Министерство науки и высшего образования Российской Федерации
Севастопольский государственный университет
Кафедра ИС

Отчет
по лабораторной работе №5
«Линейный дискриминантный анализ. Построение канонических и классификационных функций»
по дисциплине
«ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ»

Выполнил студент группы ИС/б-17-2-о
Горбенко К. Н.
Проверил
Сырых О.А.

Севастополь
2020

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 ЗАДАНИЕ НА РАБОТУ

1. Подготовить данных для дискриминантного анализа. Для проведения дискриминантного анализ необходимо разделить исходных данных 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().

5. По полученным данным составить дискриминантную функцию.

6. Провести классификацию оставшихся данных и построить матрицу неточностей.

7. По полученным результатам сделать выводы.

8. Провести шаговую процедуру выбора переменных для построения дискриминантной модели.

9. Построить дискриминантную модель с выбранными переменными, составить уравнение дискриминантной функции.

10. Вывести показатели оценки качества построенной модели: матрица неточностей, ошибку распознавания, расстояние Махалонобиса.

11. Сделать выводы по построенной модели. Сравнить полученные результаты с моделью в которой использовались все переменные.

12. Добавить в выборку данные без классификации, используя дискриминантный анализ провести классификацию.

3 ХОД РАБОТЫ

3.1 Подготовка данных для дискриминантного анализа

С помощью команды проведем кластерный анализ методом k-средних на 3 кластера:

```
1 .cluster <-  KMeans(model.matrix(~-1 + health_index + stringency_index, Dataset
    ), centers = 3, iter.max = 10, num.seeds = 10)
2
3 Dataset$KMeans <- assignCluster(model.matrix(~-1 + health_index +
    stringency_index, Dataset), Dataset, .cluster$cluster)
```

Получим данные следующего вида (рисунок 1):

	country	total_cases_by_1m	deaths_by_1m	tests_by_1m	health_index	stringency_index	KMeans
1	Spain	16003	672	272117	82	85.19	2
2	Luxembourg	13328	197	1302265	81	79.63	1
3	Moldova	12698	323	68645	69	87.04	3
4	Belgium	9942	861	272178	80	81.48	2
5	Sweden	8989	581	151525	81	46.30	5
6	France	8309	487	164025	81	87.96	2
7	Belarus	8250	87	194234	72	19.44	5
8	BosniaandHerzegovina	8243	253	71121	70	92.59	4
9	Russia	7945	140	311066	68	85.19	3
10	Iceland	7791	29	804135	82	53.70	5
11	Portugal	7265	192	250565	79	87.96	2
12	Ireland	7145	364	233290	80	90.74	2

Рисунок 1 – Разбитые на кластеры данные

3.2 Создание тренировочной выборки

Создадим выборку строк от 1 до последней с шагом 4:

```
1 Dataset.train <- Dataset [seq (1, nrow(Dataset),4),]
```

Оставшиеся данные сформируем в выборку для последующей проверки полученной классификации:

```
1 Dataset.unknow <- Dataset [-seq (1, nrow(Dataset),4),]
```

3.3 Проведение дискриминантного анализа

В качестве переменных используем столбцы 2:6, классификация проводится по столбцу 7:

```
1 dataset.lda <- lda (Dataset.train[, 2:6], Dataset.train[,7])
```

3.4 Построение дискриминантных функций

Для построения дискриминантных функций из проведенного анализ используются Коэффициенты линейных дискриминантов:

```
Coefficients of linear discriminants:
                                LD1              LD2
total_cases_by_1m -0.000139703487 -0.000287021001
deaths_by_1m       0.002212666906  0.001877240490
tests_by_1m       -0.000009048967  0.000004366039
health_index       0.069692377719  0.128238825989
stringency_index  -0.746955798249  0.014550879210
```

Получаем две дискриминантные функции:

1. $z(x) = -0.747 * stringencyIndex + 0.0697 * healthIndex - 0.000 * * testsBy1m + 0.002 * deathsBy1m - 0.000 * totalCasesBy1m$
2. $z(x) = 0.015 * stringencyIndex + 0.128 * healthIndex + 0.000 * * testsBy1m + 0.002 * deathsBy1m - 0.000 * totalCasesBy1m$

3.5 Проведем классификацию и проверку оставшихся данных

```
> dataset.ldap <- predict(dataset.lda, Dataset.unknow[, 2:6])$class
> dataset.ldap
[1] 1 1 3 1 2 1 3 1 1 3 1 3 1 1 1 1 1 1 3 3 1 3 3 2 1 1 3
Levels: 1 2 3
```

Для проверки созданной модели построим матрицу неточностей:

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

dataset.ldap  1  2  3
             1 15  0  1
             2  0  1  1
             3  1  1  7
```

По данной матрице видно, что один объект первого класса попал в 3 группу; один объект второго класса попал в 3 группу; один объект 3 класса попал в 1 группу; один объект третьего класса попал во вторую группу.

Ошибка распознавания равна:

```
> Err_S <- mean (Dataset.unknow[,7] != dataset.ldap)
> Err_S
[1] 0.1481481
```

Расстояние Махалонобиса:

```
> mahDist <- dist(dataset.lda$means %*% dataset.lda$scaling)
> mahDist
          1         2
2 31.522542
3  7.282644 24.490570
```

3.6 Пошаговое построение дискриминантной модели

Проведём шаговую процедуру выбора переменных для построения дискриминантной модели. Для этого используем функцию stepclass() из пакета klaR. Результат выполнения этой функции представлен на рисунке 2.

```
> library(klaR)
> stepclass(Dataset[,2:6], Dataset[,7], method ="lda")
 `stepwise classification', using 10-fold cross-validated correctness rate of method lda'.
36 observations of 5 variables in 3 classes; direction: both
stop criterion: improvement less than 5%.
correctness rate: 0.94167;  in: "stringency_index";  variables (1): stringency_index

 hr.elapsed min.elapsed sec.elapsed
       0.00        0.00        0.16

method      : lda
final model : Dataset[, 7] ~ stringency_index
<environment: 0x000001b1df11c768>

correctness rate = 0.9417
```

Рисунок 2 – Пошаговое построение дискриминантной модели

Пошаговой процедурой построения модели была выбрана одна переменная – строгость реакции властей на пандемию. Для одной переменной невозможно построить дискриминантную модель.

3.7 Дискриминантный анализ для данных без классификации

Добавим в датасет несколько строк без классификации на кластеры (рисунок 3).

36	Latvia	902	19	165997	71	65.74	3
37	Japan	1252	12	321213	87	78.00	NA
38	Egypt	2532	64	63412	75	68.00	NA
39	Canada	7357	139	295312	84	75.00	NA
40	Ethiopia	4578	84	61352	63	56.00	NA
41	Australia	9523	39	92355	79	64.00	NA
42	Finland	5215	46	13264	83	67.00	NA
43	Greece	4792	52	126023	74	75.00	NA
44	Iraq	11502	129	86231	66	47.00	NA

Рисунок 3 – Добавленные в датасет записи

Выполним дискриминантный анализ только для новых данных:

```
> pred <- predict(dataset.lda, Dataset[37:44,2:6])$class
> pred
[1] 3 3 3 2 3 3 3 2
Levels: 1 2 3
```

Из новых 8 записей в 3 группу попали 6 объектов, во вторую – 2, в первую – 0.

ВЫВОДЫ

В ходе выполнения лабораторной работы была создана тренировочная выборка, выполнен дискриминантный анализ и проведена классификация, а для ее проверки была построена матрица неточностей. По полученной матрице видно, что тренировочная выборка привела к построению гипотезы, согласно которой 4 объекта попали не в «свою» группу.

Во второй части работы была проведена шаговая процедура выбора переменных для построения дискриминантной модели с помощью функции stepclass() из пакета klaR.