Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ ИУ «Информатика и системы управления»

КАФЕДРА ИУ-1 «Системы автоматического управления»

ОТЧЕТ

по лабораторной работе №2

«Кластерный анализ»

по дисциплине

«Основы теории искусственного интеллекта»

Выполнил: Шевченко А.Д.

Группа: ИУ1-52Б

Проверил: Мусахан А.Д.

Работа выполнена: 28.11.2023

Отчет сдан: 04.12.2023

Оценка:

Москва 2023

Содержание

Цель работы:

Освоить применение алгоритма k-средних для кластеризации объектов и оценить оптимальное число кластеров с использованием различных индексов.

Решаемые задачи:

Вариант № 24, соотвествует номеру в группе по списку.

Освоение алгоритма k-средних для кластеризации объектов и знакомство с методами оценки качества.

Для выполнения лабораторной работы необходимо:

1. подготовить данные;
2. запустить алгоритм с различными значениями числа кластеров и проанализировать результаты, применив индекс Дэвиса-Болдуина для оценки оптимального числа кластеров и для лучшего представления;
3. сделать визуализацию в виде нескольких графиков.

Теоретическая часть

1. Кластерный анализ

Кластерный анализ многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы. Задача кластеризации относится к статистической обработке, а также к широкому классу задач обучения без учителя.

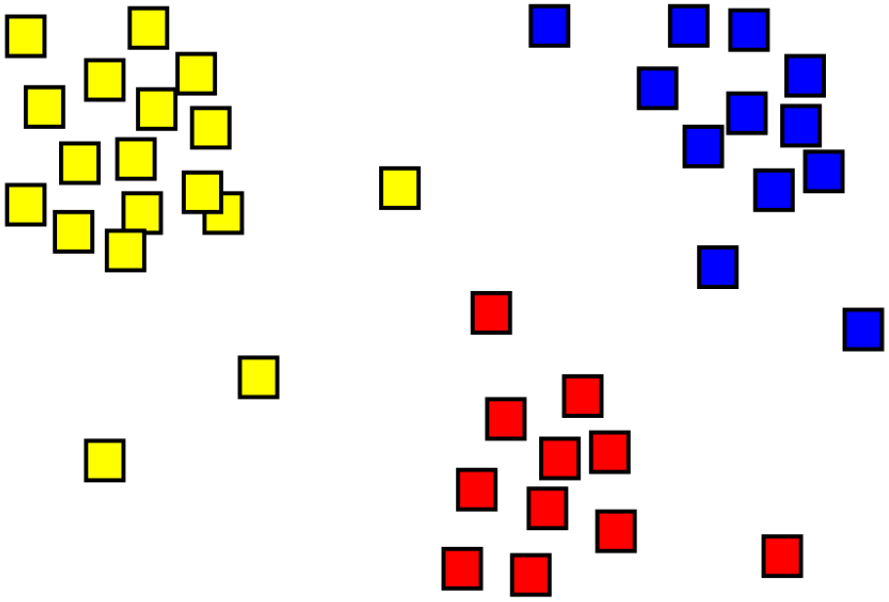

Рис. 1 Результат кластерного анализа обозначен раскрашиванием точек в соответствии с принадлежностью к одному из трёх кластеров.

Более простым языком, кластерный анализ позволяет группировать объекты на основе их сходства. В процессе кластеризации объекты, имеющие схожие характеристики, объединяются в группы, называемые кластерами.

2. Методы кластеризации

Общепринятой классификации методов кластеризации не существует, но можно выделить ряд групп подходов (некоторые методы можно отнести сразу к нескольким группам и потому предлагается рассматривать данную типизацию как некоторое приближение к реальной классификации методов кластеризации):

1. *Вероятностный подход*.

Предполагается, что каждый рассматриваемый объект относится к одному из k классов. Некоторые авторы (например, А. И. Орлов) считают, что данная группа вовсе не относится к кластеризации и противопоставляют её под названием «дискриминация», то есть выбор отнесения объектов к одной из известных групп (обучающих выборок).

- K-средних
- K-медиан
- EM-алгоритм
- Алгоритмы семейства FOREL
- Дискриминантный анализ

2. *Подходы на основе систем искусственного интеллекта*

Весьма условная группа, так как методов очень много и методически они весьма различны.

