Министерство образования и науки Российской Федерации

Севастопольский государственный университет

Институт информационных технологий

Кафедра ИС

# ОТЧЁТ

по лабораторной работе №5

ЛИНЕЙНЫЙ ДИСКРИМИНАНТНЫЙ АНАЛИЗ. ПОСТРОЕНИЕ КАНОНИЧЕСКИХ И КЛАССИФИКАЦИОННЫХ ФУНКЦИЙ

Выполнил:

ст. гр. ИС/б-21-2-о

Мовенко К. М.

Проверила:

Сырых О. А.

Севастополь

2024

# ЦЕЛЬ РАБОТЫ

Закрепить теоретические знания и приобрести практические навыки в проведении дискриминантного анализа по экспериментальным данным. Исследовать возможности языка R для проведения дискриминантного анализа.

# ЗАДАНИЕ

* 1. Подготовить данные для дискриминантного анализа. Для этого разделить исходные данные на 3 кластера;
  2. Создать тренировочную выборку из исходных данных с известной группировкой;
  3. Создать выборку оставшихся данных для последующей проверки классификации;
  4. Провести дискриминантный анализ по тренировочной выборке используя функцию lda(). По полученным данным составить дискриминантную функцию;
  5. Провести классификацию оставшихся данных и построить матрицу неточностей. Сделать выводы по полученным данным;
  6. Провести шаговую процедуру выбора переменных для построения дискриминантной модели.
  7. Построить дискриминантную модель с выбранными переменными, составить уравнение дискриминантной функции;
  8. Вывести показатели оценки качества построенной модели: матрицу неточностей, ошибку распознавания, расстояние Махалонобиса;
  9. Сделать выводы по построенной модели. Сравнить полученные результаты с моделью в которой использовались все переменные;
  10. Добавить в выборку данные без классификации. Используя дискриминантный анализ, провести классификацию.

# ХОД РАБОТЫ

Был рассмотрен набор данных из предыдущих лабораторных работ. Данные были загружены в R Commander и разделены на три кластера. Номер кластера для каждой строки таблицы был сохранён в новом столбце.

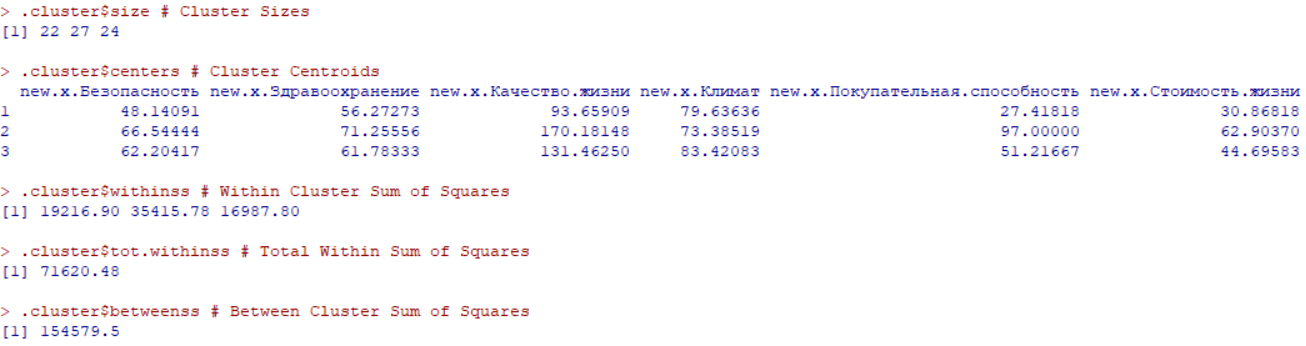


Рисунок 1 − Результат разбиения данных на три кластера

Из набора данных были взяты две выборки – обучающая и тестовая (шаг - 5). Для значений в обоих группах принадлежность к кластеру известна.

Dataset.train <- Dataset [seq (1, nrow(Dataset),5),]

Dataset.unknow <- Dataset [-seq (1, nrow(Dataset),5),]

С помощью функции lda() из библиотеки MASS по обучающей выборке были получены коэффициенты линейных дискриминантов, по ним построены дискриминантные функции.

