|  |  |  |
| --- | --- | --- |
| Министерство науки и высшего образования  Российской Федерации | | |
| Федеральное государственное бюджетное  образовательное учреждение высшего образования | | |
| «Новосибирский государственный технический университет» | | |
|  | | |
| Кафедра теоретической и прикладной информатики | | |
|  | | |
| Лабораторная работа № 1 | | |
| по дисциплине «Статистический анализ нечисловых данных» | | |
|  | | |
|  | | |
|  | | |
|  | Факультет: | ПМИ |
| Группа: | ПМИ-02 |
| Вариант: | 22 |
| Студент: | Сидоров Даниил, |
|  | Дюков Богдан |
| Преподаватель: | Тимофеева Анастасия Юрьевна. |
|  |  |
|
|  |  |
| Новосибирск | | |
| 2023 | | |

**Задание 0**

Из набора данных Вашего варианта сформировать два массива данных:

* массив количественных данных (все количественные данные оставить без изменений, для всех качественных данных кроме переменной класса произвести калибровку с учетом априорного шанса с поправкой Лапласа, значения переменной класса задать как 1, если положительный класс, 0 иначе);
* массив качественных данных (для всех качественных данных кроме переменной класса произвести калибровку с учетом априорного шанса с поправкой Лапласа, значения переменной класса задать как 1, если положительный класс, 0 иначе; для всех количественных данных произвести дискретизацию с равной частотой, в качестве границ интервалов взять выборочные квантили порядка 0, 0.2, 0.4, 0.6, 0.8, 1).

**Решение**

1. Произведем калибровку данных из столбцов A1, A3, A6, A7, A10, A17, A20 с учетом априорного шанса с поправкой Лапласа. Также зададим значения переменной класса как 1, если положительный класс, 0 иначе. В результате имеем следующую таблицу:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **A1** | **A3** | **A5** | **A6** | **A7** | **A10** | **A11** | **A16** | **A17** | **A20** | **A21** |
| 0,4228 | 0,6707 | 1795 | 0,4241 | 0,5109 | 0,5451 | 4 | 2 | 0,4571 | 0,4941 | 1 |
| 0,4228 | 0,4813 | 1092 | 0,4241 | 0,4829 | 0,5451 | 4 | 2 | 0,5242 | 0,4941 | 1 |
| 0,757 | 0,4813 | 4811 | 0,6832 | 0,597 | 0,5073 | 4 | 1 | 0,4571 | 0,4941 | 1 |
| 0,4228 | 0,423 | 4712 | 0,6832 | 0,4829 | 0,5073 | 2 | 2 | 0,4479 | 0,4941 | 1 |
| 0,757 | 0,4813 | 874 | 0,6832 | 0,4137 | 0,5073 | 1 | 1 | 0,5242 | 0,4941 | 1 |
| 0,757 | 0,4813 | 4611 | 0,4241 | 0,4137 | 0,5073 | 4 | 1 | 0,5242 | 0,4941 | 0 |
| 0,2881 | 0,423 | 6887 | 0,4241 | 0,4829 | 0,5073 | 3 | 1 | 0,5242 | 0,4941 | 0 |
| 0,4228 | 0,6707 | 4576 | 0,5741 | 0,4894 | 0,5073 | 4 | 1 | 0,5242 | 0,4941 | 1 |
| 0,757 | 0,4813 | 1238 | 0,6832 | 0,4894 | 0,5073 | 4 | 1 | 0,4479 | 0,4941 | 1 |
| 0,2881 | 0,6707 | 1382 | 0,5741 | 0,597 | 0,5073 | 1 | 2 | 0,5242 | 0,4941 | 1 |
| 0,4228 | 0,4813 | 959 | 0,4241 | 0,4829 | 0,5073 | 2 | 1 | 0,5242 | 0,6265 | 0 |
| 0,757 | 0,4813 | 2507 | 0,5898 | 0,5109 | 0,5073 | 4 | 1 | 0,4571 | 0,4941 | 1 |
| 0,4228 | 0,423 | 1209 | 0,4241 | 0,4894 | 0,5073 | 4 | 1 | 0,4479 | 0,4941 | 0 |
| 0,757 | 0,4813 | 1568 | 0,5741 | 0,4829 | 0,5073 | 4 | 1 | 0,4571 | 0,4941 | 1 |
| 0,757 | 0,4813 | 1505 | 0,4241 | 0,4829 | 0,5073 | 2 | 1 | 0,4479 | 0,4941 | 1 |
| 0,2881 | 0,4813 | 1275 | 0,6832 | 0,4829 | 0,5073 | 2 | 1 | 0,5242 | 0,4941 | 0 |
| 0,2881 | 0,4813 | 1282 | 0,5741 | 0,4829 | 0,5073 | 2 | 1 | 0,4571 | 0,4941 | 0 |
| 0,757 | 0,6707 | 1520 | 0,6832 | 0,5109 | 0,5073 | 4 | 1 | 0,5242 | 0,4941 | 1 |
| 0,757 | 0,4813 | 1736 | 0,4241 | 0,597 | 0,5073 | 4 | 1 | 0,4571 | 0,4941 | 1 |

1. Произведем дискретизацию данных из столбцов A5, A11, A16.

