Приложение 1

МИНОБРНАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Институт прикладной математики и компьютерных наук

УТВЕРЖДАЮ
Директор института прикладной
математики и компьютерных наук
_____ А.В. Замятин
«__11__» _____ноября_____ 2021 г.

Фонд оценочных средств по дисциплине

Прикладные аспекты машинного обучения

Направление подготовки
02.03.02 Фундаментальная информатика и информационные технологии
код и наименование направления подготовки

Искусственный интеллект и разработка программных продуктов
наименование профиля подготовки

Томск–2021

ФОС составил:
канд. техн. наук, доцент,
доцент кафедры теоретических основ информатики С.В. Аксёнов

Рецензент:
канд. техн. наук, доцент,
доцент кафедры теоретических основ информатики О.В. Марухина

Фонд оценочных средств одобрен на заседании учебно-методической комиссии института прикладной математики и компьютерных наук (УМК ИПМКН)

Протокол от 17 июня 2021 г. № 05

Председатель УМК ИПМКН,
д-р техн. наук, профессор С.П. Сущенко

Фонд оценочных средств (ФОС) является элементом системы оценивания сформированности компетенций у обучающихся в целом или на определенном этапе ее формирования.

ФОС разрабатывается в соответствии с рабочей программой (РП) дисциплины и включает в себя набор оценочных материалов для проведения текущего контроля успеваемости и промежуточной аттестации по дисциплине.

1. Компетенции и результаты обучения, формируемые в результате освоения дисциплины

Компетенция	**Индикатор компетенции**	**Код и наименование результатов обучения** (планируемые результаты обучения, характеризующие этапы формирования компетенций)
ПК-3. Способен осуществлять научно-исследовательские и опытно-конструкторские разработки как при исследовании самостоятельных тем, так и разработки по тематике организации	ИПК-3.1. Осуществляет проведение работ по обработке и анализу научно-технической информации и результатов исследований	ОР-3.1.1. Обучающийся будет знать: - процедуры выявления, формирования и согласования требований к результатам аналитических работ с применением технологий искусственного интеллекта и больших данных; - принципы планирования и организации аналитических работ с использованием технологий искусственного интеллекта и больших данных; ОР-3.1.2. Обучающийся сможет: - подготавливать данные для проведения аналитических работ по исследованию больших данных методами искусственного интеллекта; - проводить аналитическое исследование и разрабатывать приложения с применением технологий искусственного интеллекта и больших данных в соответствии с требованиями заказчика;

2. Этапы формирования компетенций и виды оценочных средств

№	Этапы формирования компетенций (разделы дисциплины)	Код и наименование результатов обучения	Вид оценочного средства (тесты, задания, кейсы, вопросы и др.)
1.	Раздел 1. Разведочный анализ данных. Основные положения нейросетевых вычислений. Выполнение лабораторной работы № 1 (Предварительный анализ данных)	ОР-3.1.1 ОР-3.1.2	Опрос на занятиях, подготовка к лабораторным занятиям, публичная защита лабораторной работы № 1.
2.	Раздел 2. Классификационные и регрессионные модели. Настройка архитектуры и алгоритмы настройки нейронных сетей встречного распространения. Выполнение лабораторной работы № 2 (Построение регрессора)	ОР-3.1.1 ОР-3.1.2	Опрос на занятиях, подготовка к лабораторным занятиям, публичная защита лабораторной работы № 2.
3	Раздел 3. Ансамбли моделей. Оптимизаторы обучения нейронных сетей. Выполнение лабораторной работы № 3 (Построение классификаторов)	ОР-3.1.1 ОР-3.1.2	Опрос на занятиях, подготовка к лабораторным занятиям, публичная защита лабораторной работы № 3.
4	Раздел 4. Работа с признаковым пространством. Нейронные сети с обратными связями. Выполнение лабораторной работы № 4 (Балансировка выборки и ROC)	ОР-3.1.1 ОР-3.1.2	Опрос на занятиях, подготовка к лабораторным занятиям, публичная защита лабораторной работы № 4.
5	Раздел 5 Основы нейросетевых вычислений. Сверточные нейронные сети и автоэнкодеры. Выполнение лабораторной работы № 5 (Работа с признаковым пространством)	ОР-3.1.1 ОР-3.1.2	Опрос на занятиях, подготовка к лабораторным занятиям, публичная защита лабораторной работы № 5.
6	Раздел 6. Обработка изображений сверточными нейронными сетями. Нейронные сети, обучающиеся без учителя и с подкреплением. Выполнение лабораторной работы № 6 (Определение важности признаков)	ОР-3.1.1 ОР-3.1.2	Опрос на занятиях, подготовка к лабораторным занятиям, публичная защита лабораторной работы № 6.
7	Раздел 7. Автокодировщики. Визуализация и объяснимость нейросетевых моделей. Выполнение лабораторной работы № 7 (Кластеризация данных и оценка её качества)	ОР-3.1.1 ОР-3.1.2	Опрос на занятиях, подготовка к лабораторным занятиям, публичная защита лабораторной работы № 7.
8	Раздел 8. Анализ сигналов и временных рядов. Хранение ассоциаций и управление памятью в нейросетевых моделях. Выполнение лабораторной работы № 8 (Обработка естественного языка (классификация текстов)).	ОР-3.1.1 ОР-3.1.2	Опрос на занятиях, подготовка к лабораторным занятиям, публичная защита лабораторной работы № 8.
9	Раздел 9. Использование генеративно- конкурирующих моделей. Основы генеративно-конкурирующих моделей.	ОР-3.1.1 ОР-3.1.2	Опрос на занятиях, подготовка к лабораторным занятиям, публичная защита лабораторной работы № 9.
10	Раздел 10. Практические аспекты	ОР-3.1.1	Опрос на занятиях, подготовка к

