Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования

ИРКУТСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

|  |
| --- |
| Институт информационных технологий и анализа данных |

наименование института

ОТЧЕТ ПО ЛАБОРАТОРНОЙ РАБОТЕ № 3

по дисциплине:

|  |
| --- |
| **Методы анализа данных** |
| **«Кластерный и дискриминантный анализ»** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Выполнил | АСУб-20-2 |  |  |  | Арбакова А.В. |
|  | шифр группы |  | подпись |  | Фамилия И.О. |
| Проверил |  |  |  |  | Бучнев О.С. |
|  | должность |  | подпись |  | Фамилия И.О. |

Иркутск 2022 г

**Цель лабораторной работы:**

1. Выбрать среду программирования для языка Python.
2. Получить набор данных из 100 наблюдений с параметрами, согласно номеру варианта. С использованием Python провести кластерный и дискриминантный анализ данных. При этом необходимо:

* получить значения основных показателей описательной статистики и построить диаграммы рассеяния для всех признаков;
* выполнить стандартизацию данных;
* построить и вывести на экран дендрограмму;
* провести кластеризацию данных методом k-средних, вывести на экран результаты кластеризации; показать на этих же диаграммах средние значения для каждого кластера;
* построить диаграмму средних значений для каждого кластера и каждого измерения;
* создать инструмент для графического отображения результатов дискриминантного анализа – решающих поверхностей и данных;
* провести линейный дискриминантный анализ данных, для каждой пары признаков: вывести на экран решающие поверхности и данные, оценить точность;
* провести квадратичный дискриминантный анализ данных, для каждой пары признаков: вывести на экран решающие поверхности и данные, оценить точность; сравнить с точностью линейного дискриминантного анализа.

1. Выполнить анализ полученных на каждом этапе результатов и оформить отчет по лабораторной работе.

**Язык программирования:** Python

**Номер варианта:** 3

|  |  |  |  |
| --- | --- | --- | --- |
| № Варианта | Количество кластеров | Количество признаков | Дисперсия значений |
| 3 | 3 | 3 | 2 |

# **Ход выполнения работы**

Кластеризация – разбиение совокупности объектов на однородные группы (кластеры).

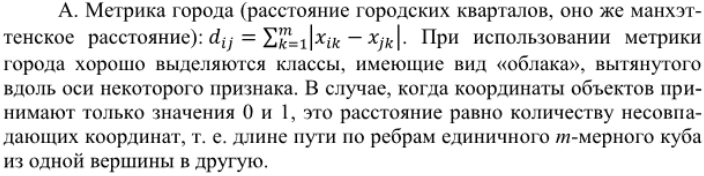
Кластерный анализ – группа методов, используемых для классификации объектов или событий в относительно однородные группы, которые называют кластерами. Кластерный анализ предназначен для разбиения исходных данных на поддающиеся интерпретации группы, таким образом, чтобы элементы, входящие в одну группу были максимально «схожи», а элементы из разных групп были максимально «отличными» друг от друга.

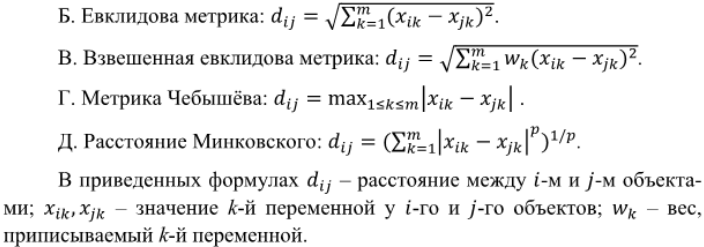
Задача кластерного анализа – выделение «сгущений» точек, разбиение совокупности на однородные подмножества объектов (сегментация).

Методы кластерного анализа позволяют решать следующие задачи:

* проведение классификации объектов с учетом признаков, отражающих сущность, природу объектов. Решение такой задачи, как правило, приводит к углублению знаний о совокупности классифицируемых объектов;
* проверка выдвигаемых предположений о наличии некоторой структуры в изучаемой совокупности объектов, т. е. поиск существующей структуры;
* построение новых классификаций для слабоизученных явлений, когда необходимо установить наличие связей внутри совокупности и попытаться привнести в нее структуру.