* Интервалы для A5: [276, 1245.8]; (1245.8, 1831.4]; (1831.4, 2750]; (2750, 4316.6]; (4316.6, 1824].
* Интервалы для A11: [1, 2]; (2, 4].
* Интервалы для A16: [1, 2]; (2, 4].

В результате имеем следующую таблицу:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **A1** | **A3** | **A5** | **A6** | **A7** | **A10** | **A11** | **A16** | **A17** | **A20** | **A21** |
| 0,4228 | 0,6707 | 0,4 | 0,4241 | 0,5109 | 0,5451 | 1 | 0,2 | 0,4571 | 0,4941 | 1 |
| 0,4228 | 0,4813 | 0,2 | 0,4241 | 0,4829 | 0,5451 | 1 | 0,2 | 0,5242 | 0,4941 | 1 |
| 0,757 | 0,4813 | 1 | 0,6832 | 0,597 | 0,5073 | 1 | 0,2 | 0,4571 | 0,4941 | 1 |
| 0,4228 | 0,423 | 1 | 0,6832 | 0,4829 | 0,5073 | 0,2 | 0,2 | 0,4479 | 0,4941 | 1 |
| 0,757 | 0,4813 | 0,2 | 0,6832 | 0,4137 | 0,5073 | 0,2 | 0,2 | 0,5242 | 0,4941 | 1 |
| 0,757 | 0,4813 | 1 | 0,4241 | 0,4137 | 0,5073 | 1 | 0,2 | 0,5242 | 0,4941 | 0 |
| 0,2881 | 0,423 | 1 | 0,4241 | 0,4829 | 0,5073 | 1 | 0,2 | 0,5242 | 0,4941 | 0 |
| 0,4228 | 0,6707 | 1 | 0,5741 | 0,4894 | 0,5073 | 1 | 0,2 | 0,5242 | 0,4941 | 1 |
| 0,757 | 0,4813 | 0,2 | 0,6832 | 0,4894 | 0,5073 | 1 | 0,2 | 0,4479 | 0,4941 | 1 |
| 0,2881 | 0,6707 | 0,4 | 0,5741 | 0,597 | 0,5073 | 0,2 | 0,2 | 0,5242 | 0,4941 | 1 |
| 0,4228 | 0,4813 | 0,2 | 0,4241 | 0,4829 | 0,5073 | 0,2 | 0,2 | 0,5242 | 0,6265 | 0 |
| 0,757 | 0,4813 | 0,6 | 0,5898 | 0,5109 | 0,5073 | 1 | 0,2 | 0,4571 | 0,4941 | 1 |
| 0,4228 | 0,423 | 0,2 | 0,4241 | 0,4894 | 0,5073 | 1 | 0,2 | 0,4479 | 0,4941 | 0 |
| 0,757 | 0,4813 | 0,4 | 0,5741 | 0,4829 | 0,5073 | 1 | 0,2 | 0,4571 | 0,4941 | 1 |
| 0,757 | 0,4813 | 0,4 | 0,4241 | 0,4829 | 0,5073 | 0,2 | 0,2 | 0,4479 | 0,4941 | 1 |
| 0,2881 | 0,4813 | 0,4 | 0,6832 | 0,4829 | 0,5073 | 0,2 | 0,2 | 0,5242 | 0,4941 | 0 |
| 0,2881 | 0,4813 | 0,4 | 0,5741 | 0,4829 | 0,5073 | 0,2 | 0,2 | 0,4571 | 0,4941 | 0 |
| 0,757 | 0,6707 | 0,4 | 0,6832 | 0,5109 | 0,5073 | 1 | 0,2 | 0,5242 | 0,4941 | 1 |
| 0,757 | 0,4813 | 0,4 | 0,4241 | 0,597 | 0,5073 | 1 | 0,2 | 0,4571 | 0,4941 | 1 |

**Задание 1**

Рассчитайте все парные показатели взаимосвязи между переменными из набора данных, соответствующего Вашему варианту. Используемый показатель взаимосвязи выбирается в соответствии с вариантом следующим образом: номер варианта нужно разделить на 5.

