Министерство науки и высшего образования Российской Федерации

ФГАОУ ВО «Севастопольский государственный университет»

Институт информационных технологий

и управления в технических системах

Лабораторная работа №5

«Линейный дискриминантный анализ. Построение канонических и классификационных функций»

по дисциплине «Интеллектуальный анализ данных»

для студентов всех форм обучения направления подготовки

09.03.02 «Информационные системы и технологии»

Севастополь

2019

Линейный дискриминантный анализ. Построение канонических и классификационных функций. Методические указания к лабораторным занятиям по дисциплине «Интеллектуальный анализ данных» / Сост.: О.А. Сырых – Севастополь: Изд-во СевГУ, 2019 – 22 с.

Методические указания предназначены для проведения лабораторных работ по дисциплине «Интеллектуальный анализ данных». Целью методических указаний является помощь студентам в изучении возможностей системы RStudio. Излагаются практические сведения необходимые для выполнения лабораторной работы, требования к содержанию отчета.

Методические указания рассмотрены и утверждены на методическом семинаре и заседании кафедры «Информационные системы» (протокол № 1 от 29 августа 2019 г.)

Лабораторная работа №5_1
Линейный дискриминантный анализ. Построение канонических и классификационных функций.

Цель:

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

Время: 2 часа

Лабораторное оборудование: персональные компьютеры, выход в сеть Internet, RStudio.

Краткие теоретические сведения

Дискриминантный анализ

С помощью дикриминантного анализа на основании некоторых признаков (независимых переменных) объект может быть причислен к одной из двух или нескольких групп (число групп определяется числом категорий зависимой переменной). В двумерном дискриминантном анализе объекты относятся к одной из двух групп, например, купившие или не купившие данных продукт. А независимыми переменными в этом случае выступают возраст, доход покупателей, и др. показатели.

Дискриминантный анализ используется для анализа данных в том случае, когда зависимая переменная категориальная, а предикторы (независимые переменные) – интервальные. В результате дискриминантного анализа строится так называемая каноническая дискриминантная функция

$$d = b_1 x_1 + b_2 x_2 + ... + b_n x_n$$

где x_1 и x_n – значения дискриминантных переменных, соответствующих рассматриваемым случаям, $b_1 ... b_n$ – коэффициенты, которые и предстоит оценить с помощью дискриминантного анализа. Коэффициенты подбираются так, чтобы по значениям дискриминантной функции можно было с максимальной четкостью провести разделение по группам.

Процедура дискриминантного анализа состоит из пяти шагов. Первый шаг – формулирование проблемы, требует определения целей, зависимой и независимых переменных. Выборку делят на две части. Анализируемую выборку используют для вычисления дискриминантной функции; проверочную – для проверки достоверности модели. Второй шаг – определение функции, включает выведение такой линейной комбинации предикторов (дискриминантных функций), чтобы группы максимально возможно различались между собой значениями предикторов.

Определение статистической значимости представляет собой третий шаг. Она включает проверку нулевой гипотезы о том, что в совокупности средние всех дискриминантных функций во всех группах равны между собой. Если нулевую гипотезу отклоняют, то имеет смысл интерпретировать результаты.

Четвертый шаг – интерпретация дискриминантных весов или коэффициентов аналогична такой же стадии во множественном регрессионном анализе.

Пятый шаг – проверка достоверности. Она включает разработку классификационной матрицы. Дискриминантные веса, определенные с помощью анализируемой выборки, умножают на значения независимых переменных в проверочной выборке, чтобы получить дискриминантные показатели для случаев в этой выборке. Затем случаи распределяют по группам, исходя из дискриминантных показателей и соответствующего правила принятия решения. Определяют процент верно классифицированных случаев и сравнивают его с процентом случаев, которое можно ожидать на основе классификации методом случайного выбора.

Для оценки коэффициентов существует два известных подхода. Прямой метод включает оценку дискриминантной функции при одновременном введении всех предикторов. Альтернативный ему пошаговый метод включает последовательное введение предсказанных переменных, исходя из их способности дискриминировать группы.

Задание и порядок выполнения лабораторной работы №5_1

Проведение дискриминантного анализа и интерпретация результатов в среде R

Дискриминантный анализ реализован в нескольких пакетах для R, в данной работе будет рассмотрено применение функции lda() из базового пакета MASS.