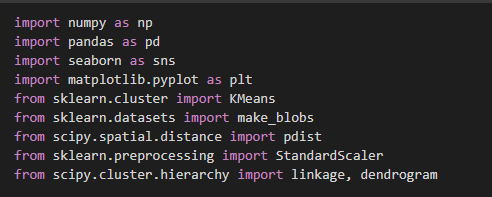
Метрики расстояния. Для каждого типа данных существует несколько способов измерения расстояния или определения меры сходства объектов. Наиболее используемыми для интервальных данных являются:





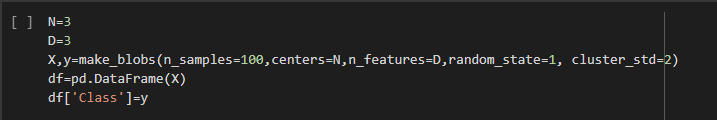
Метод кластеризации – это способ вычисления расстояний между кластерами. Существуют следующие основные методы кластеризации: межгрупповая связь, внутригрупповая связь, ближайший сосед, самый дальний сосед, центроидная кластеризация, медианная кластеризация, метод Варда.

Подключим необходимые библиотеки и модули:



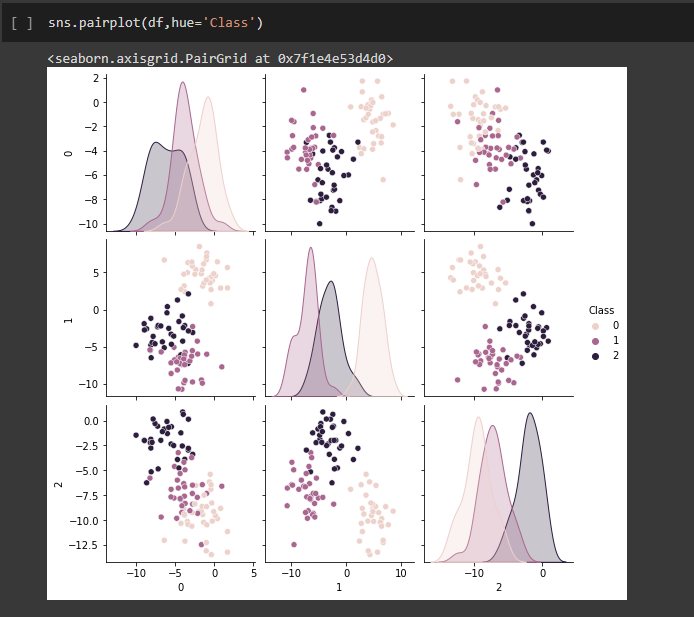
Изображение 1 – Подключаемые библиотеки и модули

Получим исходные данные. Для примера выберем число кластеров N = 3, число признаков D = 3:



Изображение 2 – Исходные данные (число кластеров, число признаков)

Построим диаграммы рассеяния:



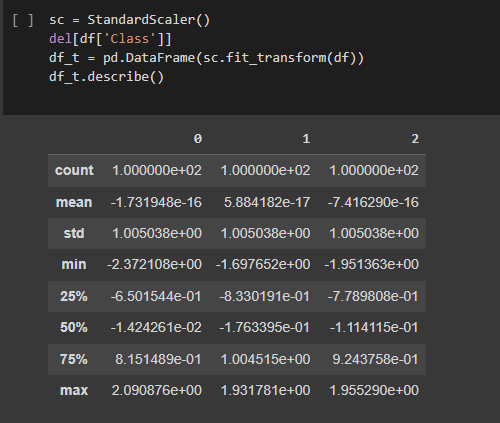
Изображение 3 – Диаграмма рессеяния

Иерархический – при маленьком количестве наблюдений.

Анализ к средних – при большом количестве наблюдений.