**Решение**

Показатель взаимосвязи при m=2 – коэффициент корреляции Пирсона. Сначала вычислялись ковариации и стандартные отклонения для каждой пары переменных, после чего находился искомый коэффициент:

В результате имеем следующую таблицу парных показателей:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **A1** | **A3** | **A5** | **A6** | **A7** | **A10** | **A11** | **A16** | **A17** | **A20** | **A21** |
| A1 |  | 0,1476 | -0,0499 | 0,3034 | 0,0426 | 0,0361 | -0,0209 | 0,0871 | 0,063 | 0,0108 | 0,3662 |
| A3 |  |  | -0,0898 | -0,0019 | 0,0762 | -0,0492 | 0,0701 | 0,3658 | 0,109 | -0,0032 | 0,2581 |
| A5 |  |  |  | 0,0554 | 0,0841 | -0,1476 | 0,0016 | -0,0323 | -0,1313 | -0,0181 | -0,1443 |
| A6 |  |  |  |  | 0,0713 | 0,0245 | 0,0717 | -0,0344 | 0,0375 | 0,0173 | 0,1948 |
| A7 |  |  |  |  |  | 0,0214 | 0,0927 | 0,1522 | 0,1176 | 0,0405 | 0,1055 |
| A10 |  |  |  |  |  |  | -0,0367 | -0,0638 | 0,0162 | 0,0743 | 0,1041 |
| A11 |  |  |  |  |  |  |  | 0,1324 | 0,0219 | -0,0634 | -0,0224 |
| A16 |  |  |  |  |  |  |  |  | 0,018 | 0,0164 | 0,0519 |
| A17 |  |  |  |  |  |  |  |  |  | -0,0027 | 0,0625 |
| A20 |  |  |  |  |  |  |  |  |  |  | 0,053 |

**Задание 2**

Проверьте гипотезы о значимости взаимосвязей между переменными.

m=2 – о значимости коэффициента корреляции Пирсона с помощью стандартной процедуры и с помощью перестановочного критерия.

**Решение**

Проверим гипотезу о значимости коэффициента корреляции Пирсона с помощью стандартной процедуры. Для каждой пары вычислим t-статистику:

Если значение t-статистики превышает критическое значение (квантиль распределения Стьюдента с 498 степенями свободы уровня 0.975 (=0.05)), то делается вывод о статистической значимости выявленной корреляционной связи.

Также выполним точную оценку силы корреляционной связи. Если абсолютное значение коэффициента корреляционной связи менее 0.3, то назовем такую силу ‘Слабой', а если абсолютное значение лежим в промежутке от 0.3 до 0.5 – ‘Умеренной’ и т.д.

В результате имеем следующую таблицу:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **A1** | **A3** | **A5** | **A6** | **A7** | **A10** | **A11** | **A16** | **A17** | **A20** | **A21** |
| A1 |  | слабая, значима | слабая, незначима | умеренная, значима | слабая, незначима | слабая, незначима | слабая, незначима | слабая, незначима | слабая, незначима | слабая, незначима | умеренная, значима |
| A3 |  |  | слабая, значима | слабая, незначима | слабая, незначима | слабая, незначима | слабая, незначима | умеренная, значима | слабая, значима | слабая, незначима | слабая, значима |
| A5 |  |  |  | слабая, незначима | слабая, незначима | слабая, значима | слабая, незначима | слабая, незначима | слабая, значима | слабая, незначима | слабая, значима |
| A6 |  |  |  |  | слабая, незначима | слабая, незначима | слабая, незначима | слабая, незначима | слабая, незначима | слабая, незначима | слабая, значима |
| A7 |  |  |  |  |  | слабая, незначима | слабая, значима | слабая, значима | слабая, значима | слабая, незначима | слабая, значима |
| A10 |  |  |  |  |  |  | слабая, незначима | слабая, незначима | слабая, незначима | слабая, незначима | слабая, значима |
| A11 |  |  |  |  |  |  |  | слабая, значима | слабая, незначима | слабая, незначима | слабая, незначима |
| A16 |  |  |  |  |  |  |  |  | слабая, незначима | слабая, незначима | слабая, незначима |
| A17 |  |  |  |  |  |  |  |  |  | слабая, незначима | слабая, незначима |
| A20 |  |  |  |  |  |  |  |  |  |  | слабая, незначима |

Проверим гипотезу о значимости коэффициента корреляции Пирсона с помощью перестановочного критерия.

Для этого мы в каждой паре переменных вычисляем коэффициенты корреляции Пирсона (первая переменная фиксируется, а во второй делается 200 перестановок, с каждой из которых вычисляется коэффициент Пирсона). Полученные коэффициенты корреляции сортируются и из них отбирается квантиль уровня 0.95, соответствующий уровню значимости 0.05. Отобранный коэффициент будет являться искомым в паре. Он будет сравниваться с критическим значением (с соответствующим коэффициентом корреляции, вычисленным ранее).

