Лабораторная работа №5

«Линейный дискриминантный анализ. Построение канонических и классификационных функций»

1 ЦЕЛЬ РАБОТЫ

– Закрепить теоретические знания и приобрести практические навыки в проведении дискриминантного анализа по экспериментальным данным;

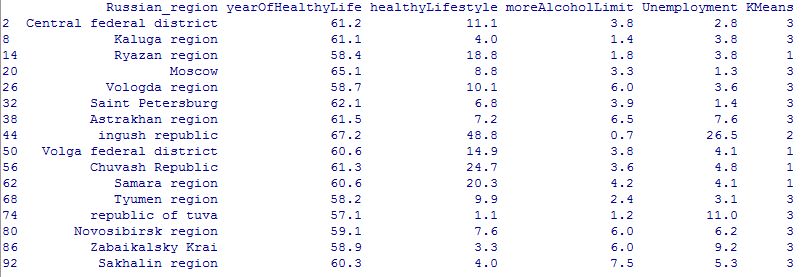
– Исследовать возможности языка R для проведения дискриминантного анализа.

2 ХОД РАБОТЫ

1. Импортируем экспериментальные данные в пакет Rcmdr. Для подготовки данных к применению дискриминантного анализа, проведём кластерный анализ методом k-средних на 3 кластера.

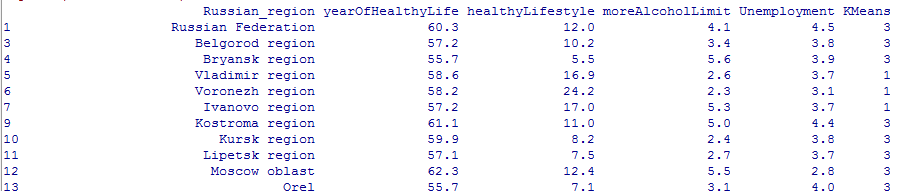
Затем создадим тренировочную выборку:





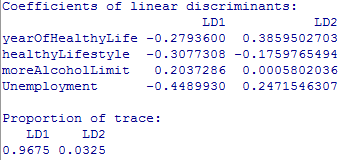
Оставшиеся данные сформируем в выборку для последующей проверки полученной классификации:





Проведём дискриминантный анализ с помощью функции lda() пакета MASS. В качестве переменных используем столбцы 2:5, классификацию проводим по 6 столбцу, в котором располагаются кластеры:

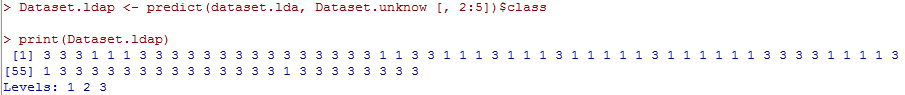




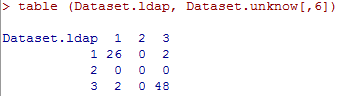
По результатам проведённого анализа построим дискриминантные функции, используя полученные коэффициенты линейных дискриминантов:

1. *z(x) =* –0.449*x1* + 0.204*x2* – 0.308*x3* – 0.279*x4*
2. *z(x) =* 0.247*x1* + 0.0006*x2* – 0.176*x3* + 0.386*x4*

Проведём классификацию и проверку оставшихся данных. Результаты классификации:



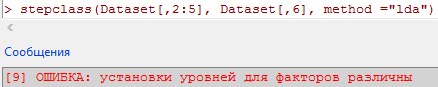
Матрица неточностей для проверки полученной модели:



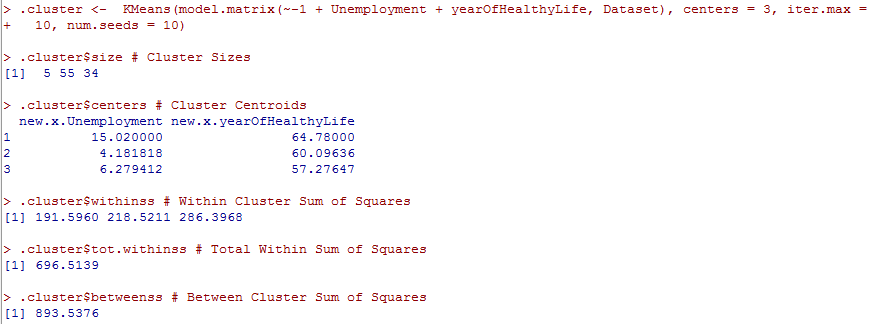
Проанализировав данную матрицу, можно сделать вывод, что тренировочная выборка привела к построению гипотезы, по которой 2 объекта первого класса попали не в «свою» группу и 2 объекта из третьего класса также попали не в «свою» группу.