Стандартизация данных – это рабочий процесс обработки данных, который преобразует структуру различных наборов данных в один общий формат данных. Позволяет сравнивать значения, измеренные в разных шкалах.

Выполним стандартизацию данных и выведем результат:

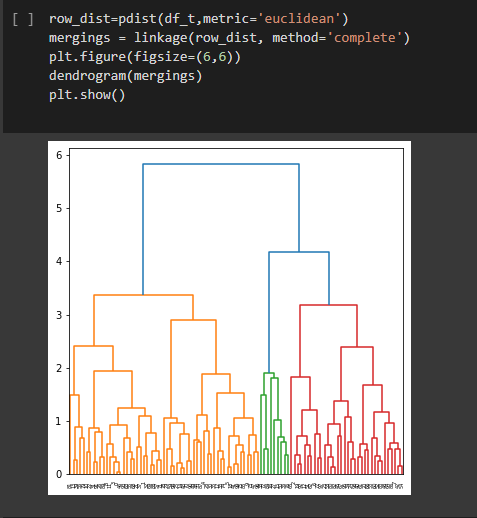


Изображение 4 – Стандартизация данных

Иерархический кластерный анализ.

Эта процедура предназначена для выявления относительно однородных групп наблюдений по заданным характеристикам при помощи алгоритма, который вначале рассматривает каждое наблюдение как отдельный кластер, а затем последовательно объединяет кластеры, пока не останется только один.

Выполним иерархическую кластеризацию:



Изображение 5 – Иерархическая кластеризация

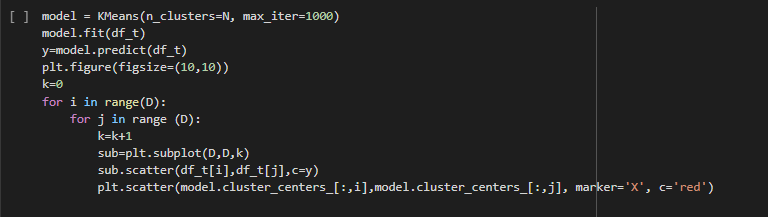
Алгоритм k-means.

K-means – это алгоритм обучения без учителя. Смысл алгоритма заключается в наблюдении за набором немаркированных данных для автоматического обнаружения скрытой структуры, а также для обнаружения закономерности в немаркированных данных.

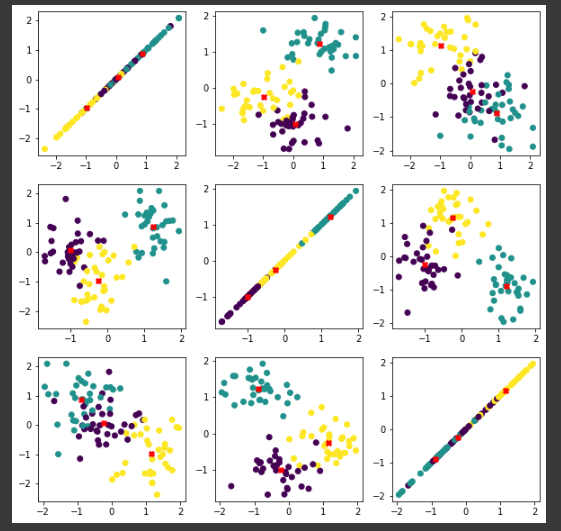
Задача минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров.

Сначала определяется центр кластера, а затем группируют все объекты в пределах заданного от центра порогового значения.

Выполним кластеризацию методом k-средних, выведем диаграммы рассеяния, нанесем центры кластеров:



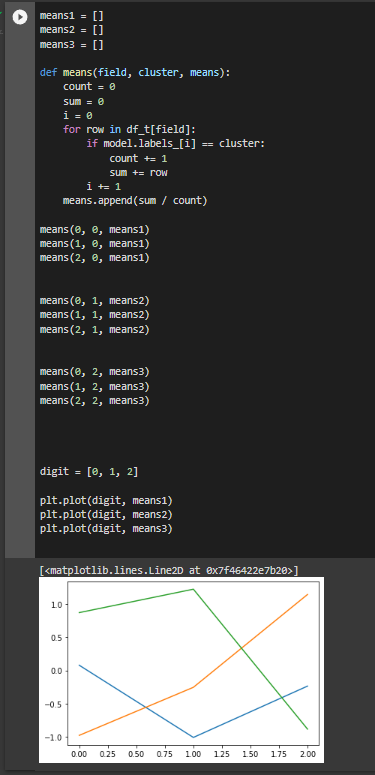
Изображение 6 – Кластеризация методом k-средних



