МИНОБРНАУКИ РОССИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ НАУКИ
ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР
«КОЛЬСКИЙ НАУЧНЫЙ ЦЕНТР РОССИЙСКОЙ АКАДЕМИИ НАУК»
(ФИЦ КНЦ РАН)

МЕТОДИЧЕСКИЕ УКАЗАНИЯ К ВЫПОЛНЕНИЮ САМОСТОЯТЕЛЬНЫХ РАБОТ

По дисциплине Б1.В.03 Проблемно-ориентированные информационные системы
указывается цикл (раздел) ОП, к которому относится дисциплина, название дисциплины

для направления подготовки (специальности) 09.04.02 Информационные системы и технологии
код и наименование направления подготовки (специальности)

направленность программы (профиль) Информационные системы предприятий и учреждений
наименование профиля /специализаций/образовательной программы

Квалификация выпускника, уровень подготовки
Магистр
указывается квалификация (степень) выпускника в соответствии с ФГОС ВО

Апатиты

2020

Лист согласования

1 Разработчик:

доцент	УАиМ		Н.А. Тоичкин
должность		подпись	И.О. Фамилия

2 Методические указания рассмотрены и одобрены на заседании учебно-методической комиссии управления аспирантуры и магистратуры 29 июня 2020 г., протокол № 02.

Председатель УМК УАиМ

29.06.2020		Л.Д. Кириллова
дата	подпись	И.О.Фамилия

Пояснительная записка

1. **Методические указания** составлены в соответствии с требованиями федерального государственного образовательного стандарта по образовательной программе высшего образования – программе магистратуры по направлению подготовки 09.04.02 Информационные системы и технологии, утвержденного приказом Минобрнауки России от 19.09.2017 № 917.

2. **Цель дисциплины:** изучение современных технологий анализа информации и методов машинного обучения и их применение на практике.

Задачи дисциплины:

- изучить основные методы и алгоритмы машинного обучения;
- получить навыки применения алгоритмов машинного обучения в задачах анализа информации;
- осуществлять математическую и информационную постановку задач по обработке информации.

3. Требования к уровню подготовки обучающегося в рамках данной дисциплины.

Процесс изучения дисциплины (модуля) «Проблемно-ориентированные информационные системы» направлен на формирование элементов следующих компетенций в соответствии с ФГОС ВО 09.04.02 Информационные системы и технологии (уровень магистратуры), представленных в таблице 1.

Таблица 1 – Компетенции, формируемые в процессе изучения дисциплины «Проблемно-ориентированные информационные системы**»**

№ п/п	Код компетенции	Содержание компетенции
1.	ПК-1	Способен проводить экспертизу и оказывать информационно-аналитическую поддержку в решении профессиональных задач в научной деятельности.

4. Планируемые результаты обучения по дисциплине (модулю) «Проблемно-ориентированные информационные системы».

Результаты формирования компетенций и обучения представлены в таблице 2.

Таблица 2 – Планируемые результаты обучения

№ п/п	Код компетенции	Компоненты компетенции, степень их реализации	Результаты обучения

1.	ПК-1	Компоненты компетенции соотносятся с содержанием дисциплины и компетенция реализуется полностью.	***знать*** – формализацию задачи машинного обучения; – понятие больших данных и их свойства; – постановку задачи – классификации и регрессии; – понятие обобщенного метрического классификатора; – алгоритмы метрической классификации; – основные принципы построения логических алгоритмов классификации; – алгоритм построения дерева классификации ID 3; – линейные методы классификации. ***уметь*** – использовать алгоритмы обработки информации для различных приложений; – выполнять постановку задачи машинного обучения; – выбирать методы и средства для решения задач машинного обучения;
			владеть – инструментальными средствами решения задач машинного обучения; – методами интеллектуального анализа информации.

ПЕРЕЧЕНЬ РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ

Основная литература:

1. Архипенков С. Я. , Голубев Д. , Максименко О. Хранилища данных: от концепции до внедрения. М.: Диалог-МИФИ, 2002, 528 с. Режим доступа: https://biblioclub.ru/index.php?page=book_red&id=89285&sr=1
2. Чубукова И. А. Data Mining. М.: Интернет-Университет Информационных Технологий, 2008, 383 с. Режим доступа: https://biblioclub.ru/index.php?page=book_red&id=233055&sr=1

Дополнительная литература:

3. Введение в анализ данных с помощью Pandas. Режим доступа: https://habrahabr.ru/post/196980/

СОДЕРЖАНИЕ ПРОГРАММЫ И МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ

Тема 1. Большие данные и машинное обучение. Большие данные. Свойства больших данных. Машинное обучение. Типы задач машинного обучения. Модель алгоритмов. Метод обучения. Этап обучения и этап применения. Функционалы качества. Сведение задачи обучения к задаче оптимизации. Переобучение и обобщение. Пример переобучения (Рунге). Эмпирические оценки обобщающей способности..

