Лабораторная работа №2

«Обнаружение аномалий»

Цель работы:

Получить практические навыки создания, обучения и применения искусственных нейронных сетей типа автокодировщик. Исследовать влияние архитектуры автокодировщика и количества эпох обучения на области в пространстве признаков, распознаваемые автокодировщиком после обучения. Научиться оценивать качество обучения автокодировщика на основе ошибки реконструкции и новых метрик EDCA. Научиться решать актуальную задачу обнаружения аномалий в данных с помощью автокодировщика как одноклассового классификатора.

Подготовка к работе:

1) Определить свой набор данных по таблице исходя из номера бригады k:

$$N = k \bmod 3$$

Вариант (N)	0	1	2
Название набора данных (name)	Cardio	Letter	WBC

2) Подготовить программную среду Google Colaboratory для выполнения лабораторной работы:

- Разрешить доступ к своим файлам на Google Диск (Содержание > Файлы > Подключить Google Диск);
- Создать на Google Диске папку для данной лабораторной работы с именем *is_lab2*;
- Скачать библиотеку *lab02_lib.py*, необходимую для выполнения лабораторной работы, в папку *is_lab2*;
- Скачать в *is_lab2* файлы обучающей и тестовой выборки: *name_train.txt*, *name_test.txt*, где *name* – название набора данных;
- Создать в *is_lab2* папку для результатов с именем *out*.

Задание 1:

1) В среде Google Colab создать новый блокнот (notebook). Импортировать необходимые для работы библиотеки и модули.
2) Сгенерировать индивидуальный набор двумерных данных в пространстве признаков с координатами центра (k, k), где k – номер бригады. Вывести полученные данные на рисунок и в консоль.
3) Создать и обучить автокодировщик AE1 простой архитектуры, выбрав небольшое количество эпох обучения. Зафиксировать в таблице вида табл.1 количество скрытых слоёв и нейронов в них.
4) Зафиксировать ошибку MSE, на которой обучение завершилось. Построить график ошибки реконструкции обучающей выборки. Зафиксировать порог ошибки реконструкции – порог обнаружения аномалий.
5) Создать и обучить второй автокодировщик AE2 с усложненной архитектурой, задав большее количество эпох обучения.
6) Зафиксировать ошибку MSE, на которой обучение завершилось. Построить график ошибки реконструкции обучающей выборки. Зафиксировать второй порог ошибки реконструкции – порог обнаружения аномалий.
7) Рассчитать характеристики качества обучения EDCA для AE1 и AE2. Визуализировать и сравнить области пространства признаков, распознаваемые автокодировщиками AE1 и AE2. Сделать вывод о пригодности AE1 и AE2 для качественного обнаружения аномалий.
8) Если автокодировщик AE2 недостаточно точно аппроксимирует область обучающих данных, то подобрать подходящие параметры автокодировщика и повторить шаги (6) – (8).
9) Изучить сохраненный набор данных и пространство признаков. Создать тестовую выборку, состоящую, как минимум, из 4ёх элементов, не входящих в обучающую выборку. Элементы должны быть такими, чтобы AE1 распознавал их как норму, а AE2 детектировал как аномалии.
10) Применить обученные автокодировщики AE1 и AE2 к тестовым данным и вывести значения ошибки реконструкции для каждого элемента тестовой выборки относительно порога на график и в консоль.
11) Визуализировать элементы обучающей и тестовой выборки в областях пространства признаков, распознаваемых автокодировщиками AE1 и AE2.

12) Результаты исследования занести в таблицу:

Табл. 1 Результаты задания №1

	Количество скрытых слоев	Количество нейронов в скрытых слоях	Количество эпох обучения	Ошибка MSE_stop	Порог ошибки реконструкции	Значение показателя Excess	Значение показателя Approx	Количество обнаруженных аномалий
AE1								
AE2								

13) Сделать выводы о требованиях к:
- данным для обучения,
- архитектуре автокодировщика,
- количеству эпох обучения,
- ошибке MSE_stop, приемлемой для останова обучения,
- ошибке реконструкции обучающей выборки (порогу обнаружения аномалий),
- характеристикам качества обучения EDCA одноклассового классификатора,

для качественного обнаружения аномалий в данных.

Задание 2:

1) Изучить описание своего набора реальных данных, что он из себя представляет;
2) Загрузить многомерную обучающую выборку реальных данных *name_train.txt*.
3) Вывести полученные данные и их размерность в консоли.
4) Создать и обучить автокодировщик с подходящей для данных архитектурой. Выбрать необходимое количество эпох обучения.
5) Зафиксировать ошибку MSE, на которой обучение завершилось. Построить график ошибки реконструкции обучающей выборки. Зафиксировать порог ошибки реконструкции – порог обнаружения аномалий.
6) Сделать вывод о пригодности обученного автокодировщика для качественного обнаружения аномалий. Если порог ошибки реконструкции слишком велик, то подобрать подходящие параметры автокодировщика и повторить шаги (4) – (6).
7) Изучить и загрузить тестовую выборку *name_test.txt*.
8) Подать тестовую выборку на вход обученного автокодировщика для обнаружения аномалий. Вывести график ошибки реконструкции элементов тестовой выборки относительно порога.

9) Если результаты обнаружения аномалий не удовлетворительные (обнаружено менее 70% аномалий), то подобрать подходящие параметры автокодировщика и повторить шаги (4) – (9).
10) Параметры наилучшего автокодировщика и результаты обнаружения аномалий занести в таблицу:

Табл. 2 Результаты задания №2

Dataset name	Количество скрытых слоев	Количество нейронов в скрытых слоях	Количество эпох обучения	Ошибка MSE_stop	Порог ошибки реконструкции	% обнаруженных аномалий

11) Сделать выводы о требованиях к:
    - данным для обучения,
    - архитектуре автокодировщика,
    - количеству эпох обучения,
    - ошибке MSE_stop, приемлемой для останова обучения,
    - ошибке реконструкции обучающей выборки (порогу обнаружения аномалий),

    для качественного обнаружения аномалий в случае, когда размерность пространства признаков высока.

Контрольные вопросы

1) Что такое автокодировщик и какова особенность его архитектуры?
2) Для каких задач используется автокодировщик?
3) Что такое ошибка обучения MSE?
4) Что такое ошибка реконструкции и порог ошибки реконструкции?
5) Какой вывод об области, распознаваемой автокодировщиком, можно сделать на основе значений характеристик качества обучения EDCA (например, по Excess и Approx)?
6) Для каких прикладных областей задача обнаружения аномалий наиболее актуальна?
7) Как влияет архитектура автокодировщика на качество обнаружений аномалий?
8) Как влияет количество эпох обучения автокодировщика на качество обнаружений аномалий?
9) Как влияет порог ошибки реконструкции на качество обнаружений аномалий?