|  |  |
| --- | --- |
| ФБМІ НТУУ «КПІ»  Кафедра БМК  Дисципліна:  Біомедична кібернетика-2. Моделювання складних систем | Виконав: Войник Б.О.  Факультет\_ФБМІ\_гр. БС-71мп  Дата\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Бали\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Прийняв /\_\_\_\_\_\_\_\_/\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

**Комп’ютерний практикум №1**

**Тема:** «Статичні методи класифікації»

**Мета роботи:** побудова та порівняння моделей, отриманих в результаті класифікації

**Завдання**

1. Проведення логістичної регресії
2. Проведення дискримінантного аналізу
3. Побудова статистичних моделей
4. Застосування методів поліпшення якості класифікаторів

**Логістична регресія**

Регресійний аналіз один із розділів математичної статистики та являє собою метод моделювання вимірюваних даних і дослідження їх властивостей. Дані складаються з пар значень залежної змінної (змінної відгуку) і незалежної змінної (пояснюватиме змінної). Регресійна модель являє собою функцію незалежної змінної і параметрів із додаванням випадкової змінної. Параметри моделі налаштовуються таким чином, щоб модель найкращим чином наближала дані. Критерієм якості наближення (цільовою функцією) зазвичай є середньоквадратична помилка: сума квадратів різниці значень моделі і залежною змінною для всіх значень незалежної змінної в якості аргументу. Залежна змінна є сумою значень деякої моделі і випадкової величини. Щодо характеру розподілу цієї величини робляться припущення, які називаються гіпотезами породження даних. Для підтвердження або спростування цієї гіпотези виконуються статистичні тести, так званий аналіз залишків. При цьому передбачається, що незалежна змінна не містить помилок. Регресійний аналіз використовується для прогнозу, аналізу часових рядів, тестування гіпотез і виявлення прихованих взаємозв'язків між даними. Бінарна логістична регресія є однією з ризновидів регресійного аналізу.

За допомогою методу бінарної логістичної регресії можна дослідити залежність дихотомічних змінних від незалежних змінних, які мають будь-який вид шкали.

Як правило, у випадку з дихотомічними змінними йдеться про деяку подію, яка може відбутися або не відбутися; бінарна логістична регресія в такому випадку розраховує ймовірність настання події в залежності від значень незалежних змінних.

Ймовірність настання події для деякого випадку розраховується за формулою (1) та формулою (2):

|  |  |
| --- | --- |
| , | (1) |

де

|  |  |
| --- | --- |
| *z=b1\*X1 + b2\*Х2 + ...+ bn\*Xn + a*, | (2) |

де *X1* - значення незалежних змінних;

*b1* - коефіцієнти, розрахунок яких є завданням бінарної логістичної регресії;

*а* - деяка константа.

Якщо для *р* вийде значення менше 0,5, то можна припустити, що подія не настане; в іншому випадку передбачається настання події. Розрахована ймовірність *р* завжди вказує на виконання пророкування, яке відповідає більшій з двох кодувань залежних змінних.

Через залучення до аналізу великої кількості змінних комп'ютер повинен вирішити, які з них в кінцевому випадку будуть відібрані для використання в рівнянні ймовірності. Тому потрібно вибирати не метод вкладення, який включає в розрахунок всі змінні, а один з покрокових методів.

Метод прямої селекції починається з використання одних лише констант на стартовому етапі, а потім послідовно підключаються змінні, які демонструють сильну кореляцію з залежними змінними. Далі знову слідує перевірка того, які змінні повинні бути виключені, причому як критерій перевірки вибирається або статистика Вальдовского (Wald), або функція правдоподібності, або один з варіантів, які називаються умовною статистикою. Метод зворотної селекції спочатку бере до уваги всі змінні, а потім у зворотному порядку відбувається виключення малозначущих змінних.