Рекомендуемая литература: [1], [2], [3].

Вопросы для самоконтроля знаний:

1. Свойства больших данных.
2. Формализация задачи машинного обучения.
3. Признаковое описание объекта.
4. Примеры задач машинного обучения: задачи классификации и регрессии; задачи ранжирования.
5. Эксперименты в машинном обучении: эксперименты на реальных и синтетических данных
6. Использование Phyton для анализа данных. Дистрибутив Anaconda.
7. Основные библиотеки Phyton: Scikit-learn, NumPy, SciPy, matplotlib, pandas.
8. Pandas: базовые методы, индексация и извлечение данных, применение функций к ячейкам, столбцам и строкам.
9. Pandas: группировка данных, таблицы сопряженности, сводные таблицы.
10. DataFrame в Pandas.
11. Как выполняется загрузка данных в DataFrame в библиотеке Pandas?
12. Как выполняется доступ к столбцам DataFrame в библиотеке Pandas?

Тема 2. Метрические методы классификации

Формализация задачи. Обобщенный метрический классификатор. Метод ближайшего соседа. Метод k взвешенных ближайших соседей. Метод парзеновского окна. Метод потенциальных функций. Отбор эталонных объектов. Понятие отступа объекта. Типы объектов в зависимости от отступа. Отбор эталонов, алгоритм STOLP. Задача выбора метрики. Жадное добавление признаков.

Рекомендуемая литература: [2], [3].

Вопросы для самоконтроля знаний:

1. Метрические методы классификации.Признаковые описания объекта.
1. Гипотеза компактности. В чем идея гипотезы компактности?
2. Метрики, виды метрик. Приведите формулу Метрики Минковского. Что является ее параметром?
2. Весовая Евклидова метрика, метрика Минковского.
3. Масштабирование признаков.
3. Метод k ближайших соседей. Реализация kNN в классе sklearn.neighbors.KNeighborsClassifier. Какой параметр метода k ближайших соседей, задает число соседей для построения прогноза?
4. В каком классе Scikit-learn реализован метод kNN?
4. Кросс-валидация. Алгоритм выполнения кросс-валидации по блокам. В чем смысл кросс-валидации?
5. Вычисление ошибки на разбиениях.
6. Как вычисляется весовая Евклидова метрика?

Тема 3. Логические методы классификации

Логическая закономерность. Основы вопросы построения логических алгоритмов классификации. Виды закономерностей. Критерии информативности: простые критерии, статистический критерий, энтропийный критерий. Где находятся закономерности в *(p, n)*-плоскости. Схема локального поиска информативных закономерностей. Определение бинарного решающего дерева. Жадный алгоритм построения дерева ID 3. Варианты критериев ветвления в ID 3. Обработка пропусков, алгоритм обработки пропусков на этапе обучения и этапе классификации. Алгоритм ID3: достоинства и недостатки. Стратегии редукции решающих деревьев. Небрежные решающие деревья. Бинаризация вещественного признака.

Рекомендуемая литература: [2].

Вопросы для самоконтроля знаний:

1. Логическая закономерность.
2. Основы вопросы построения логических алгоритмов классификации.
3. Определение бинарного решающего дерева.
4. Реализация решающих деревьев в библиотеке scikit-learn.
5. Важность признаков.
6. Пропуски в данных.
7. В каких классах scikit-learn реализуются решающие деревья для задач классификации и регрессии?
8. С помощью какой функции scikit-learn реализуется обучение модели решающих деревьев?

Тема 4. Линейные методы классификации.

Задача построения разделяющей поверхности. Задача построения разделяющей поверхности. Минимизация эмпирического риска. Непрерывные аппроксимации пороговой функции потерь. Линейный классификатор. Персептрон. Устройство нервной клетки. Линейная модель нейрона МакКаллока-Питтса. Алгоритм Stohastic Gradient. Дельта-правило ADALINE. Правило Хебба. SG: инициализация весов. SG: проблемы переобучения. Принцип максимума правдоподобия. Оптимальная разделяющая гиперплоскость. Метод

SVM. Нелинейное обобщение SVM.