- Метод нечеткой кластеризации C-средних
- Нейронная сеть Кохонена
- Генетический алгоритм

3. *Логический подход.*

Построение дендрограммы осуществляется с помощью дерева решений.

4. *Теоретико-графовый подход.*

- Графовые алгоритмы кластеризации

5. *Иерархический подход.*

Предполагается наличие вложенных групп (кластеров различного порядка). Алгоритмы в свою очередь подразделяются на агломеративные (объединительные) и дивизивные (разделяющие). По количеству признаков иногда выделяют монотетические и политетические методы классификации.

- Иерархическая дивизивная кластеризация или таксономия. Задачи кластеризации рассматриваются в количественной таксономии.

6. *Другие методы.*

Не вошедшие в предыдущие группы.

- Статистические алгоритмы кластеризации
- Ансамбль кластеризаторов
- Алгоритмы семейства KRAB
- Алгоритм, основанный на методе просеивания
- DBSCAN и др.

3. Метод k-средних

Метод *k*-средних (англ. *k-means*) — наиболее популярный метод кластеризации. Был изобретён в 1950-х годах математиком Гуго Штейнгаузом и почти одновременно Стюартом Ллойдом. Особую популярность приобрёл после работы Маккуина.

Действие алгоритма таково, что он стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров.

$$V = \sum_{i=1}^{k} \sum_{x \in S_i} (x - \mu_i)^2$$

где k — число кластеров, S_i — полученные кластеры, $i = 1, 2, \ldots, k$, а μ_i — центры масс всех векторов x из кластера S_i.

Алгоритм представляет собой версию EM-алгоритма, применяемого также для разделения смеси гауссиан. Он разбивает множество элементов векторного пространства на заранее известное число кластеров k.

Основная идея заключается в том, что на каждой итерации перевычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике.

Алгоритм завершается, когда на какой-то итерации не происходит изменения внутрикластерного расстояния. Это происходит за конечное число итераций, так как количество возможных разбиений конечного множества конечно, а на каждом шаге суммарное квадратичное отклонение V уменьшается, поэтому зацикливание невозможно.

