**Лабораторная работа № 7**

**Дискриминантный анализ в системе STATISTICA**

Знакомство с возможностями проведения дискриминантного анализа в системе STATISTICA лучше всего начать с разбора апробированного примера. Таким примером является классический **пример Фишера** – анализ цветков ириса.

Задача состоит в том, чтобы по результатам измерения длины и ширины чашелистиков и лепестков цветков ириса отнести ирис к одному из трех типов: SETOSA, VERSICOL, VIRGINIC.

Данные для этого примера имеются в файле Irisdat.sta. В файле содержатся результаты измерений 150 цветков ириса, по 50 каждого типа.

**Шаг 1**. Нажмите кнопку **Open Data (открыть данные)** и откройте файл данных *Irisdat.sta* из каталога **Examples (примеры)**. Следующий файл данных появится на экране (рис. 1).



*Рис. 1 – Файл данных Iris.sta*

**Шаг 2.** Из **Переключателя модулей** STATISTICA откройте стартовую панель модуля **Discriminate function analysis** **(дискриминантный функциональный анализ)** (рис. 2).



*Рис. 2 –Стартовая панель модуля* ***Дискриминантный анализ***

**Шаг 3.** Нажмите кнопку **Variables (переменные)** и выберите переменные для анализа. В качестве **Grouping variable (группирующая переменная)** выберите переменную **IRISTYPE (тип ириса)**. В качестве **Independent variables (независимые переменные)** выберите переменные **ДЛЧАШЕЛ**, **ШИРЧАШЕЛ**, **ДЛЛЕПЕСТ**, **ШИРЛЕПЕСТ**.

Сделайте установки, как показано на рис. 2.

**Шаг 4**. Нажмите кнопку **ОK** и откройте диалоговое окно **Model Definition (определение модели)** – рис. 3.



*Рис. 3 – Окно определения модели дискриминантного анализа*

Сделайте установки, как показано на рис. 3. Нажмите кнопку **OK** и запустите вычислительную процедуру, реализующую пошаговый метод включения.

**Шаг 5**. Всесторонне просмотрите итоги в диалоговом окне **Discriminant Function Analysis Results (результаты дискриминантного анализа)** (рис. 4).



*Рис. 4 – Окно результатов дискриминантного анализа данных из файла Iris.sta*

Информационная часть окна сообщает, что использован:

- **Stepwise analysis (пошаговый анализ), Step 4 Final step (шаг 4 заключительный шаг)**;

- **Number of variables in the model (число переменных в модели): 4**;

- **Last variable entered (последняя включенная переменная): ДЛЧАШЕЛ**, соответствующее значение статистики **F-критерия F(2, 144) = 4,72**, уровень значимости **р < 0,0103**;

- **Wilks lambda (значение лямбды Уилкса): 0,0234**;

- **approx. F (4,292) = 199,1454 (приближенное значение F- статистики)**, связанной с **лямбдой Уилкса**;

- **р** – уровень значимости **F-критерия** для значения **199,1454**;

- значения статистики **лямбда Уилкса** лежат в интервале [0, 1].

Значения статистики Уилкса, лежащие около 0, свидетельствуют о хорошей дискриминации. Значения статистики Уилкса, лежащие около 1, свидетельствуют о плохой дискриминации. Иными словами, это можно выразить следующим образом: если значения лямбды Уилкса близки к 0, то мощность дискриминации (мощность = 1– вероятность ошибки) близка к 1, если лямбда Уилкса близка к 1, то мощность близка к 0.

**Шаг 6**. Нажмите кнопку Variables in the model (переменные, включенные в модель). На экране появится итоговая таблица анализа (рис. 5).



*Рис. 5 – Итоговая таблица анализа данных из файла iris.sta*

**Шаг 7**. Просмотрите разделение групп на графике. Для этого инициируйте кнопку **Perform canonical analysis (Канонический анализ)**. В появившемся диалоговом окне **Canonical Analysis (канонический анализ)** нажмите кнопку **Scatterplot of canonical scores (диаграмма рассеяния канонических значений)**. На экране появится следующий график (рис. 6).