№	Этапы формирования компетенций (разделы дисциплины)	Код и наименование результатов обучения	Вид оценочного средства (тесты, задания, кейсы, вопросы и др.)
	использования обучения с подкреплением. Обучение с подкреплением.	ОР-3.1.2	лабораторным занятиям, публичная защита лабораторной работы № 10.
11	Промежуточная аттестация (по результатам выполнения лабораторных работ (min 70%) и презентации индивидуального проекта -2-3 мин/чел.)	ОР-3.1.1 ОР-3.1.2	Публичное представление и защита результатов индивидуального проекта.

3. Типовые контрольные задания или иные материалы, необходимые для оценки образовательных результатов обучения

3.1. Типовые задания для проведения текущего контроля успеваемости по дисциплине «Прикладные аспекты машинного обучения».

1. Предварительный анализ данных

Написать программу на Python, которая загружает набор данных, производит исследовательский анализ этих данных и визуализирует ряд зависимостей между признаками в нижеперечисленных вариациях с помощью библиотек matplotlib и sns.

Для своего варианта анализа необходимо посмотреть последнюю цифру номера своей зачетной книжки (или студенческого билета) и выполнить следующие корректировки:

• если последняя цифра 0 или 5: датасет – Лесные пожары (https://archive.ics.uci.edu/ml/datasets/Forest+Fires), предсказываемое значение – площадь пожара (Area);

• если последняя цифра 1 или 6: датасет – Качество вина (https://archive.ics.uci.edu/ml/datasets/Wine+Quality) предсказываемое значение – качество (Quality), для датасета с красным вином, winequality-red.csv;

• если последняя цифра 2 или 7: датасет – Качество вина (https://archive.ics.uci.edu/ml/datasets/Wine+Quality) предсказываемое значение – качество (Quality), для датасета с белым вином, winequality-white.csv;

• если последняя цифра 3 или 8: датасет – Аренда велосипедов (https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset), предсказываемое значение – количество аренд велосипедов в сутки (Area), датасет day.csv;

• если последняя цифра 4 или 9: датасет – Аренда велосипедов (https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset), предсказываемое значение – количество аренд велосипедов в час (Area), датасет hour.csv;

1. Получить описание набора данных и список атрибутов, получить число пропущенных значений в для каждого атрибута.

2. Обработать пропущенные значения (удалить строки/ удалить атрибуты/ выполнить импутацию значений)

4. Построить pairplot для набора данных

Для каждого изображения в заданиях ниже добавить легенду и подпись

5. Выбрать не менее 2-х признаков с неким распределением значений, и отобразить это распределение с помощью гистограмм hist и kdeplot, jointplot

6. Выбрать часть признаков датасета и отобразить корреляцию между ними в виде тепловой карты (heatmap). После этого построить тепловую карту, которые будут отображать лишь высокие значения прямой и обратной корреляции.

