Смирнова;
- 4) критерий Фишера.

19. В ходе анализа таблицы сопряжённости выполняется:

- 1) проверка на соответствие;
- 3) проверка на непротиворечивость;
- 2) проверка на монотонность;
- 4) проверка на значимость.

20. Максимальная размерность таблицы сопряжённости может быть:

- 1) 3;
- 3) 5;
- 2) 10;
- 4) какая угодно.

21. Определение ранга пожара на основе анализа первичных данных является задачей ...

- 1) регрессии
- 2) классификации
- 3) кластеризации

6.1.2. Промежуточной аттестации

Примерный перечень вопросов, выносимых на экзамен

1. Понятие большие данные. Роль цифровой информации в 21 веке. Проблемы анализа и обработки большого объема данных.

2. Базовые принципы обработки больших данных.

3. Определение тиражирования знаний. Процесс построения модели.

4. Вопросы для определения базового уровня:

5. Технологии обработки больших данных: NoSQL, MapReduce, Hadoop, R.

6. Методика извлечения знаний Knowledge Discovery in Databases (KDD). Этапы KDD.

7. Data Mining. Постановка основных задач.

8. Машинное обучение. Бизнес-решения с помощью алгоритмов Data Mining.

9. Классификация ПО в области Data Mining и KDD. Типовая схема системы на базе аналитической платформы.

10. Понятие ассоциативного правила и транзакции. Определение поддержки и достоверности. Определение значимости и полезности ассоциативных правил, показатели их характеризующие.

11. Формальная постановка задачи кластеризации. Цели кластеризации.

12. Основные шаги алгоритма k-means. Условие остановки алгоритма k-means. Преимущества и недостатки алгоритма k-means.

13. Кластеризация с помощью самоорганизующейся карты Кохонена

14. Этапы проведения классификации. Обзор методов классификации и регрессии.

15. Задачи линейной и логистической регрессии.
16. Определение дерева решений. Структура дерева решений. Выбор атрибута разбиения в узле.
17. Алгоритм ID3.
18. Алгоритм C4.5.
19. Обучение с учителем и обучение без учителя.
20. Классы задач машинного обучения: регрессия, классификация, кластерный анализ.
21. Постановка задачи регрессионного анализа.
22. Парная линейная регрессия.
23. Множественная линейная регрессия.
24. Точечный и интервальный прогноз по модели регрессии.
25. Линейные алгоритмы классификации.
26. Логистическая регрессия.
27. Понятие о деревьях решений.
28. Метрики качества классификации
29. Постановка задачи кластерного анализа.
30. Метод K-средних

6.2. Шкала оценивания результатов промежуточной аттестации и критерии выставления оценок

Система оценивания включает:

Форма контроля	Показатели оценивания	Критерии выставления оценок	Шкала оценивания
экзамен	правильность и полнота ответа	дан правильный, полный ответ на поставленный вопрос, показана совокупность осознанных знаний по дисциплине, доказательно раскрыты основные положения вопросов; могут быть допущены недочеты, исправленные самостоятельно в процессе ответа.	отлично
		дан правильный, недостаточно полный ответ на поставленный вопрос, показано умение выделить существенные и несущественные признаки, причинно-следственные связи; могут быть допущены недочеты, исправленные с помощью преподавателя.	хорошо
		дан недостаточно правильный и полный ответ; логика и последовательность изложения имеют нарушения; в ответе отсутствуют выводы.	удовлетворительно
		ответ представляет собой разрозненные знания с существенными ошибками по вопросу; присутствуют фрагментарность, нелогичность	неудовлетворительно

		изложения; дополнительные и уточняющие вопросы не приводят к коррекции ответа на вопрос.	
--	--	--	--

7. Ресурсное обеспечение дисциплины

7.1. Лицензионное и свободно распространяемое программное обеспечение

Перечень лицензионного и свободно распространяемого программного обеспечения, в том числе отечественного производства:

Microsoft Windows 7 Professional – ПО-BE8-834 [Лицензионное]

Microsoft Office Standard 2010 – ПО-413-406 [Лицензионное]

7-Zip – ПО-F33-948 [Свободно распространяемое]

Adobe Acrobat Reader – ПО-F63-948 [Свободно распространяемое]

Google Chrome – ПО-F2C-926 [Свободно распространяемое]

МойОфис Образование – ПО-41В-124 [Свободно распространяемое - Отечественное]

7.2. Профессиональные базы данных и информационные справочные системы

Информационная справочная система — Сервер органов государственной власти Российской Федерации <http://россия.рф/> (свободный доступ); профессиональные базы данных — Портал открытых данных Российской Федерации <https://data.gov.ru/> (свободный доступ); федеральный портал «Российское образование» <http://www.edu.ru> (свободный доступ); система официального опубликования правовых актов в электронном виде <http://publication.pravo.gov.ru/> (свободный доступ); федеральный портал «Совершенствование государственного управления» <https://ar.gov.ru> (свободный доступ); электронная библиотека университета <http://elib.igps.ru> (авторизованный доступ); электронно-библиотечная система «ЭБС IPR BOOKS» <http://www.iprbookshop.ru> (авторизованный доступ).

7.3. Литература

Основная литература:

1. Адлер, Ю. П. Статистическое управление процессами. «Большие данные» : учебное пособие / Ю. П. Адлер, Е. А. Черных. — Москва : Издательский Дом МИСиС, 2016. — 52 с. — ISBN 978-5-87623-969-3. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: <https://www.iprbookshop.ru/64199.html>

2. Воронов, В. И. Data Mining - технологии обработки больших данных : учебное пособие / В. И. Воронов, Л. И. Воронова, В. А. Усачев. — Москва : Московский технический университет связи и информатики, 2018. — 47 с. — ISBN 2227-8397. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: <https://www.iprbookshop.ru/81324.html>

Дополнительная литература:

1. Железнов, М. М. Методы и технологии обработки больших данных : учебно-методическое пособие / М. М. Железнов. — Москва : МИСИ-МГСУ, ЭБС АСВ, 2020. — 46 с. — ISBN 978-5-7264-2193-3. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: <https://www.iprbookshop.ru/101802.html>
2. Воронова, Л. И. Machine Learning: регрессионные методы интеллектуального анализа данных : учебное пособие / Л. И. Воронова, В. И. Воронов. — Москва : Московский технический университет связи и информатики, 2018. — 82 с. — ISBN 2227-8397. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: <https://www.iprbookshop.ru/81325.html>

7.4 Материально-техническое обеспечение дисциплины

Для проведения и обеспечения занятий используются помещения, которые представляют собой учебные аудитории для проведения учебных занятий, предусмотренных программой магистратуры, оснащенные оборудованием и техническими средствами обучения: автоматизированное рабочее место преподавателя, маркерная доска, мультимедийный проектор, документ-камера, посадочные места обучающихся.

Помещения для самостоятельной работы обучающихся оснащены компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа к электронной информационно-образовательной среде университета.

Авторы: канд. техн. наук, доцент Матвеев А.В., канд. техн. наук, доцент Максимов А.В.