*Рис. 6 – Разделение трех типов ириса*

**Шаг 8**. Просмотрите функции классификации. В диалоговом окне **Результаты дискриминантного анализа** нажмите кнопку **Classification functions (функции классификации)** (рис. 7).



*Рис. 7 – Функции классификации, построенные пошаговым методом вперед (Forward stepwise)*

С помощью этих функций можно вычислить классификационные значения (метки) для вновь наблюдаемых цветков по формулам:

**SETOSA = – 16,43\*ДЛ + 23,69\*ШЧ – 17,4\*ШЛ + 23,54\*ДЧ – 86,31;**

**VERSICOL = 5,21\*ДЛ + 7,07\*ШЧ – 6,43\*ШЛ + 15,70\*ДЧ – 72,85;**

**VIRGINIC = 12,76\*ДЛ + 3,69\*ШЧ – 21,08\*ШЛ + 12,5\*ДЧ – 104,37,**

где:

**- ДЛ – ДЛЛЕПЕСТ;**

**- ШЧ – ШИРЧАШЕЛ;**

**- ШЛ – ШИРЛЕПЕСТ;**

**- ДЧ – ДЛЧАШЕЛ.**

Пусть вы имеете новый цветок со значениями: ДЛЛЕПЕСТ, ШИРЧАШЕЛ, ШИРЛЕПЕС, ДЛЧАШЕЛИ.

К какому типу ириса его отнести? Формально следует подставить эти значения в приведенные выше формулы и вычислить классификационные значения SETOSA, VERSICOL, VIRGINIC.

Новый цветок относится к тому классу, для которого классификационное значение максимально. Конечно, построенные классификационные функции могут быть определены в электронных таблицах как формулы, и для каждого добавленного случая по ним могут быть вычислены классификационные метки. Таким образом, каждый новый объект автоматически относится к определенному классу.

**Шаг 9.** **Расстояния Махаланобиса**

Нажмите кнопку **Squared Mahalanobis distance (квадрат расстояния Махаланобиса)** и вы увидите таблицу с квадратами расстояния Махаланобиса от точек (случаев) до центров групп (рис. 8).



*Рис. 8 – Расстояния Махаланобиса для данных из файла Iris.sta*

Случай относится к группе, до которой расстояние Махаланобиса минимально.

**Шаг 10**. **Апостериорные вероятности**.

Рассмотрите группу опций справа внизу диалогового окна **Результаты дискриминантного анализа: A priori classifications probabilities (априорные вероятности классификации)**. До анализа вы задаете для каждого случая (в данном примере цветка) вероятность, с какой он принадлежит к определенному классу. После того как анализ выполнен, можно пересчитать эти вероятности и получить апостериорные вероятности классификации. Нажав кнопку **Posterior probabilities (апостериорные вероятности)**, вы увидите таблицу с апостериорными вероятностями принадлежности объекта к определенному классу (рис. 9).



*Рисунок 9 – Таблица апостериорных вероятностей*

Интерпретация данной таблицы очень проста. В первом столбце указан тип ириса для каждого случая. Во втором, третьем, четвертом столбцах даны апостериорные вероятности отнесения каждого цветка к определенному типу.

Цветок относится к группе с максимальной апостериорной вероятностью.

Знаком \* отмечаются неправильно классифицированные при использовании данного правила случаи (5, 9, 12).

**Шаг 11.** **Классификация новых случаев**.

Не закрывая диалога Результаты дискриминантного анализа, добавьте в таблицу исходных данных новый случай (рис. 10).



*Рисунок 10 – Новое наблюдение в данных Iris.sta*

Для того чтобы понять, к какому классу относится этот объект, нажмите кнопку **Posterior probabilities (апостериорные вероятности)**, вы увидите ту же таблицу с постериорными вероятностями, к которой будет добавлена строка (рис. 11).