7. Выбрать 3 признака (имеющие некое распределение значений), целевую переменную (категориальную) и отобразить для них графики размаха (ящик с усами, box plot).

8. Визуализировать некоторые статистики, для разных атрибутов с использованием следующих инструментов: violinplot, countplot, FacetGrid, stripplot, swarmplot, catplot, pie.

Напишите короткое заключение о наиболее интересных зависимостях, которые Вы обнаружили в данных.

2. Построение регрессора

Написать программу на Python, которая обучает три регрессионных модели, построенных на наборе с помощью трёх алгоритмов: линейный регрессор, полиномиальный регрессор и регрессор, основанный на случайном лесе.
Выбрать признаки, использующиеся при обучении, и, если необходимо, выполнить их предобработку. Разделить выборку на обучающую и тестовую.
В работе необходимо исследовать работу алгоритма случайный лес с разными значениями гиперпараметров и степенью полинома для полиномиальной регрессионной модели.
Для модели случайный лес вывести значения важности признаков.
Написать короткий отчет по работе, включив в него программу с комментариями, значения качества моделей (коэффициент детерминации, среднюю квадратичную и среднюю абсолютную ошибки).
Выбрать наилучшую модель из полученных регрессоров.

Для своего варианта необходимо посмотреть последнюю цифру номера своей зачетной книжки (или студенческого билета) и выполнить следующие корректировки:
• если последняя цифра 0 или 5: датасет – Лесные пожары (https://archive.ics.uci.edu/ml/datasets/Forest+Fires), предсказываемое значение – площадь пожара (Area);
• если последняя цифра 1 или 6: датасет – Качество вина (https://archive.ics.uci.edu/ml/datasets/Wine+Quality) предсказываемое значение – качество (Quality), для датасета с красным вином, winequality-red.csv;
• если последняя цифра 2 или 7: датасет – Качество вина (https://archive.ics.uci.edu/ml/datasets/Wine+Quality) предсказываемое значение – качество (Quality), для датасета с белым вином, winequality-white.csv;
• если последняя цифра 3 или 8: датасет – Аренда велосипедов (https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset), предсказываемое значение – количество аренд велосипедов в сутки (Area), датасет day.csv;
• если последняя цифра 4 или 9: датасет – Аренда велосипедов (https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset), предсказываемое значение – количество аренд велосипедов в час (Area), датасет hour.csv;

3. Построение классификаторов

Написать программу на Python, которая загружает набор данных и выполняет задачи построения классификаторов, подбора гиперпараметров моделей и анализа качества работы классифицирующих моделей.

Выборка для классификатора Covertype Data Set (https://archive.ics.uci.edu/ml/datasets/Covertype).

Для этого необходимо посмотреть последнюю цифру номера своей зачетной книжки (или студенческого билета) и выполнить следующие корректировки:

Метка класса – Cover_Type. Так как необходимо создать бинарные классификаторы, а возможных классов – 7, то сначала необходимо изменить значение метки Cover_Type.

Для этого необходимо посмотреть последнюю цифру номера своей зачетной книжки (или студенческого билета) и выполнить следующие корректировки: если последняя цифра 0 или 5: метку 0 заменить на класс A, метки 1, 2, 3, 4 заменить на класс B;

если последняя цифра 1 или 6: метку 1 заменить на класс A, метки 0, 2, 3, 4 заменить на класс B;

если последняя цифра 2 или 7: метку 2 заменить на класс A, метки 0, 1, 3, 4 заменить на класс B;

если последняя цифра 3 или 8: метку 3 заменить на класс A, метки 0, 1, 2, 4 заменить на класс B;

если последняя цифра 4 или 9: метку 4 заменить на класс A, метки 0, 1, 2, 3 заменить на класс B.

1. Выполнить предварительную обработку набора данных.

2. Построить классифицирующие модели с использованием GridSearch для алгоритмов RandomForest, XGBoost, LogisticRegression и SVC. Провести эксперименты с регуляризацией.

Выберите лучшее сочетание алгоритма и гиперпараметров, сделайте выводы и отразите проделанную работу в отчёте.

