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 – 5 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 №5.2
Линейный дискриминантный анализ. Проведение дискриминантного анализа и интерпретация результатов.

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

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

где x_1 и x_n – значения дискриминантных переменных, соответствующих рассматриваемым случаям, $b_1 \ldots b_n$ – 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_2

Проведение дискриминантного анализа и интерпретация результатов в среде R

При проведении дискриминантного анализа может возникнуть вопрос, какие из имеющихся признаков являются информативными при разделении, а какие – сопутствующим балластом.

По построенной модели необходимо выводить важные показатели для оценки ее качества: матрицы неточностей на обучающей выборке, ошибку распознавания и расстояние Махалонобиса между центроидами двух классов

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

Шаговая процедура выбора переменных при классификации, реализованная функцией stepclass() из пакета klaR, основана на вычислении сразу четырех параметров качества моделей-претендентов:

а) индекса ошибок (correctness rate),

б) точности (accuracy), основанной на евклидовых расстояниях между векторами "факта" и "прогноза",

в) способности к разделимости (ability to seperate), также основанной на расстояниях,

г) доверительных интервалах центроидов классов.

```
stepclass(Dataset[,2:7], Dataset[,8], method ="lda")
```

Все эти параметры оцениваются в режиме многократной перекрестной проверки.

2. Построить дискриминантную модель с выбранными переменными, составить уравнение дискриминантной функции.
3. Вывести показатели оценки качества построенной модели: матрица неточностей, ошибку распознавания, расстояние Махалонобиса.

```
table (dataset.ldap, Dataset.unknow[,8])
Err_S <- mean (Dataset.unknow[,8] != dataset.ldap)
mahDist <- dist(dataset.lda$means %*% dataset.lda$scaling)
```

4. Сделать выводы по построенной модели. Сравнить полученные результаты с моделью в которой использовались все переменные.
5. Добавить в выборку данные без классификации, используя дискриминантный анализ провести классификацию.

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