Лабораторная работа 5.2

Произведём пошаговую процедуру выбора переменных для построения дискриминантной модели с помощью функции stepclass() из пакета klaR:

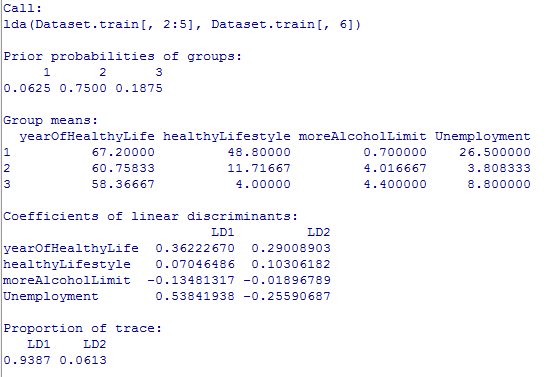


Так как возникла ошибка на данном этапе, выберем переменные самостоятельно. Проведём кластерный анализ для переменных yearOfHealthyLife и Unemployment:



Построим дискриминантную модель для выбранных переменных:

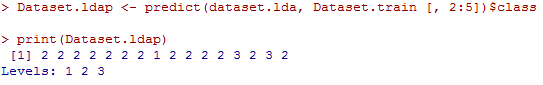




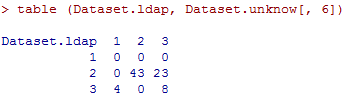
По результатам проведённого анализа построим дискриминантные функции, используя полученные коэффициенты линейных дискриминантов:

1. *z(x) =*  0.538*x1* – 0.135*x2* + 0.07*x3* + 0.362*x4*
2. *z(x) = –*0.256*x1* – 0.019*x2* + 0.103*x3* + 0.29*x4*

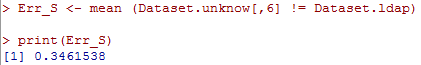
Классификация данных по выбранных переменным:



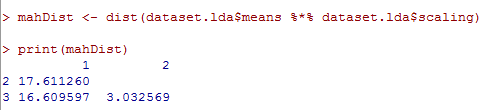
Выведем показатели оценки качества построенной модели. Матрица неточностей:



Ошибка распознавания:



Расстояние Махалонобиса:



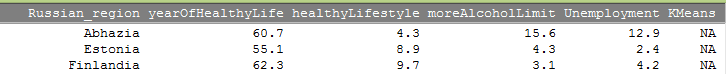
Проанализируем построенную модель. По матрице неточности можно сделать вывод, что тренировочная выборка привела к построению гипотезы, по которой значительное количество (23 объекта) второго класса попали не в «свою» группу и 4 объекта из третьего класса также попали не в «свою» группу, а в первом классе не оказалось объектов вовсе.

Ошибка распознавания также велика – 35%.

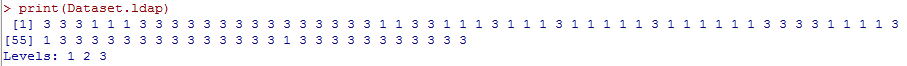
С помощью расстояния Махаланобиса можно определять сходство неизвестной и известной [выборки](https://ru.wikipedia.org/wiki/%D0%92%D1%8B%D0%B1%D0%BE%D1%80%D0%BA%D0%B0). Оно отличается от [расстояния Евклида](https://ru.wikipedia.org/wiki/%D0%95%D0%B2%D0%BA%D0%BB%D0%B8%D0%B4%D0%BE%D0%B2%D0%B0_%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0) тем, что учитывает [корреляции](https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D1%80%D1%80%D0%B5%D0%BB%D1%8F%D1%86%D0%B8%D1%8F) между переменными и инвариантно к масштабу. На основе полученных расстояний между центроидами классов, можно сделать вывод, что наименьшее расстояние между 2 и 3 классами.

Общий вывод данной модели: модель, построенная на основе выбранных переменных значительно хуже той, что была получена при использовании всех переменных в модели.

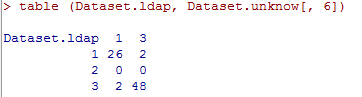
Добавим в выборку данные без классификации:



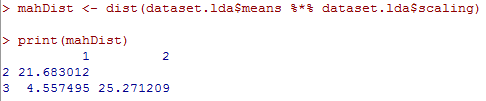
Результаты классификации и проверки данных, в которые вошли новые данные без классификации:



Матрица неточностей:



Расстояние Махалонобиса:



На основании Dataset.ldap можно сделать вывод, что новые данные причислены к 3 кластеру, так как последние 3 строки таблицы данных соответствуют трём крайним цифрам Dataset.ldap.

ВЫВОДЫ

В ходе данной лабораторной работы были закреплены теоретические знания и приобретены практические навыки в проведении дискриминантного анализа по экспериментальным данным. Также были исследованы возможности языка R для проведения дискриминантного анализа.