4. Балансировка выборки и ROC

В качестве исследуемых данных берется набор из предыдущей лабораторной работы (ЛР №3).

Ход работы:

1. Произвести изменение количества объектов разных классов для получения трёх сбалансированных выборок с использованием разных методов перебалансировки и выполнить шаг 2 из предыдущей работы (Построить классифицирующие модели с использованием GridSearch для алгоритмов RandomForest, XGBoost, LogisticRegression и SVC. Провести эксперименты с регуляризацией.) для новых выборок.

2. Выполнить K-блочную стратифицированную проверку для указанных алгоритмов с значениями гиперпараметров, полученных ранее для сбалансированных и первоначальной выборок.

3. Визуализировать результаты ROC-анализа. Рассчитать среднюю ROC и отклонение (standard deviation).

Выберите лучшее сочетание алгоритма и гиперпараметров, сделайте выводы и отразите проделанную работу в отчёте.

5. Работа с признаковым пространством

1.Для выполнения задания необходимо загрузить набор данных с репозитория https://archive.ics.uci.edu/ml/datasets.php для задачи любой задачи (регрессии, кластеризации или классификации), который содержит как минимум 350 объектов в выборке и количество числовых атрибутов, которые подлежат анализу не менее 12.

2.Построить модели PCA, T-SNE, SOM для выбранных данных.

3.Для PCA получить значения выборочной дисперсии, процента дисперсии компонент, кумулятивного процента дисперсии, формулы для получения главных компонент, 20 первых векторов новых признаков главных компонент для исследуемой выборки. Для SOM разброс по кластерам, а также 20 первых векторов новых признаков.

6. Определение важности признаков

В качестве исследуемых данных берется набор из предыдущей лабораторной работы (ЛР №4).

1. Визуализировать значения важности признаков как вклад в показатель прироста информации, используемый при расщеплении, для моделей, использующих деревья для сбалансированных и первоначальной выборок.

2. Визуализировать значения коэффициентов логистической регрессии.

3. Получить значения Шепли и произвести их визуализацию с помощью средств библиотеки shap.

4. Отобразить важность признаков с помощью средств библиотеки dalex.

5. Выберите наиболее важные признаки, влияющие на целевую переменную, сделайте выводы по результатам исследования важности и отразите проделанную работу в отчёте.

7. Кластеризация данных и оценка её качества

Напишите программу на Python, которая выполняет анализ алгоритмов кластеризации. В качестве набора данных выберите один из тех, которые вы рассмотрели на предыдущих лабораторных работах. Для кластеризации выделите два произвольных численных признака. При необходимости приведите признаки к стандартному масштабу. Используйте следующие алгоритмы кластеризации: K-means, агломеративная

кластеризация и DBScan. Проведите эксперимент по выявлению оптимального количества кластеров, для каждого результата выведите метрику качества (например, коэффициент силуэта или др.), покажите на графике кластеры и центроиды для каждого алгоритма и их гиперпараметров. Выберите лучшее сочетание алгоритма и гиперпараметров, сделайте выводы и отразите проделанную работу в отчёте.

8. Обработка естественного языка (классификация текстов)

Собрать статьи из электронной газеты (по Вашему выбору) соответствующие пяти тематикам (например, путешествия, еда, автомобили, здоровье, культура и т.д.). Написать программу на PythonSpark, которая загружает набор данных по вашему варианту и производит построение классификаторов статей отзывов на основе четырех алгоритмов классификации с использованием трех любых алгоритмов векторизации (Bag-of-Words, TF-IDF, …) текстов.

Напишите короткое заключение о наиболее интересных результатах и ошибках, возникших при выполнении работы.

Темы индивидуальных проектов:

Для укрепления изученного материала предусмотрено выполнение индивидуального проекта в рамках часов самостоятельной работы. Проект может быть выполнен как индивидуально, так и в мини-группе (2-3 чел.), при условии, что объем работы также будет увеличен. В конце семестра по каждому проекту представляется мини-презентация о результатах работы.

Тематика индивидуального проекта связана с темой ВКР магистранта. Цель работы – использование методов глубинного обучения в своей научной работе.

Темы опросов на занятиях:

Связаны с материалом предыдущих лекций, а также личным опытом студентов. Студенты могут предлагать варианты решений поставленной преподавателем задачи, а также инструменты решения.