Library(MASS)

dataset.lda <- lda (Dataset.train[, 1:6], Dataset.train[,7])

dataset.lda

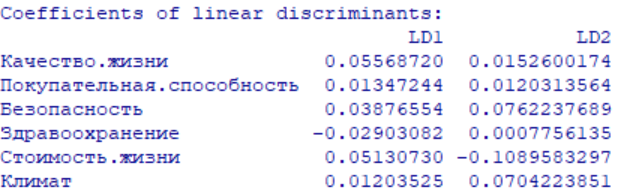


Рисунок 2 – Коэффициенты линейных дискриминантов

Была проведена классификация данных из второй выборки. Для проверки результата была построена матрица неточностей (рисунок 4).

dataset.ldap <- predict(dataset.lda, Dataset.unknow [, 1:6])$class

dataset.ldap

table (dataset.ldap, Dataset.unknow[,7])

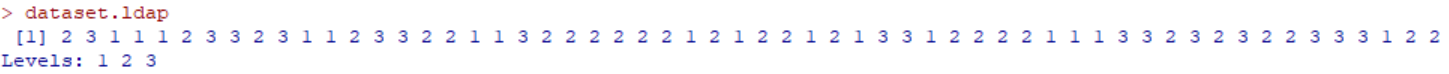


Рисунок 3 – Классификация тестовой выборки

По матрице видно, что тренировочная выборка привела к построению гипотезы, по которой были неправильно классифицированы 3 объекта первого класса, 1 объект второго класса и 8 объектов третьего класса. Процент ошибки составил 7%.

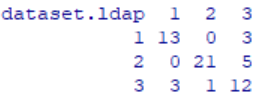


Рисунок 4 – Матрица неточностей классификации

Также для построенной модели были выведены ошибка распознавания и расстояние Махалонобиса.

Err\_S <- mean (Dataset.unknow[,7] != dataset.ldap)

mahDist <- dist(dataset.lda$means %\*% dataset.lda$scaling)

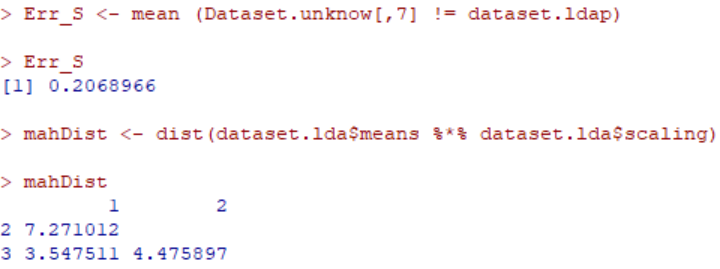


Рисунок 5 – Показатели оценки качества модели (по всем параметрам)

С помощью функции stepclass() из библиотеки klaR было проведено пошаговое построение дискриминантной модели. Процедурой были выбраны две переменные – Качество жизни и Покупательная способность.

library(klaR)

stepclass(Dataset[,1:6], Dataset[,7], method ="lda")

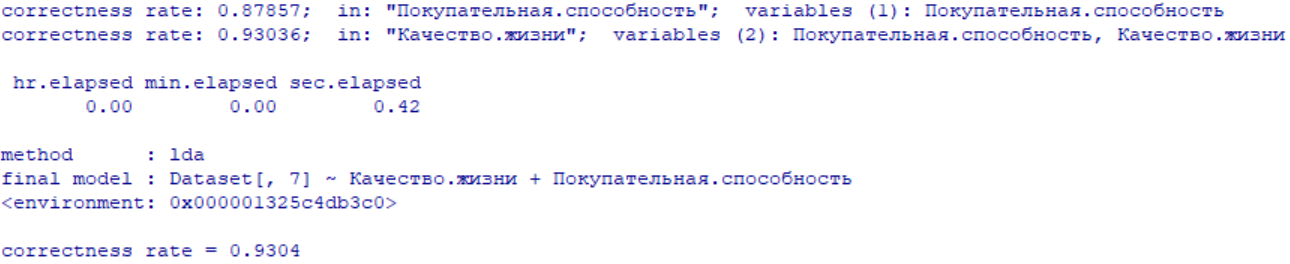


Рисунок 6 – Построение дискриминантной модели

По выбранным переменным были получены коэффициенты линейных дискриминантов, построены дискриминантные функции.

dataset.lda <- lda (Dataset.train[, 1:2], Dataset.train[,7])

dataset.lda

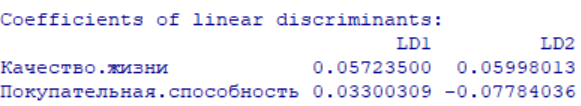


Рисунок 7 − Коэффициенты линейных дискриминантов новой модели

Данные из тестовой выборки были классифицированы по новым значениям дискриминантных функций.

dataset.ldap <- predict(dataset.lda, Dataset.unknow [, 1:2])$class

dataset.ldap

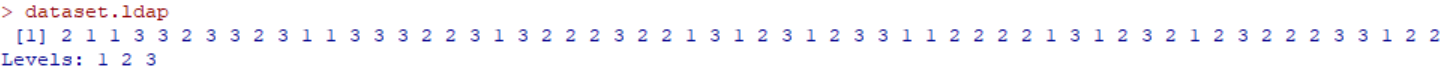


Рисунок 8 – Повторная классификация тестовой выборки

Для проверки результата была построена матрица неточностей (рисунок 9). По ней видно, что неправильно классифицированы были 2 объекта первого класса и 2 объекта третьего. Процент ошибки составил 2%, что сильно лучше предыдущей модели.