Демонстрация алгоритма

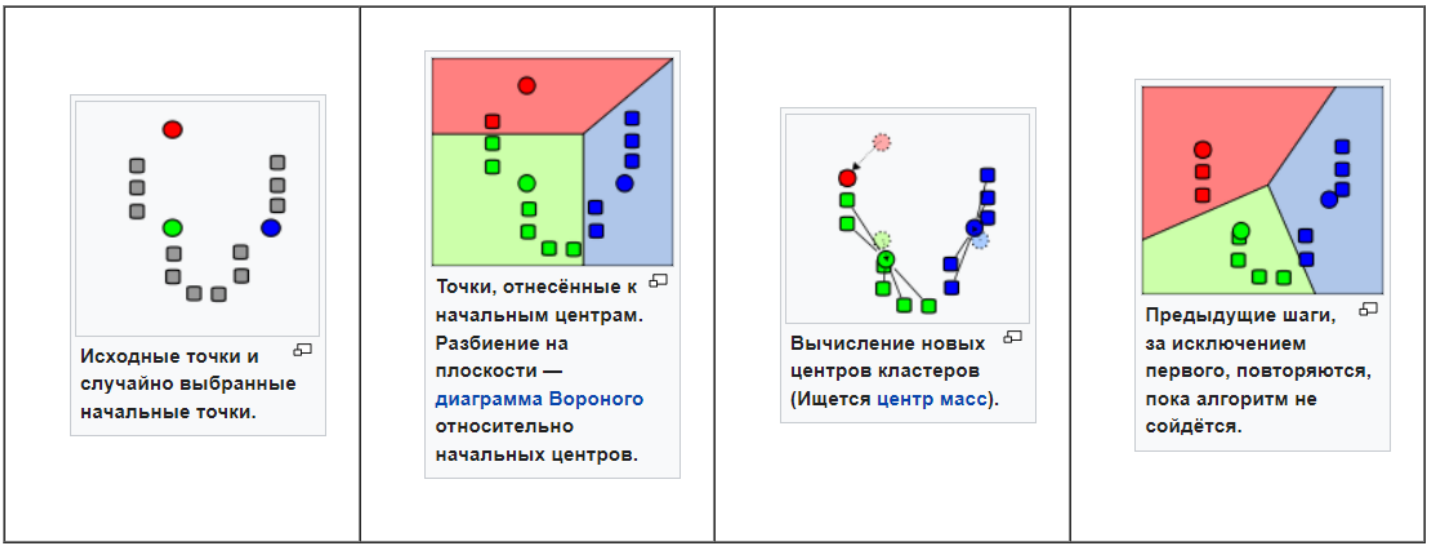

Рис. 2 Демонстрация работы алгоритма k средних

4. Идекс Девида-Болдуина (DBI)

Это показатель для оценки алгоритмов кластеризации. Это схема внутренней оценки, в которой проверка того, насколько хорошо была выполнена кластеризация, производится с использованием величин и характеристик, присущих набору данных. Недостатком этого метода является то, что хорошее значение, сообщаемое этим методом, не означает наилучшего поиска информации.

Учитывая n размерных точек, пусть C_i — скопление точек данных. Пусть X_j — n-мерный вектор признаков, присвоенный кластеру C_i.

$$S_i = \left(\frac{1}{T_i} \sum_{j=1}^{T_i} \|X_i - A_i\|_p^q \right)^{1/q}$$

Здесь — центроид C_i, а T_i — размер кластера i. S_i — q-й корень из q-го момента точек в кластере i относительно среднего значения.

Если q = 1 когда S_i — среднее расстояние между векторами признаков в кластере i и центром тяжести кластера. Обычно значение p равно 2, что делает расстояние функцией евклидова расстояния. Чем ближе данная оценка к 1, тем лучше произошло разделение объектов на кластеры.

5. Модификация k-средних++

Модификация внесена в шаг 1 инициализации центроидов.

Улучшенная инициализация центроидов:

Вместо случайного выбора одной точки в качестве первого центроида, k-средних++ использует более разумный метод. Первый центроид выбирается случайным образом из набора данных, затем для каждой следующей точки выбирается с вероятностью пропорциональной квадрату расстояния от этой точки до ближайшего уже выбранного центроида. Это уменьшает вероятность выбора точек, находящихся далеко от уже выбранных центроидов.

Преимущества модификации:

Улучшенная сходимость - благодаря более разумной инициализации, алгоритм k-средних++ часто достигает сходимости быстрее по сравнению с классическим методом. Более стабильные результаты - избегание случайных выборов в начале может сделать алгоритм менее чувствительным к исходной инициализации и повысить стабильность результатов кластеризации.

Модификация k-средних++ часто применяется в практике кластерного анализа для улучшения производительности и результатов алгоритма k-средних.

4. Описание выполнения работы

a. Импортируются необходимые библиотеки:
   - numpy
   - matplotlib
   - sklern

b. Подготавливаются данные в виде координат в трёхмерном пространстве.

c. Выводим график начальных точек, что требуется в задании.