Примеры вопросов:

1.Какая нейросетевая модель из перечисленных в лучшей степени подходит для прогнозирования временных последовательностей?

a) Single-Layer Perceptron	b) CNN
c) LSTM	d) Multi-layer Perceptron

2. Как называется несколько примеров из обучающей выборки, использующихся для одномоментного расчета градиента и весов сети?

3. Почему модели на сверточных нейронных сетях показывают наилучшие показатели по классификации объектов на изображениях по сравнению с другими моделями?

a)Они в высокой степени оптимизированы для обработки векторов с числовыми, а не категориальными признаками	b)Они обладают широким набором инструментов преобразования признакового пространства, которые может варьировать разработчик в модели

c)Они учитывают корреляцию смежных компонент вектора	d)Они используют существенно б*о*льшее число настраиваемых параметров, по сравнению с другими моделями

3.2. Типовые задания для проведения промежуточной аттестации по дисциплине «Нейронные сети»

Зачет выставляется на основе представления и защиты индивидуального проекта.

Студент выполняет презентацию, а также демонстрирует программный код. Вопросы по результатам могут задавать все студенты группы, не только преподаватель.

Таблица критериев выставления зачета

Оценка	Критерии
Зачтено	Студент активно работал в течение семестра, выступил с презентацией индивидуального проекта, посещал лекционные занятия, лабораторные работы сданы в срок.
Не зачтено	Студент не работал во время семестра, не выступал с презентацией индивидуального проекта, не посещал лекционные занятия, лабораторные работы не сданы или сданы на менее чем 10 баллов.

4. Методические материалы, определяющие процедуры оценивания образовательных результатов обучения

№ п/п	Авторы / составители	Заглавие	Издательство	Год издания
1.	Джоэл Грас	Data Science: Наука о данных с нуля. 2-е издание. ISBN 978-5-9775-6731-2	СПб: БХВ-Петербург	2021
2.	Себастьян Рашка, Вахид Мирджалили	Python и машинное обучение. ISBN 978-5-907203-57-0	М.: Диалектика	2020
3.	Ameet V. Joshi	Machine Learning and Artificial Intelligence. ISBN 978-3-030-26621-9	Springer Nature Switzerland AG	2020
4.	Denis Rothman	Artificial Intelligence by Example. Second Edition. ISBN 978-1-83921-153-9	Packt Publishing	2020
6	Stuart Russel, Peter Norvig	Artificial Intelligence. A Modern Approach. 4th Edition. ISBN: 978-0-13-461099-3	Hoboken: Pearson	2021
7	Эндрю Гласснер	Глубокое обучение без	М.: ДМК Пресс	2020

		математики. Том 1. Основы. ISBN 978-5-97060-701-5		
8	Эндрю Гласснер	Глубокое обучение без математики. Том 2. Практика ISBN 978-5-97060-767-1	М.: ДМК Пресс	2020
9	Ян Гудфеллоу, Иошуа Бенджио, Аарон Курвилль	Глубокое обучение. Второе цветное издание, исправленное. ISBN 978-5-97060-618-6	М.: ДМК Пресс	2018
10	Roman Shirkin	Artificial Intelligence. The Complete Beginners' Guide to Artificial Intelligence. ISBN: 9798609154415	Amazon KDP Printing and Publishing	2020
11	Франсуа Шолле	Глубокое обучение на Python. ISBN 978-5-4461-0770-4	СПб: Питер	2018

Рейтинговая система для оценки текущей успеваемости обучающихся

Таблица – Балльные оценки для элементов контроля

Элементы учебной деятельности	Максимальный балл с начала семестра	Оцениваемая компетенция
Подготовка к лабораторным занятиям и защита отчета по лабораторной работе	15*4=60	ИОПК-2.1. ИОПК-2.2.
Защита индивидуальных проектов	40	ИОПК-2.3.
Зачет		

Пересчет баллов в оценки промежуточной успеваемости

Баллы на дату контрольной точки	Оценка
≥ 90% от максимальной суммы баллов	5 (зачтено)
От 70% до 89% от максимальной суммы баллов	4 (зачтено)
От 60% до 69% от максимальной суммы баллов	3 (зачтено)
< 60% от максимальной суммы баллов	2 (незачтено)