*Рис. 11 – Классификация нового наблюдения*

Итак, новое наблюдение с вероятностью 0,999 можно отнести к типу SETOSA.

Нажмите кнопку **Squared Mahalanobis distance (квадрат расстояния Махаланобиса)**, и вы увидите таблицу (рис. 12) с квадратами расстояния Махаланобиса. В последней строке этой таблицы показаны расстояния нового случая до групповых центров.



*Рис. 12 – Расстояние Махаланобиса от нового наблюдения до центров групп*

Опять расстояние от нового наблюдения до центра групп минимально для группы SETOSA. Следовательно, с высокой степенью вероятности новый цветок – это ирис типа SETOSA.

## Задание 1

Создайте пустую электронную таблицу Spreadsheet.sta. Внесите данные для выполнения расчетов, находящиеся в Приложении. Выполните процедуры дискриминантного анализа. Дайте объяснение полученным результатам.

**Задание 2** Из таблицы внесите новые данные. Запустите процедуру дискриминантного анализа. Внесите поправки в исходную таблицу, пополнив обучающую выборку новой информацией.

Таблица – Классификация новых случаев

|  |  |  |
| --- | --- | --- |
| 1 задание | 2 задание | 3 задание |
| Класс | 1 призн. | 2 призн. | 3 призн. | 4 призн. | Класс | 1 призн. | 2 призн. | Класс | 1 призн. | 2 призн. |
|  | 1,14 | 1,26 | 0,99 | 2,06 |  | 0,738 | 0,658 |  | 36,63 | 31,29 |
|  | 0,79 | 0,84 | 1,17 | 2,72 |  | 0,612 | 0,243 |  | 24,84 | 19,63 |
|  | 1,01 | 1,16 | 1,06 | 1,4 |  | 0,774 | 0,233 |  | 17,78 | 13,00 |
|  | 0,97 | 1,11 | 0,73 | 0,98 |  | 0,933 | 0,271 |  | 5,17 | 1,92 |

