МИНОБРНАУКИ РОССИИ

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

МЕТОДИЧЕСКИЕ УКАЗАНИЯ К ВЫПОЛНЕНИЮ ПРАКТИЧЕСКИХ 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.

Таблица 3 - Перечень практических работ

№ п/п	Наименование практических работ	Количество часов	Наименование темы по табл. 4
1.	Предобработка данных в Pandas	2	1
2.	Метрические методы классификации в Scikit-learn	4	2
3.	Деревья решений. Важность признаков	4	3
4.	Линейная классификация. Нормализация признаков	4	4
Итого часов		**14**	

Рекомендации к выполнению практических работ

Практическая работа № 1. Предобработка данных в Pandas

1. Анализ данных по доходу населения UCI Adult. В задании предлагается с помощью Pandas ответить на несколько вопросов по данным репозитория UCI Adult. Список вопросов:
    - Каков средний возраст (признак age) женщин?
    - Какова доля граждан Германии (признак native-country)?
    - Постройте гистограмму распределения (bar plot) образования людей (признак education).
    - Каковы средние значения и среднеквадратичные отклонения возраста тех, кто получает более 50K в год (признак salary) и тех, кто получает менее 50K в год?
    - Правда ли, что люди, которые получают больше 50k, имеют как минимум высшее образование? (признак education - Bachelors, Prof-school, Assoc-acdm, Assoc-voc, Masters или Doctorate)
    - Выведите статистику возраста для каждой расы (признак race) и каждого пола. Используйте groupby и describe. Найдите таким образом максимальный возраст мужчин расы Amer-Indian-Eskimo.
    - Среди кого больше доля зарабатывающих много (>50K): среди женатых или холостых мужчин (признак marital-status)? Женатыми считаем тех, у кого marital-status начинается с Married (Married-civ-spouse, Married-spouse-absent или Married-AF-spouse), остальных считаем холостыми.
    - Какое максимальное число часов человек работает в неделю (признак hours-per-week)? Сколько людей работают такое количество часов и каков среди них процент зарабатывающих много?
    - Посчитайте среднее время работы (hours-per-week) зарабатывающих мало и много (salary) для каждой страны (native-country).
2. Анализ данных по пассажирам Титаника. В задании предлагается с помощью Pandas ответить на несколько вопросов по данным репозитория UCI Titanic. Список вопросов:
    - Какое количество мужчин и женщин ехало на корабле?
    - Какой части пассажиров удалось выжить? Посчитайте долю выживших пассажиров.
    - Какую долю пассажиры первого класса составляли среди всех пас- сажиров?
    - Какого возраста были пассажиры?
    - Посчитайте среднее и медиану возраста пассажиров.
    - Коррелируют ли число братьев/сестер с числом родителей/детей? Посчитайте корреляцию Пирсона между признаками SibSp и Parch.
    - Какое самое популярное женское имя на корабле? Извлеките из полного имени пассажира (колонка Name) его личное имя (First Name).

Практическая работа № 2. Метрические методы классификации в Scikit-learn.
Задание 1.

1. В этом задании нужно подобрать оптимальное значение k для алгоритма kNN. Будем использовать набор данных Wine, где требуется предсказать сорт винограда, из которого изготовлено вино, используя результаты химических анализов.
2. Выполните следующие шаги:

- Загрузите выборку Wine по адресу https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data
- Извлеките из данных признаки и классы. Класс записан в первом столбце (три варианта), признаки — в столбцах со второго по последний. Более подробно о сути признаков можно прочитать по адресу https://archive.ics.uci.edu/ml/datasets/Wine
- Оценку качества необходимо провести методом кроссвалидации по 5 блокам (5-fold). Создайте генератор разбиений, который перемешивает выборку перед формированием блоков (shuffle=True). Для воспроизводимости результата, создавайте генератор KFold с фиксированным параметром random_state=42. В качестве меры качества используйте долю верных ответов (accuracy).
- Найдите точность классификации на кросс-валидации для метода k ближайших соседей (sklearn.neighbors.KNeighborsClassifier), при k от 1 до 50. При каком k получилось оптимальное качество? Чему оно равно (число в интервале от 0 до 1)?
- Произведите масштабирование признаков с помощью функции sklearn.preprocessing.scale. Снова найдите оптимальное k на кросс-валидации.
- Какое значение k получилось оптимальным после приведения признаков к одному масштабу? Как изменилось значение качества? Приведите ответы на вопросы.