Таблица полученных коэффициентов корреляции с помощью перестановочного критерия:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **A1** | **A3** | **A5** | **A6** | **A7** | **A10** | **A11** | **A16** | **A17** | **A20** | **A21** |
| A1 |  | 0,075 | 0,0675 | 0,0687 | 0,0694 | 0,0837 | 0,0613 | 0,0638 | 0,0649 | 0,0722 | 0,0809 |
| A3 | 0,0693 |  | 0,0804 | 0,0685 | 0,0752 | 0,0651 | 0,0677 | 0,0721 | 0,0806 | 0,0731 | 0,0639 |
| A5 | 0,0739 | 0,0836 |  | 0,0816 | 0,0862 | 0,068 | 0,0622 | 0,0911 | 0,079 | 0,0838 | 0,0691 |
| A6 | 0,0805 | 0,068 | 0,0763 |  | 0,0779 | 0,0813 | 0,0806 | 0,0941 | 0,0744 | 0,0599 | 0,0645 |
| A7 | 0,0727 | 0,0852 | 0,0766 | 0,0739 |  | 0,0653 | 0,0655 | 0,0799 | 0,0613 | 0,0719 | 0,0755 |
| A10 | 0,0883 | 0,0596 | 0,0684 | 0,0814 | 0,0658 |  | 0,0652 | 0,0591 | 0,0733 | 0,0465 | 0,084 |
| A11 | 0,0715 | 0,0665 | 0,0692 | 0,0759 | 0,0736 | 0,0872 |  | 0,0724 | 0,0686 | 0,0787 | 0,0845 |
| A16 | 0,0727 | 0,0725 | 0,0752 | 0,075 | 0,0738 | 0,0585 | 0,0724 |  | 0,0644 | 0,0738 | 0,0663 |
| A17 | 0,0703 | 0,071 | 0,0883 | 0,0658 | 0,0672 | 0,0737 | 0,0717 | 0,0633 |  | 0,0678 | 0,0727 |
| A20 | 0,0728 | 0,0612 | 0,0797 | 0,0793 | 0,0767 | 0,0465 | 0,0787 | 0,0929 | 0,0678 |  | 0,0542 |
| A21 | 0,0764 | 0,0729 | 0,0645 | 0,0677 | 0,0685 | 0,077 | 0,0734 | 0,0738 | 0,0703 | 0,0773 |  |

Результаты сравнения данных коэффициентов с коэффициентами корреляции из предыдущего пункта:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **A1** | **A3** | **A5** | **A6** | **A7** | **A10** | **A11** | **A16** | **A17** | **A20** | **A21** |
| A1 |  | значима | незначима | значима | незначима | незначима | незначима | значима | незначима | незначима | значима |
| A3 | значима |  | значима | незначима | значима | незначима | значима | значима | значима | незначима | значима |
| A5 | незначима | значима |  | незначима | незначима | значима | незначима | незначима | значима | незначима | значима |
| A6 | значима | незначима | незначима |  | незначима | незначима | незначима | незначима | незначима | незначима | значима |
| A7 | незначима | незначима | значима | незначима |  | незначима | значима | значима | значима | незначима | значима |
| A10 | незначима | незначима | значима | незначима | незначима |  | незначима | значима | незначима | значима | значима |
| A11 | незначима | значима | незначима | незначима | значима | незначима |  | значима | незначима | незначима | незначима |
| A16 | значима | значима | незначима | незначима | значима | значима | значима |  | незначима | незначима | незначима |
| A17 | незначима | значима | значима | незначима | значима | незначима | незначима | незначима |  | незначима | незначима |
| A20 | незначима | незначима | незначима | незначима | незначима | значима | незначима | незначима | незначима |  | незначима |
| A21 | значима | значима | значима | значима | значима | значима | незначима | незначима | незначима | незначима |  |

**Задание 3**

Дайте интерпретацию полученным результатам исходя из практических

соображений. Уделите особое внимание интерпретации взаимосвязей между

объясняющими переменными и переменной класса.

По результатам выполнения лабораторной работы имеем следующие выводы:

1. Если смотреть на полученные коэффициенты корреляционной связи, то можно увидеть, что теснота корреляционной связи между переменными преимущественно слабая. Это говорит нам о том, что изменения в одних переменных слабо связаны с изменениями в других переменных.
2. По результатам стандартной процедуры можно увидеть, что корреляционная связь между A20 и остальными переменными статистически незначима. Связь между A7 и остальными переменными (за исключением A21) также статистически незначима.
3. Результаты перестановочного критерия практически полностью совпадает с результатами стандартной процедуры.