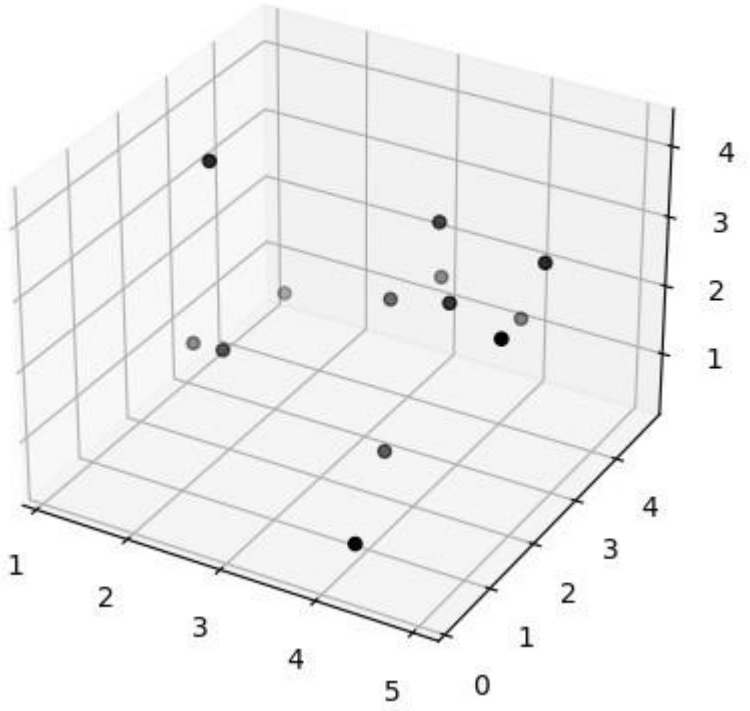

Рис. 3 График отображения начальных точек в 3D пространстве

d. Запускается сам алгоритм k-means — для каждого значения k алгоритм подгоняется к данным, и полученные метки и центроиды сохраняются. Оценка “DBI” или “индекс Дэвида-Болдуина” рассчитывается для каждой кластеризации и сохраняется.

e. Отображаем кластеры, для каждого значения k создается точечная диаграмма, на которой точки данных показаны цветом в соответствии с их назначением в кластере.