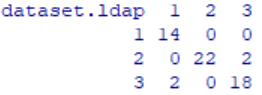


Рисунок 9 – Матрица неточностей для повторной классификации

Также были вычислены ошибка распознавания и расстояние Махалонобиса.

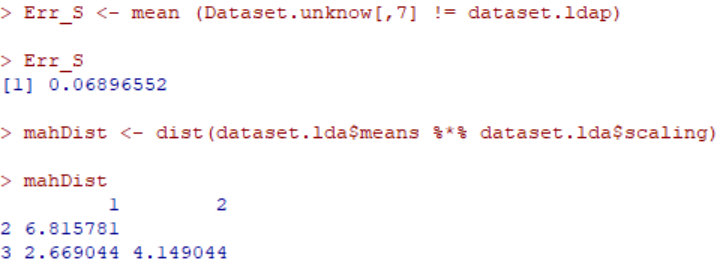


Рисунок 10 – Показатели оценки качества модели (по значимым параметрам)

В набор были добавлены новые неклассифицированные данные.

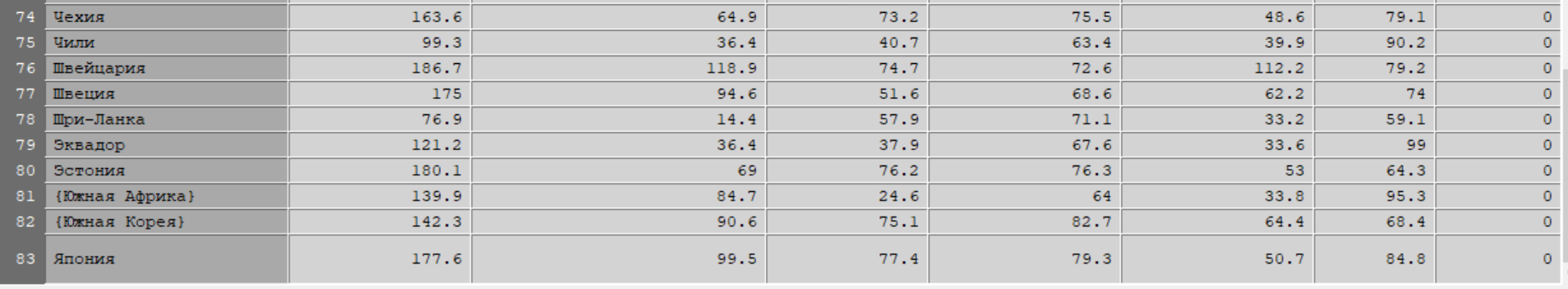


Рисунок 11 – Новые данные

Был выполнен дискриминантный анализ неклассифицированных данных. В итоге из 10 объектов 2 были отнесены к первому классу, 7 ко второму, 1 к третьему.

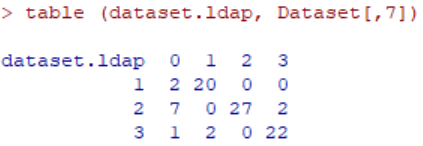


Рисунок 12 – Матрица неточностей с новыми данными

К первому классу отнесены 2 и 5 строки. Ко второму классу отнесены 1, 3, 4, 7, 8, 9, 10 строки. К третьему классу отнесена 6 строка.

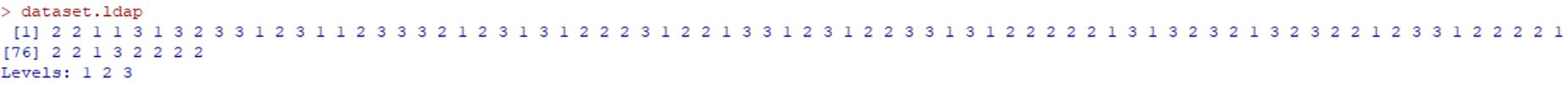


Рисунок 13 – Классификация набора с новыми данными

# ВЫВОД

В ходе работы исходные данные были разделены на 3 кластера. Из них была взята обучающая выборка, по которой был выполнен дискриминантный анализ и построены дискриминантные функции. По результатам анализа были классифицированы объекты тестовой выборки. В ходе проверки классификации было определено, что неправильно классифицированы были 3 объекта первого класса, 1 объект второго класса и 8 объектов третьего класса. Процент ошибки составил 7%.

Далее были выбраны наиболее существенные для анализа переменные: Индекс качества жизни и Покупательная способность. По ним для исходного набора была построена новая дискриминантная модель. В результате проверки классификации неправильно классифицированы были 2 объекта первого класса и 2 объекта третьего. Процент ошибки составил 2%.

Таким образом сужение числа рассматриваемых параметров и выбор наиболее информативных позволил увеличить процент верно классифицируемых объектов.

.