Рекомендуемая литература: [2], [1].

Вопросы для самоконтроля знаний:

1. Линейные алгоритмы классификации. Сформулируйте постановку задачи линейной классификации.
2. Как выполняется стандартизация признаков?
3. Персептрон. В каком классе scikit-learn реализуется персептрон?
4. Нормализация признаков. Стандартизация признаков. Каким классом удобно воспользоваться для стандартизации признаков?
5. Для чего используется функция sklearn.metrics.accuracy_score?
6. Реализация линейных классификаторов в библиотеке scikit-learn.
7. Метрика качества.
8. Метод опорных векторов.
9. Опорные объекты. Какие объекты называют опорными?
10. На что направлен функционал, который он оптимизирует метод опорных векторов?

КОНТРОЛЬНЫЕ ВОПРОСЫ

Итоговый уровень знаний обучающихся, приобретенный при изучении дисциплины «Проблемно-ориентированные информационные системы», проверяется на зачете.

1. Основные понятия – информация, данные, знания. Виды информации. Обработка данных и ее виды. Data Mining. Классификация задач Data Mining.
2. Модели процессов обработки данных. Модель: конечные автоматы.
3. Модели процессов обработки данных. Модель: сети Петри.
4. Задачи обработки данных различных типов. Прикладные области обработки данных. Оцифровка сигналов. Теорема Котельникова.
5. Базы данных. OLTP – системы. Неэффективность OLTP для анализа данных. Определение и свойства хранилищ данных.
6. Физические и виртуальные хранилища данных (ХД). Основные проблемы создания ХД.
7. Витрины данных.
8. Данные в хранилищах данных. ETL процесс.
9. Представление данных в виде гиперкуба. Операции над гиперкубом. Пример. Технология OLAP. Тест FASMI.
10. Многомерное представление данных и многомерный куб. Представление данных в виде гиперкуба. Пример.
11. Основные понятия гиперкубов (OLAP кубов). Структура OLAP куба. Операции над гиперкубом.
12. Архитектура OLAP. Компоненты OLAP. MOLAP, ROLAP, HOLAP.
13. Задача анализа текстов. Этапы анализа. Предобработка текста.
14. Извлечение ключевых понятий из текста.
15. Классификация текстовых документов. Методы классификации текстовых документов.
16. Большие данные. Свойства больших данных.
17. Машинное обучение, формализация задачи машинного обучения.
18. Признаковое описание объекта. Ответы и типы задач машинного обучения. Модель алгоритмов. Метод обучения. Этап обучения и этап применения.
19. Функционалы качества. Сведение задачи обучения к задаче оптимизации.
20. Переобучение и обобщение. Пример переобучения (Рунге). Эмпирические оценки обобщающей способности.
21. Примеры задач машинного обучения: задачи классификации.
22. Примеры задач машинного обучения: задачи регрессии.
23. Примеры задач машинного обучения: задача ранжирования.
24. Эксперименты в машинном обучении: эксперименты на реальных и синтетических данных.
25. Формализация метрической классификации. Обобщенный метрический классификатор.
26. Метод ближайшего соседа.
27. Метод k взвешенных ближайших соседей.
28. Метод парзеновского окна.
29. Метод потенциальных функций.
30. Отбор эталонных объектов. Понятие отступа объекта. Типы объектов в зависимости от отступа.
31. Отбор эталонов, алгоритм STOLP.
32. Логическая закономерность. Основы вопросы построения логических алгоритмов классификации. Виды закономерностей.

33. Критерии информативности: простые критерии, статистический критерий, энтропийный критерий. Схема локального поиска информативных закономерностей.
34. Определение бинарного решающего дерева. Жадный алгоритм построения дерева ID 3.
35. Варианты критериев ветвления в ID 3.
36. Алгоритм ID3: достоинства и недостатки.
37. Обработка пропусков в ID 3, алгоритм обработки пропусков на этапе обучения и этапе классификации.
38. Стратегии редукции решающих деревьев.
39. Небрежные решающие деревья.
40. Бинаризация вещественного признака.

Рекомендуемая литература: [1], [2], [3]