**Приложение**

|  |  |  |
| --- | --- | --- |
| 1 задание | 2 задание | 3 задание |
| Класс | 1 призн. | 2 призн. | 3 призн. | 4 призн. | Класс | 1 призн. | 2 призн. | Класс | 1 призн. | 2 призн. |
| a | 1,01 | 0,48 | 0,87 | 1,12 | d | 0,247 | 0,295 | a | 9,23 | 5,26 |
| a | 1,14 | 1,11 | 1,38 | 1,17 | d | 0,491 | 0,495 | c | 19,82 | 14,00 |
| a | 1,22 | 1,44 | 1,02 | 0,79 | c | 0,768 | 0,240 | a | 11,59 | 8,06 |
| a | 1,09 | 0,94 | 0,91 | 1,1 | c | 0,838 | 0,354 | a | 11,06 | 6,99 |
| a | 0,65 | 0,89 | 1,09 | 1,12 | c | 0,921 | 0,320 | a | 7,58 | 2,83 |
| a | 1,21 | 1,22 | 1,11 | 1,4 | c | 0,837 | 0,409 | a | 9,35 | 5,73 |
| a | 0,95 | 1,04 | 0,76 | 1,15 | b | 0,642 | 0,309 | a | 10,27 | 6,24 |
| b | 2,01 | 1,24 | 0,78 | 0,8 | b | 0,754 | 0,288 | d | 25,95 | 20,80 |
| b | 1,64 | 0,93 | 0,96 | 1,48 | c | 0,844 | 0,340 | a | 11,04 | 6,83 |
| b | 2,13 | 1,38 | 1,15 | 0,93 | d | 0,433 | 0,242 | b | 16,03 | 11,53 |
| a | 1,41 | 0,85 | 1,15 | 0,73 | c | 0,923 | 0,289 | e | 31,41 | 26,44 |
| a | 1,22 | 1,03 | 1,1 | 0,87 | c | 0,824 | 0,243 | b | 16,92 | 11,96 |
| b | 2,81 | 1,13 | 1 | 0,65 | c | 0,963 | 0,335 | b | 16,02 | 11,47 |
| b | 2,05 | 1,25 | 0,91 | 0,88 | c | 0,800 | 0,223 | b | 17,74 | 12,96 |
| b | 2,31 | 0,87 | 1,14 | 1,37 | a | 0,828 | 0,696 | b | 17,00 | 11,96 |
| b | 1,8 | 0,84 | 0,94 | 0,89 | c | 1,000 | 0,366 | b | 18,36 | 13,36 |
| a | 1,15 | 1,27 | 1,25 | 1,16 | b | 0,715 | 0,354 | b | 15,73 | 11,25 |
| a | 0,74 | 2 | 0,86 | 0,85 | a | 0,885 | 0,573 | b | 17,98 | 13,18 |
| a | 0,89 | 2,05 | 0,9 | 0,87 | b | 0,658 | 0,278 | b | 17,33 | 12,11 |
| a | 0,68 | 1,39 | 0,71 | 0,99 | a | 0,739 | 0,632 | a | 12,62 | 8,28 |
| a | 0,94 | 2,07 | 1,09 | 0,78 | b | 0,684 | 0,357 | c | 21,71 | 16,34 |
| a | 1,06 | 1,92 | 1,01 | 0,82 | a | 0,684 | 0,592 | c | 22,13 | 16,88 |
| a | 0,76 | 2,27 | 1,09 | 1,17 | d | 0,512 | 0,428 | e | 30,27 | 26,10 |
| a | 1,12 | 1,86 | 0,9 | 1,26 | a | 0,763 | 0,777 | c | 20,92 | 15,50 |
| a | 1,11 | 1,9 | 1,33 | 0,93 | d | 0,452 | 0,339 | e | 30,17 | 25,93 |
| a | 1,03 | 1,25 | 0,91 | 1,02 | d | 0,468 | 0,428 | c | 22,83 | 17,16 |
| a | 1,57 | 2,1 | 0,94 | 1,09 | c | 0,849 | 0,298 | c | 21,02 | 15,63 |
| a | 1,05 | 0,56 | 1,82 | 0,95 | c | 0,761 | 0,252 | c | 21,84 | 16,83 |
| a | 0,92 | 0,93 | 2,17 | 1,17 | d | 0,298 | 0,201 | e | 31,82 | 26,62 |
| a | 1,11 | 0,9 | 2,36 | 1,03 | c | 0,820 | 0,438 | d | 25,06 | 19,97 |
| a | 1,44 | 1,29 | 2,11 | 0,95 | c | 0,886 | 0,252 | d | 26,06 | 21,15 |
| a | 0,81 | 1,03 | 2,36 | 1,12 | a | 0,761 | 0,534 | c | 20,60 | 15,17 |
| a | 0,69 | 1,29 | 1,79 | 0,92 | c | 0,827 | 0,361 | d | 24,51 | 19,51 |
| a | 1,04 | 0,83 | 2,76 | 1,26 | a | 0,894 | 1,000 | e | 28,94 | 23,35 |
| a | 0,98 | 1,17 | 1,62 | 0,93 | d | 0,381 | 0,459 | d | 26,12 | 21,25 |
| a | 1,23 | 0,88 | 2,69 | 0,63 | b | 0,750 | 0,301 | d | 23,90 | 18,94 |
| a | 1,25 | 1,11 | 2,52 | 1,38 | a | 0,613 | 0,512 | d | 23,74 | 18,34 |
| c | 1,26 | 0,95 | 0,66 | 2,44 | d | 0,443 | 0,369 | d | 25,30 | 20,05 |
| c | 1,03 | 1,27 | 1,39 | 2,15 | c | 0,853 | 0,231 | c | 20,51 | 15,11 |