Изображение 6 – Кластеризация методом k-средних

График средних. Основным способом визуализации результата работы алгоритма, когда число измерений т больше двух, является график средних, где линии показывают масштабированные средние по каждому из т измерений для каждого кластера. По горизонтали отложены участвующие в классификации переменные, а по вертикали – средние значения переменных для каждого кластера.

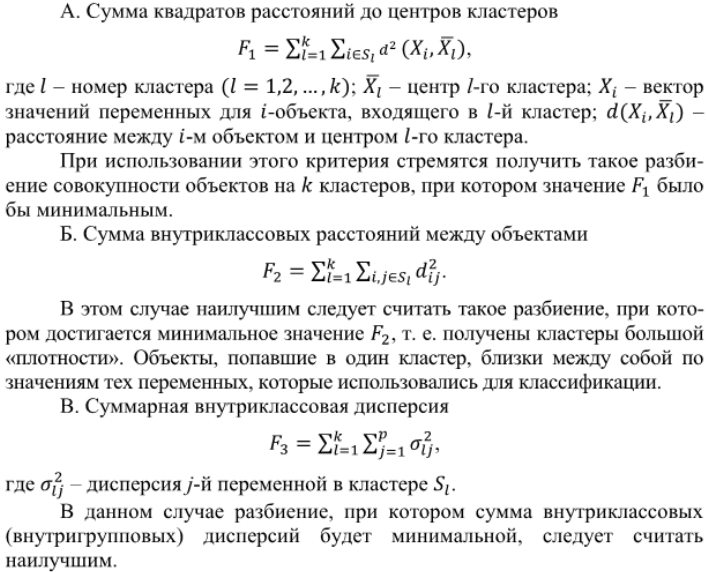
Выведем графики средних для каждого измерения каждого кластера:



Изображение 7 – График средних

Таким образом, после завершения процедуры классификации необходимо оценить полученные результаты. Для этой цели используется мера качества, которую называют критерием качества. Наилучшим по выбранному критерию считают такое разбиение, при котором критерий качества принимает экстремальное (минимальное или максимальное) значение.

Наиболее часто используются следующие критерии:



Дискриминантный анализ – общий термин, относящийся к нескольким тесно связанным статистическим процедурам. Условно эти процедуры можно разделить на методы интерпретации межгрупповых различий (дискриминации) и методы классификации наблюдений по группам.

Задачи дискриминантного анализа:

1. Задачи первого типа – задачи дискриминации.
2. Второй тип задачи относится к ситуации, когда признаки принадлежности объекта к той или иной группе потеряны, и их нужно восстановить.
3. Задачи третьего типа связаны с предсказанием будущих событий на основании имеющихся данных.

Ограничения. Зависимая переменная должна быть представлена в номинативной шкале, а независимые измерены в метрической шкале (интервалов или отношений) и иметь нормальное распределение.

Дискриминантный анализ позволяет решить две группы проблем:

1. Интерпретировать различия между классами, то есть ответить на вопросы: насколько хорошо можно отличить один класс от другого, используя данный набор переменных; какие из этих переменных наиболее существенны для различения классов.

2. Классифицировать объекты, то есть отнести каждый объект к одному из классов, исходя только из значений дискриминантных переменных.

Дискриминантная функция – выведенная посредством дискриминантного анализа линейная комбинация независимых переменных, с помощью которой можно наилучшим образом различить (дискриминировать) категории зависимой переменной.