Кількість утворюваних "фіктивних" дихотомічних змінних має бути завжди на 1 менше, ніж число кількість заданих категорій. Категорія, яка виявилася зайвою, називається еталонної категорією і, відповідно до попередніх установками, є останньою категорією. За допомогою поля контрастів (Contrast) можна управляти особливостями залучення в аналіз освічених фіктивних змінних; при контрасті рівному Deviation (Відхилення) всі категорії, крім еталонної, будуть перевіряються щодо сумарного ефекту.

**Проведення розрахунків**

Для побудови моделей було використано базу даних студентів (чоловіків), що пройшли пробу Мартіне більше одного разу. База даних містить показники артеріального тиску та пульсу в стані спокою та за п’ять хвилин після навантаження

Оскільки класифікатор студентів є не бінарною змінною, тому було прийнято рішення розбити пацієнтів на групи методом «один проти всіх» та вирівняти дані в групах, де були отримані показники класифікації занадто асиметричні.

На рис. 1 зображено встановлення параметрів, які водилися для отримання класифікації даних.

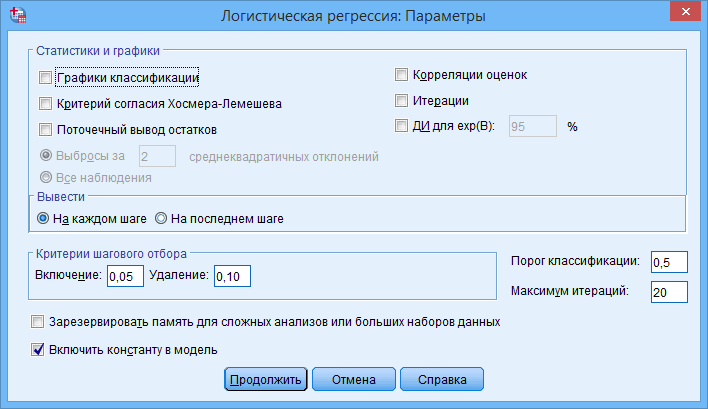
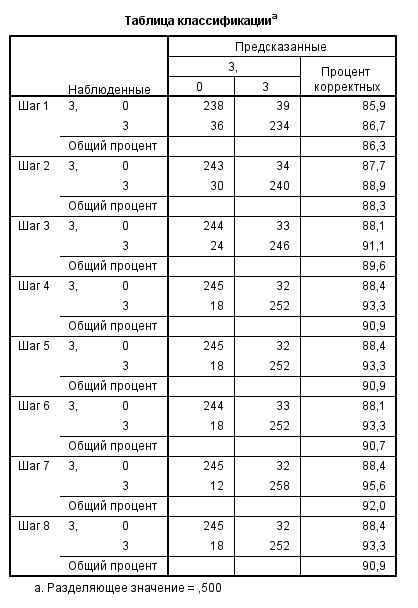


Рисунок 1 – Зображення параметрів

**Розбиття на класи «один проти всіх»**

У таблиці 1 наведені результати класифікації при об’єднанні у групи: перша група – студенти, що знаходяться в 3 кластері, друга – студенти в інших кластерах (3 проти 1, 2, 4, 5, 6, 7)

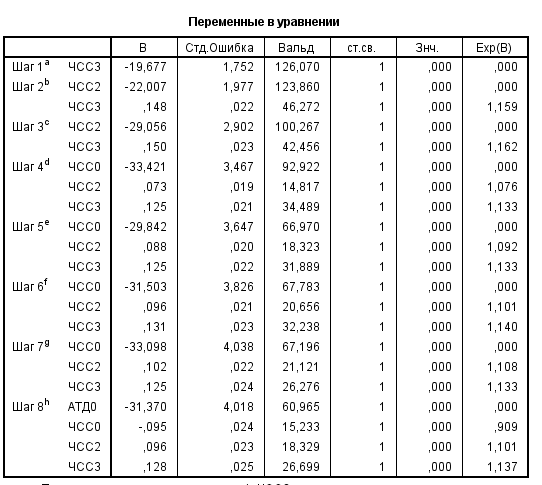
Таблиця 1 – Результати класифікації ЛР



За результатами дослідження ми бачимо, що відсоток коректно спрогнозованих даних складає 90,9%.