Все процедуры дискриминантного анализа можно разбить на две группы: первая группа позволяет интерпретировать различия между имеющимися группами (сравнивая средние), вторая – проводить классификацию новых объектов в тех случаях, когда неизвестно заранее, к какому из существующих классов они принадлежат.

1. Подготовка данных для дискриминантного анализа. Для проведения дискриминантного анализ необходимо иметь разделение исходных данных на группы (классы). В данной работе в качестве классов (групп) возьмем разбиение выборки на кластеры.
2. Создать тренировочную выборку из исходных данных с известной группировкой.

Для того чтобы работать с методами классификации с обучением, надо сначала освоить технику «обучения». Для этого выбирается часть данных с известной групповой принадлежностью. На основании анализа этой части (тренировочной выборки) строится гипотеза о том, как должны распределяться по группам остальные, неклассифицированные данные

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

$$d = b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

6. Провести классификацию оставшихся данных и построить матрицу неточностей.
7. По полученным результатам сделать выводы.

Пример. Провести дискриминантный анализ и проверку построенной модели на экспериментальных данных.

1. Подготовка данных для дискриминантного анализа:

Загрузить файл Данные.xls

Провести кластерный анализ методом k-средних на 3 кластера (согласно проведенному анализу в лабораторной работе 4)

Результаты разбиения на кластеры добавить к данным.

2. Создание тренировочной выборки

Создадим выборку строк от 1 до последней с шагом 5

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

Оставшиеся данные сформируем в выборку для последующей проверки полученной классификации

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

3. Проведение дискриминантного анализа.

В качестве переменных используются столбцы 2:7, классификация проводится по столбцу 8:

```
dataset.lda <- lda (Dataset.train[, 2:7], Dataset.train[,8])
```

4. Построение дискриминантных функций:

Для построения дискриминантных функций из проведенного анализ используются Коэффициенты линейных дискриминантов

```
Coefficients of linear discriminants:
                              LD1       LD2
Х.ожидаемая.продол.ть -1060.816 -613.7209
образовании           -1047.838 -677.6436
бедность              -1120.301 -682.0631
безработица           -1079.061 -670.3226
реальное.ВВП.         -1194.556 -675.2506
общест...ое.развитие   5585.588 3383.7527
```

Исходя из результатов получаем две дискриминантные функции

1) $z(x)=5586x_1 - 1195x_2 - 1079x_3 - 1120x_4 - 1048x_5 - 1061x_6$

2) $z(x)=3384x_1 - 675x_2 - 670x_3 - 682x_4 - 678x_5 - 614x_6$

5. Провести классификацию и проверку оставшихся данных:

```
dataset.ldap <- predict(dataset.lda, Dataset.unknow [, 2:7])$class
```

Результат полученной классификации:

```
> dataset.ldap
 [1] 2 2 2 1 2 2 2 2 2 1 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[42] 2 2 2 1
Levels: 1 2 3
```

Для проверки созданной модели построим матрицу неточностчей:

```
table (dataset.ldap, Dataset.unknow[,8])
```

Результат:

dataset.ldap	1	2
1	3	1
2	6	31
3	0	0

По данной матрице видно, что тренировочная выборка привела к построению гипотезы, по которой в 6 объектов первого класса попали не в «свою» группу и 1 объект из второго класса.

Библиография

1. Алексей Шипунов и др. Наглядная статистика. Используем R! – М.: ДМК Пресс, 2014. – 298 с. [Электронный ресурс]. Режим доступа: http://ashipunov.info/shipunov/school/books/rbook.pdf.
2. Зарядов И.С. Введение в статистический пакет R: типы переменных, структуры данных, чтение и запись информации, графика. М.: Издательство Российского университета дружбы народов, 2010. – 207 с.

3. Роберт И. Кабаков R в действии. Анализ и визуализация данных в программе R / пер. с англ. Полины А. Волковой. – М.: ДМК Пресс, 2014. – 588 с.: ил.
4. Оффициальный сайт RStudio. Режим доступа: https://www.rstudio.com.
5. Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных [Электронный ресурс]. Режим доступа: http://machinelearning.ru.
6. Мастицкий С.Э., Шитиков В.К. (2014) Статистический анализ и визуализация данных с помощью R. – Электронная книга. Режим доступа: http://r-analytics.blogspot.com