Линейный дискриминантный анализ (LDA) – это обобщение линейного дискриминанта Фишера, метода, используемого в статистике и других областях, для нахождения линейной комбинации признаков, которая характеризует или разделяет два или более классов объектов или событий.

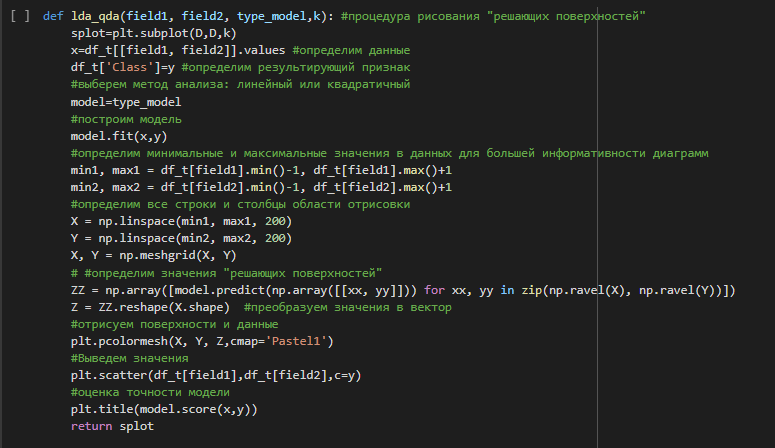
Линейный дискриминантный анализ предполагает построение линейных дискримнантных функций. Для некоторых наборов данных результаты классификации можно улучшить, если использовать нелинейные дискриминантные функции. Поэтому разработан квадратичный дискриминантный анализ (QDA), где, как следует из названия, используются дискримнантные функции второго порядка. В квадратичном дискриминантном анализе, как и в линейном, предполагается, что наблюдения в каждом классе распределены по нормальному закону.

Оценка качества модели дискриминантного анализа.

Для того чтобы выяснить, насколько хорошо построенные дискриминирующие функции могут предсказывать, к какому классу принадлежит конкретное наблюдение, анализируют таблицу сопряженности «Факт – Прогноз» и оценивают вероятность ошибочной классификации каждого класса. Если процент ошибок высок, то причиной этого является либо нарушение предположений о нормальности многомерного распределения и равенстве ковариационных матриц, либо использование плохих дискриминантных переменных.

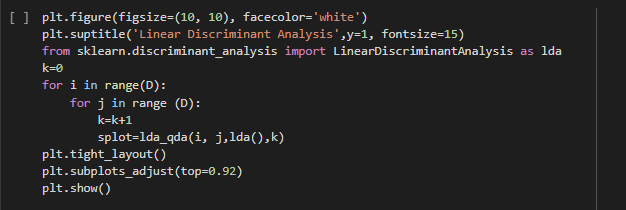
Также можно использовать следующий прием: один из элементов обучающей выборки исключается и строится классификация по оставшимся элементам и удаленный элемент используется в качестве экзаменующего. Затем удаленный возвращается в выборку, удаляется второй элемент, и процедура повторяется. Алгоритм заканчивает свою работу, когда удалению подвергнутся все элементы выборки по очереди. Ошибка дискриминации складывается из ошибок дискриминации каждого экзаменующего элемента.

Создадим функцию, реализующую инструмент для графического отображения результатов дискриминантного анализа, – решающих поверхностей и отображения данных:

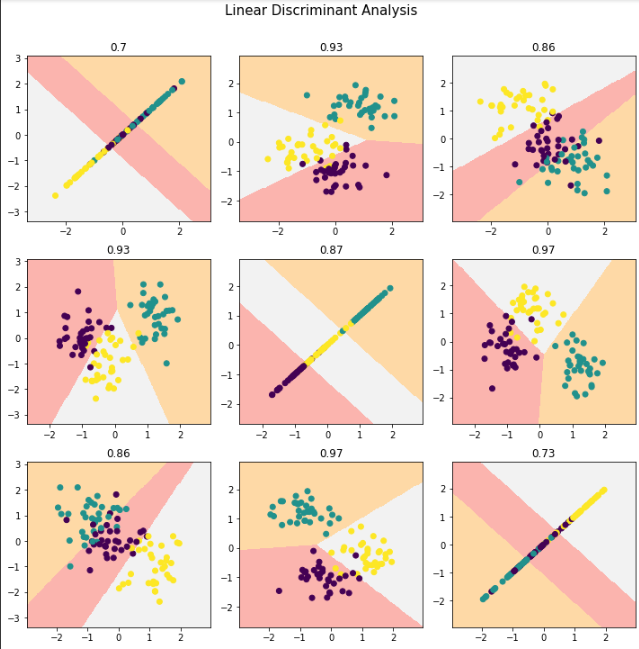


Изображение 8 – Функция, реализующая решающие поверхности и отображение данных

Проведем линейный дискриминантный анализ данных. Для каждой пары признаков выведем на экран решающие поверхности и данные, для каждой пары признаков оценим точность модели дискриминантного анализа:

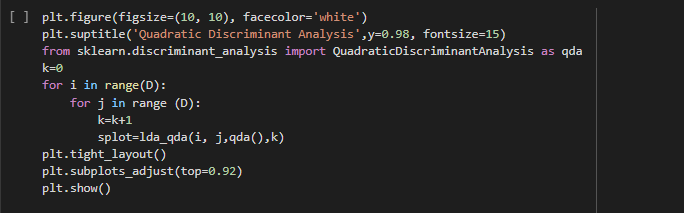


Изображение 9 – Код линейного дискриминантного анализа

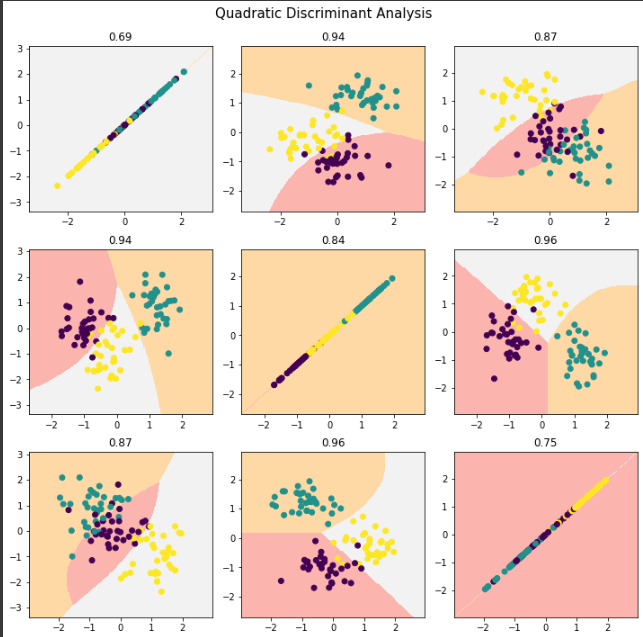


Изображение 10 – Линейный дискриминантный анализ

Выполним предыдущий пункт, используя модель квадратичного дискриминантного анализа:



Изображение 11 – Код квадратичного дискриминантного анализа



Изображение 12 – Квадратичный дискриминантный анализ

# **Заключение**

В ходе выполнения лабораторной работы по теме «Кластерный и дискриминантный анализ» была выполнена цель работы проведению кластерного и дискриминантного анализ данных.

Было изучены такие темы как: назначение кластерного анализа, метрики расстояния, иерархический кластерный анализ, алгоритм k-means, график средних, оценка качества кластеризации, назначение дискриминантного анализа, дискриминантные функции, линейный дискриминантный анализ, квадратичный дискриминантный анализ и оценка качества модели дискриминантного анализа.

Знания были закреплены во время использования среды разработки Google Colab и использования языка программирования Python, в котором было выполнено задание лабораторной работы.

# **Список литературы**

1. Бучнев О. С. Методы анализа данных: лабораторный практикум – Иркутск, 2022. – 115 с
2. <https://colab.research.google.com>
3. С.Э. Мастицкий, В.К. Шитиков Статистический анализ и визуализация данных с помощью языка R, Хайдельберг – Лондон – Тольятти, 2014 г. – 401 с