Задание 2.

1. Нам понадобится решать задачу регрессии с помощью метода k ближайших соседей — воспользуемся для этого классом sklearn.neighbors.KNeighborsRegressor.
2. Метрика задается с помощью параметра metric, нас будет интересовать значение ’minkowski’. Параметр метрики Минковского задается с помощью параметра p данного класса.
3. Инструкция по выполнению

- Мы будем использовать в данном задании набор данных Boston, где нужно предсказать стоимость жилья на основе различных характеристик расположения (загрязненность воздуха, близость к дорогам и т.д.). Подробнее о признаках можно почитать по адресу https://archive.ics.uci.edu/ml/datasets/Housing
- Загрузите выборку Boston с помощью функции sklearn.datasets.load_boston(). Результатом вызова данной функции является объект, у которого признаки записаны в поле data, а целевой вектор — в поле target.
- Приведите признаки в выборке к одному масштабу при помощи функции sklearn.preprocessing.scale.
- Переберите разные варианты параметра метрики p по сетке от 1 до 10 с таким шагом, чтобы всего было протестировано 200 вариантов (используйте функцию numpy.linspace). Используйте KNeighborsRegressor с n_neighbors=5 и weights=’distance’ - данный параметр добавляет в алгоритм веса, зависящие от расстояния до ближайших соседей. В качестве метрики качества используйте среднеквадратичную ошибку (параметр scoring=’mean_squared_error’ у cross_val_score; при использовании библиотеки scikit-learn версии 18.0.1 и выше необходимо указывать scoring=’neg_mean_squared_error’). Качество оценивайте, как и в предыдущем задании, с помощью кросс-валидации по 5 блокам с random_state = 42, не забудьте включить перемешивание выборки (shuffle=True).
- Определите, при каком p качество на кросс-валидации оказалось оптимальным. Обратите внимание, что cross_val_score возвращает массив показателей качества по

блокам; необходимо макет массив показателей качества по блокам; необходимо максимизировать среднее этих показателей.

Практическая работа № 3. Деревья решений. Важность признаков

План:

1. Загрузите выборку из файла titanic.csv с помощью пакета Pandas.
2. Оставьте в выборке четыре признака: класс пассажира (Pclass), цену билета (Fare), возраст пассажира (Age) и его пол (Sex).
3. Обратите внимание, что признак Sex имеет строковые значения.
4. Выделите целевую переменную — она записана в столбце Survived.
5. В данных есть пропущенные значения — например, для некоторых пассажиров неизвестен их возраст. Такие записи при чтении их в pandas принимают значение nan. Найдите все объекты, у которых есть пропущенные признаки, и удалите их из выборки.
6. Обучите решающее дерево с параметром random_state=241 и остальными параметрами по умолчанию.
7. Вычислите важности признаков и найдите два признака с наибольшей важностью. Их названия будут ответами для данной задачи (в качестве ответа укажите названия признаков через запятую без пробелов).

Практическая работа № 4. Линейная классификация. Нормализация признаков.

Задание 1.

1. Загрузите обучающую и тестовую выборки из файлов perceptrontrain.csv и perceptron-test.csv. Целевая переменная записана в первом столбце, признаки — во втором и третьем.
2. Обучите персептрон со стандартными параметрами и random_state=241.
3. Подсчитайте качество (долю правильно классифицированных объектов, accuracy) полученного классификатора на тестовой выборке.
4. Нормализуйте обучающую и тестовую выборку с помощью класса StandardScaler.
5. Обучите персептрон на новых выборках. Найдите долю правильных ответов на тестовой выборке.
6. Найдите разность между качеством на тестовой выборке после нормализации и качеством до нее.

Задание 2.

1. Загрузите выборку из файла svm-data.csv. В нем записана двумерная выборка (целевая переменная указана в первом столбце, признаки — во втором и третьем).
2. Обучите классификатор с линейным ядром, параметром C=100000 и random_state=241. Такое значение параметра нужно использовать, чтобы убедиться, что SVM работает с выборкой как с линейно разделимой. При более низких значениях параметра алгоритм будет настраиваться с учетом слагаемого в функционале, штрафующего за маленькие отступы, из-за чего результат может не совпасть с решением классической задачи SVM для линейно разделимой выборки.
3. Найдите номера объектов, которые являются опорными (нумерация с единицы).

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/