Для побудови рівняння регресії було використано наступну таблицю:

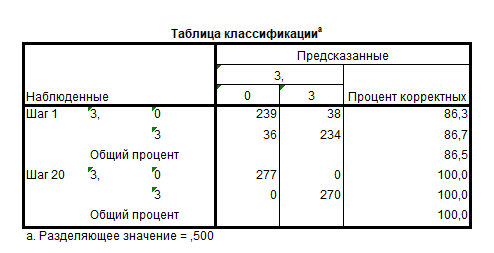
Таблиця 2. Змінні для рівняння регресії



Таким чином, рівняння регресії набуває вигляду:

Для поліпшення якості класифікаторів було вирішено розширити матрицю змінних за допомогою нелінійних перетворень. Тому необхідним кроком було встановлення в параметрах аналізу покрокового режиму для відбору змінних

Таблиця 3 – Результати класифікації ЛР з нелінійними перетвореннями



За результатами дослідження ми бачимо, що відсоток коректно спрогнозованих даних складає 100,0%. Але слід зазначити, що це при умові неповної моделі, оскільки при обробці даних SPSS видав наступне попередження:

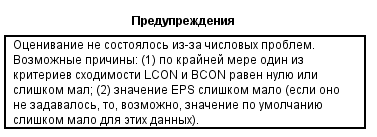
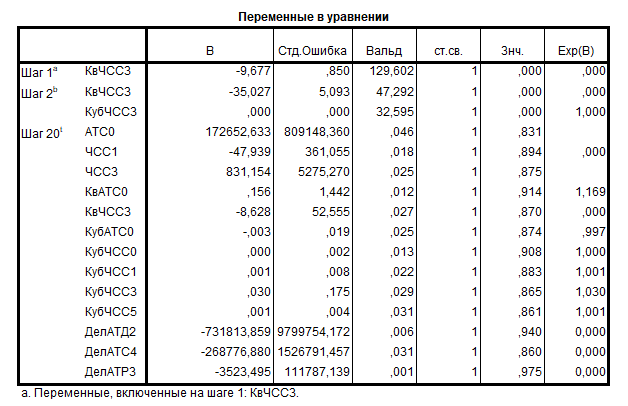


Рисунок 2 – попередження

Таблиця 4. Змінні для рівняння регресії з нелінійними перетвореннями



Таким чином, рівняння регресії має вигляд:

Порівнюючи дві моделі ми можемо дійти висновку, що модель, яка побудована на базі даних з додатковими змінними нелінійних перетворень є більш складною, хоча й не повною, але водночас дає вищий результат в порівнянні з моделлю, що включають лише істинні змінні.

**Дискримінантний аналіз**

Дискримінаційний чи дискримінантний аналіз використовується в тому випадку, якщо є дані, класифіковані на кілька груп, і необхідно знайти одну або більше функцій кількісних вимірів, які допоможуть віднести спостереження до однієї з цих груп. В дискримінантному аналізі розрізняють дві мети:

1. інтерпретація;
2. класифікація.

Метою інтерпретації є визначення кількості, значущості дискримінантних функцій і їх значень для пояснення відмінностей між класами. Метою класифікації є визначення класу, до якого належить новий об'єкт – це і є мета даної роботи.

В дискримінаційному аналізі, на відміну від кластерного, є навчальна вибірка, в якій відомо до яких класів відносяться об'єкти. За навчальною вибіркою необхідно отримати правила, які в подальшому дозволять визначити, до якого класу відносяться нові об'єкти.

Найбільш загальне застосування дискримінантного аналізу є включення у дослідження багатьох змінних з метою визначення тих з них, які найкращим чином поділяють сукупності між собою. Наприклад, дослідник в галузі освіти, який цікавиться прогнозом вибору, який зроблять випускники середньої школи щодо своєї подальшої освіти, зробить з метою одержання найбільш точних прогнозів реєстрацію більшої кількості параметрів учнів.