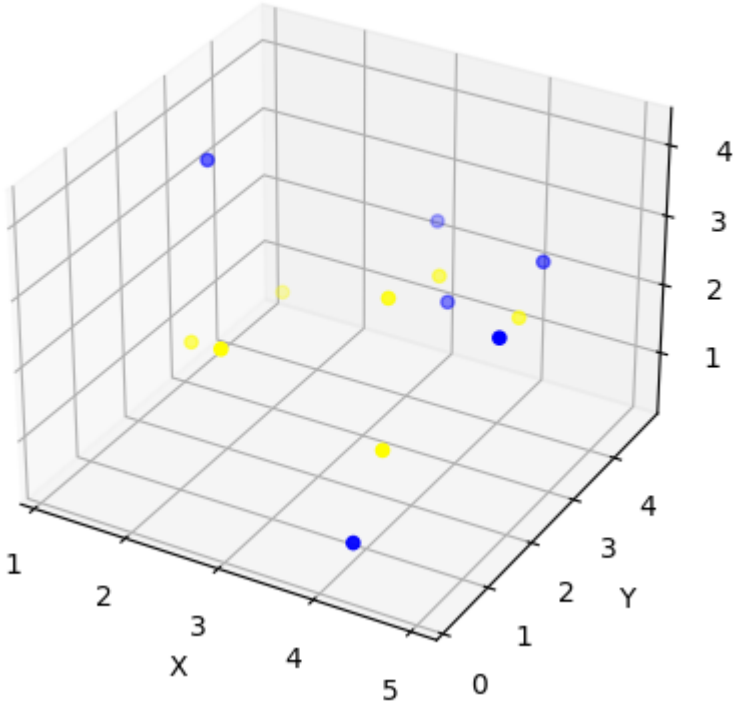

Рис. 4 График распределения по кластерам для k=2

DBI = 1.2299418658346146

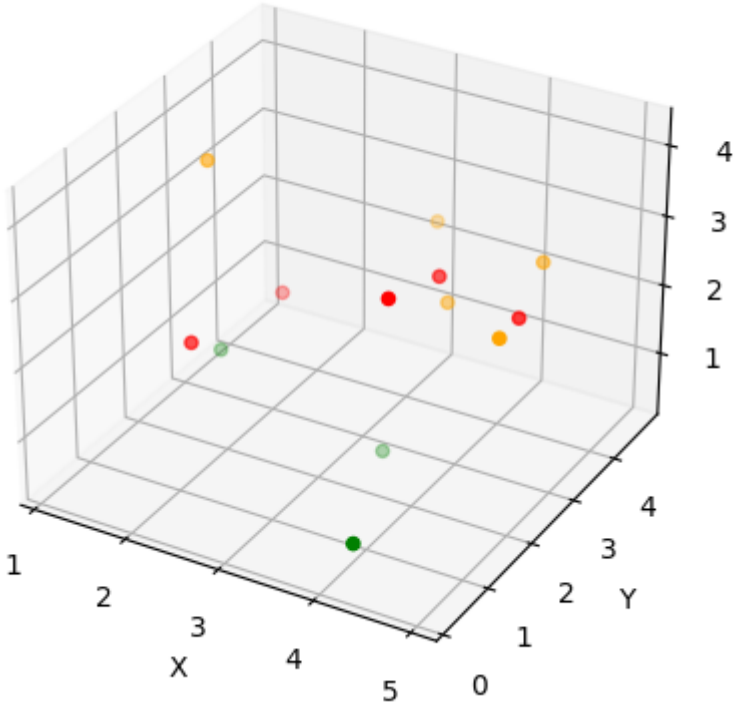

Рис. 5 График распределения по кластерам для k=3

DBI = 1.0353276154373376

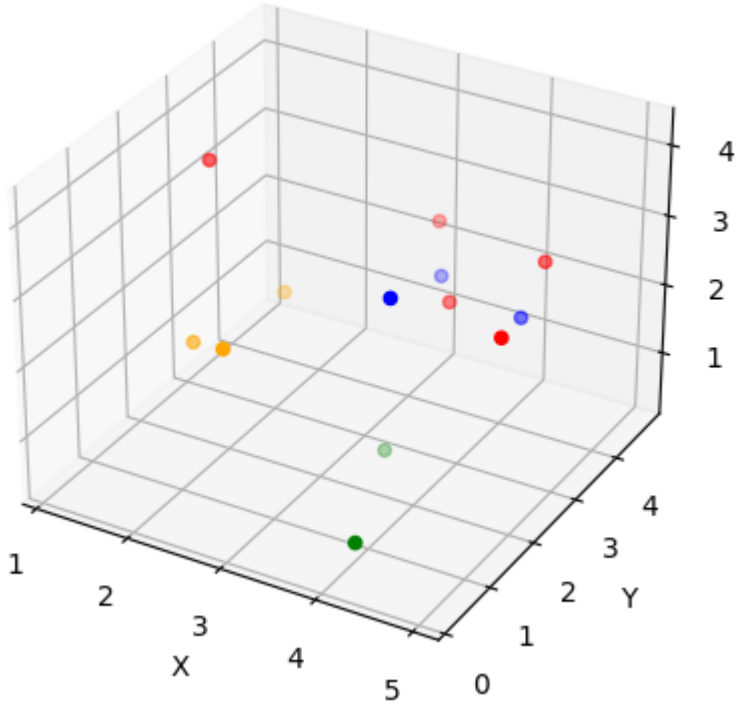

Рис. 6 График распределения по кластерам для k=4

DBI = 0.8184046977229166

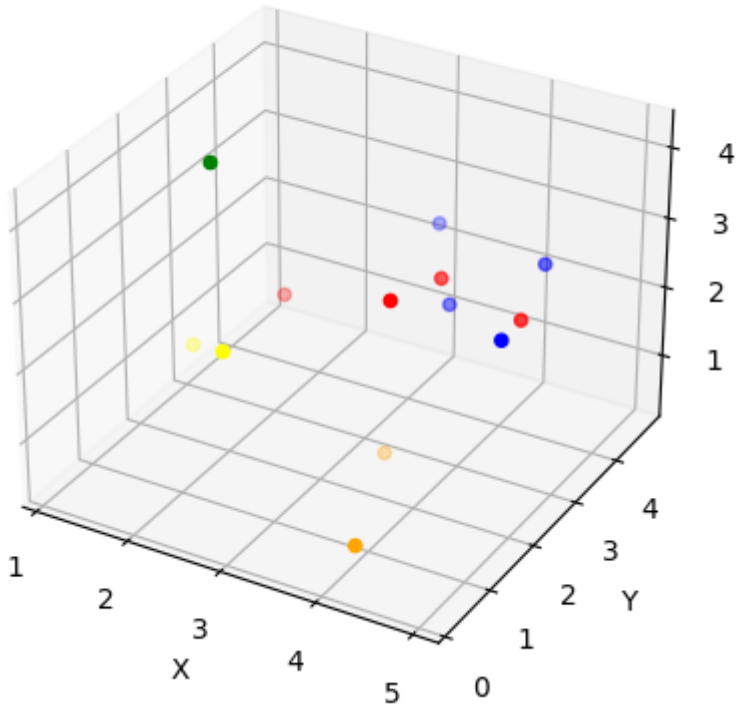

Рис. 7 График распределения по кластерам для k=5

DBI = 0.6420823435785749

f. Создаётся линейный график, показывающий оценку “DBI” для каждого значения количества кластеров.

