МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

НИЖЕГОРОДСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ

УНИВЕРСИТЕТ им. Р.Е.АЛЕКСЕЕВА



ОТЧЁТ

По лабораторной работе №2

по дисциплине

«методы Data Mining»

Руководитель:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Суркова А.С.

Студент:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Докукин Д.В.

Группа:

20-ПО

Работа защищена «\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_

С оценкой \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Цель работы**

На практике применить метод снижения размерности - Метод главных компонентов.

Ход работы

Задание 1

Применить метод главных компонентов к данным (задан dataset 3x10). определить количество главных компонентов для описания 70% данных.

|  |  |  |  |
| --- | --- | --- | --- |
| № | Масса травоядного животного, кг | Масса потребляемой пищи, кг | Продолжительности жизни, года |
| 1 | 10 | 1,6 | 45 |
| 2 | 125 | 12 | 25 |
| 3 | 65 | 5 | 9 |
| 4 | 8 | 1 | 15 |
| 5 | 46 | 3,5 | 32 |
| 6 | 5 | 0,8 | 86 |
| 7 | 79 | 7 | 50 |
| 8 | 20 | 2,3 | 12 |
| 9 | 39 | 3 | 59 |
| 10 | 98 | 8 | 34 |
| среднее | 49,5 | 4,42 | 36,7 |

|  |  |
| --- | --- |
| σ1 = | 41,39041757 |
| σ2 = | 3,611340164 |
| σ3 = | 24,00948886 |

|  |  |  |  |
| --- | --- | --- | --- |
| № | X`1 | X`2 | X`3 |
| 1 | -0,9543271684 | -0,7808735462 | 0,3456966555 |
| 2 | 1,824093702 | 2,098943787 | -0,4873073336 |
| 3 | 0,3744828129 | 0,1606051974 | -1,153710525 |
| 4 | -1,002647531 | -0,9470168539 | -0,9038093281 |
| 5 | -0,08456063518 | -0,2547530718 | -0,1957559374 |
| 6 | -1,075128076 | -1,002397956 | 2,053354833 |
| 7 | 0,7127253536 | 0,7144162231 | 0,5539476527 |
| 8 | -0,7127253536 | -0,5870396872 | -1,028759927 |
| 9 | -0,2536819055 | -0,3932058282 | 0,9287994478 |
| 10 | 1,171768802 | 0,9913217359 | -0,1124555385 |

Матрица ковариаций:

|  |  |
| --- | --- |
| cov12, cov21 = | 0,9857472175 |
| cov13, cov31 = | -0,2416740433 |
| cov23, cov32 = | -0,2341746063 |

S =

|  |  |  |
| --- | --- | --- |
| σ1^2 | cov12 | cov1,3 |
| cov21 | σ2^2 | cov23 |
| cov31 | cov32 | σ3^2 |

=

|  |  |  |
| --- | --- | --- |
| 1 | 0,98574 | -0,2416740 |
| 0,98574 | 1 | -0,234174 |
| -0,2416740 | -0,2341746 | 1 |

Собственные числа:

λ1 = 0,014

λ2 = 0,894

λ3 = 2,090

Характеристические уравнения и собственные вектора

λ1 = 0,014

Собственный вектор [-2,292; -2,288; 1]

q = 0,29

w [-0,66468;-0,66352;0,29]

λ2 = 0,894

Собственный вектор [-0,214; -0,223; 1]

q = 0,95

w [-0,2033;-0,21185;0,95]

λ3 = 2,090

Собственный вектор [123,906; 123,659; 1]

q = 0,00571

w [-0,707;-0,706;0,00571]

Линейные комбинации

Анализ

|  |  |  |  |
| --- | --- | --- | --- |
| Главная компонента y | Y1 | Y2 | Y3 |
| Собственное число | 0,014 | 0,894 | 2,090 |
| Вклад j-й главной компоненты | 0,47 | 29,8 | 69,73 |
| Суммарный вклад | 0,47 | 30,27 | 100 |

Задание 2

Проверить правильность задания 1 в программной среде Python/R. Построить scatter-plot (попарно или в 3d), гистограмму, box-plot и qq-plot каждого признака. Сделать вывод о признаках. Есть ли взаимосвязь между признаками? Целесообразно ли использовать метод главных компонентов к этим данным?