Дискримінантний аналіз використовують, якщо дослідник хоче побудувати модель, яка дозволить краще всього передбачити, до якої сукупності належатиме той чи інший показник. У наступному міркуванні термін модель буде використовуватися для того, щоб позначати змінні, використовувані в пророкуванні приналежності до сукупності, а про невикористовуванні змінні говорять, що вони знаходяться поза межами моделі.

У ролі дискримінантного аналізу найчастіше береться лінійна функція записана у вигляді формули (3):

|  |  |
| --- | --- |
| , | (3) |

де *Х1,Х2,…,Хm*– значення ознак у даного об'єкта;

*С1,С2,…,Сm* – дискримінанті множники.

За допомогою дискримінантних множників виконуємо перехід від m-мірного простору первинних показників до одновимірного простору.

Лінійну функцію можна розглядати як проекцію даного об'єкта на деяку (одновимірну) дискримінантну вісь.

У процедурі дискримінантного аналізу дискримінантні множники визначаються таким чином, щоб забезпечити найбільшу відмінність між проекціями першої та другої вибірок на дискримінантну вісь.

Дискримінантний аналіз потрібно проводити з використанням мінімальної кількості функцій. Їхня кількість залежить від конфігурації класів в багатовимірному просторі дискримінантних змінних. Щоб визначити, скільки функцій необхідно, використовують перевірку функцій на значимість. Для оцінки значущості використовують або А-статистику Уілкса або ксі – квадрат [11].

Критерій значення Уілкса обчислюють за формулою (4):

|  |  |
| --- | --- |
| , | (4) |

де *К –* кількість значень;

*k –* число вже обчислених дискримінаційних функцій.

Чим ближче значення критерію *К*, тим краща відмінності класів, а чим ближче до 1, тим відмінність гірша.

Значення ксі-квадрат розраховують за формулою (5):

|  |  |
| --- | --- |
| , | (5) |

де *р –* кількість членів у дискримінатної функції, виключаючи вільний член функції.

Якщо це значення більше критичного із заданим рівнем значущості і числом ступенів свободи *(р-к) (К-k-1)*, то значимість підтверджується.

Канонічна дискримінантна функція для загального випадку k класів записана у формулі (6):

|  |  |
| --- | --- |
| , | (6) |

де *fki* — значення канонічної дискримінантної функції для 1-го об'єкта в *k*-му класі;

uj — шукані коефіцієнти дискримінантної функції;

*Хjki* — значення дискримінантної змінної *Хj* для i-го об'єкта в класі *k*.

Функцію будують таким чином, щоб її середні значення для різних класів якомога більше розрізнялися. При цьому сукупність функцій повинна утворювати ортогональний простір, тобто функції - незалежні один від одного. З цього випливає, що кількість функцій нe може бути більше кількості класів мінус 1 або числа дискримінантних змінних (в залежності від того, яка з цих величин менше).

Розраховані значення канонічної дискримінантної функції fki, розглядають як точки в деякому просторі. Для кожної групи можна розрахувати центр групування. Тому в цій новій системі координат для нового об'єкта розраховують відстань від нього до кожної точки групування. Зазвичай для цього використовують квадрат відстані Махаланобіса.

**Проведення розрахунків**

Всі дані було класифіковано методом дискримінантного аналізу, використовуючи покроковий відбір. На рис. 3 показаний вибір відстані та критерій розпізнавання.