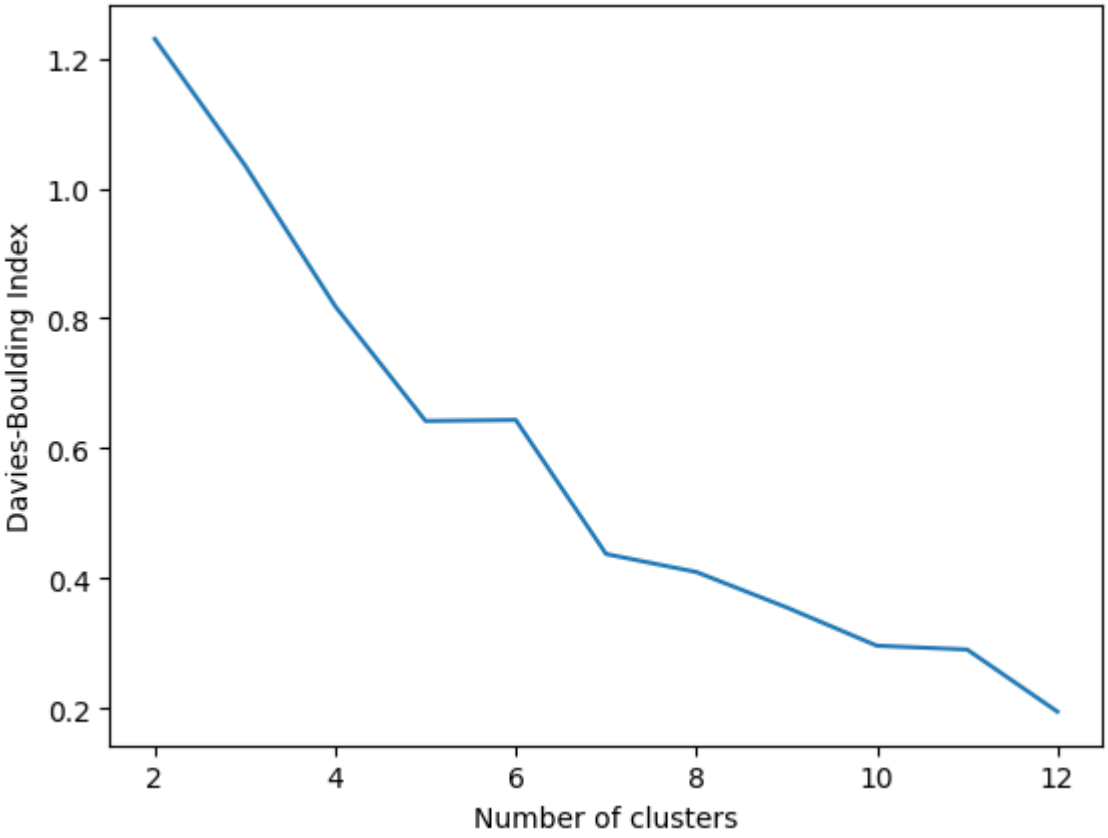

Рис. 8 График изменения индекса Дэвида-Болдана в зависимости от количества кластеров

g. По графику можно оценить оптимального количества кластеров.

Заметим, что самый лучший показатель индекса Девида-Болдана для исходных данных достигается при количестве кластеров равном 3. При этом значении клестеров DBIndex = 1.0353276154373376, что имеет наибольшее приближение к оптималдьному значению разбиения = 1.

Вывод:

В ходе данной лабораторной работы были изучены темы кластеризации, и такой вероятсностный метод кластеризации, как k-means, а так же изучен и использован такой метод оценки разбиения объектов на кластеры, как индекс Девида-Бэлдона.

Список источников:

https://en.wikipedia.org/wiki/Davies–Bouldin_index

https://ru.wikipedia.org/wiki/Кластерный_анализ

k-means clustering - Wikipedia

Приложение (код):

https://github.com/BMSTU-Automatic-Control-Systems-IU1-1/semesters/tree/semester-5/theory-of-artificial-intelligence