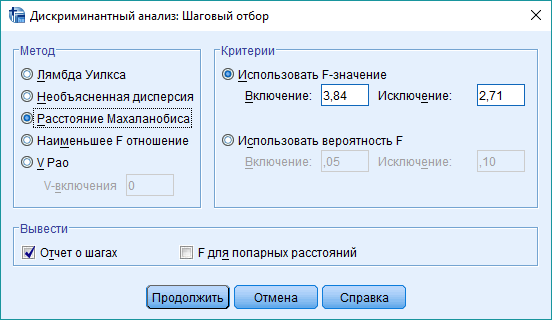


Рисунок – 3 Вибір методів

У вікні вибору класифікації, що зображено на рис 4. показаний вибір розрахунку класифікації.

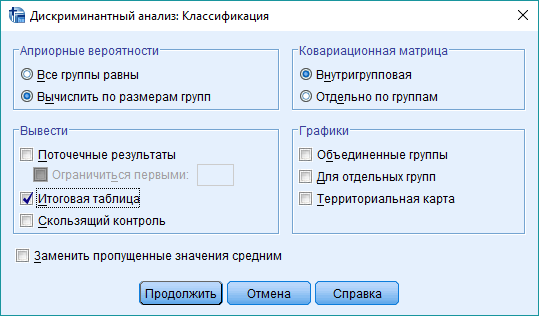
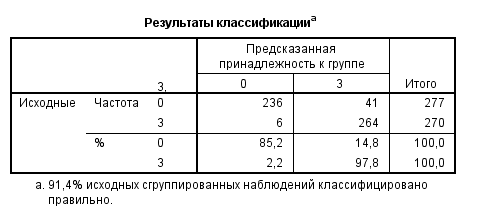


Рисунок 4 – Вибір класифікації

Результати аналізу для груп «3 проти 1,2,4,5,6,7»

Таблиця 5. Результати класифікації



За результатами дискримінантного аналізу ми бачимо, що 91,4% вихідних згрупованих спостережень класифіковано правильно.

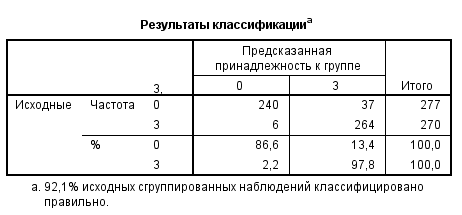
Для побудови дискримінантної функції застосовуються дані з наступної таблиці:

Таблиця 6. Коефіцієнти для дискримінантної функції



Таким чином, дискримінантна функція набуває вигляду:

Таблиця 7. Результати класифікації з нелінійними перетвореннями



З таблиці ми бачимо, що відсоток правильно класифікованих спостережень складає 92,1%

Таблиця 8. Коефіцієнти для дискримінантної функції з нелінійними перетвореннями



Таким чином, дискримінантна функція має вигляд:

Результати дослідження показують, що модель, побудована з додатковими змінними дає біль високий результат класифікації даних. Слід зазначити, що приріст є незначним і складає 0,7%, а модель при цьому стала складнішою в 2,25 рази

**Висновок**. Порівнюючи результати логістичної регресії і дискримінантного аналізу ми можемо зробити висновок, що для вихідних змінних результати класифікації методом логістичної регресії складає 90,9%, при цьому в рівняння включено 4 змінні. Класифікація методом дискримінантного аналізу складає 91,4%, кількість змінних також рівна 4. Таким чином, дискримінантний аналіз дає більший відсоток правильно спрогнозованих даних.

При проведенні логістичної регресії та дискримінантно аналізу на даних з нелінійними перетвореннями, отримаємо наступні результати: ЛР – 100%, при умові неповної моделі, при цьому до рівняння включено 13 змінних; ДА – 92,1%, при цьому функція включає 9 змінних. Беручи до уваги, що ЛР не змогла побудувати якісну модель на даних з нелінійними перетворення, то перевагу надаємо ДА. Порівняльна характеристика ДА на вихідних даних та ДА з додатковими даними нелінійного перетворення свідчить про те, що відсоток коректно спрогнозованих змінних при нелінійних перетвореннях складає <1%, та модель є більш складною. Тому для подальшого аналізу перевагу було надано дискримінатному аналізу на вихідних даних.