|  |  |  |
| --- | --- | --- |
|  | | |
| 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» | | |
| Место для ввода текста. | | |
|  | | |
|  | Бригада 11 | Мещанинов Григорий |
| Группа ПМ-14 | Хамитова екатерина |
| Вариант 11 |  |
|  |  |
|  |  |
|  |  |
| Преподаватель | Тимофеева анастасия юрьевна |
|  |  |
| Новосибирск, 2024 | | |

**Задание 1.**

Из набора данных Вашего варианта сформируйте два массива данных:

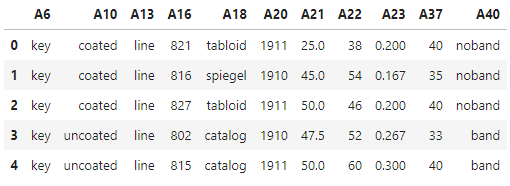
- массив количественных данных (все количественные данные оставить без изменений, для всех качественных данных кроме переменной класса1 произвести калибровку с учетом априорного шанса с поправкой Лапласа, значения переменной класса задать как 1, если положительный класс, 0 иначе);

- 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).

Данные:

|  |  |  |  |
| --- | --- | --- | --- |
| Столбец | Наименование | Тип | Значения |
| A6 | ink color | nominal | key  type |
| A10 | paper type | nominal | uncoated  coated  super |
| A13 | solvent type | nominal | xylol  lactol  naptha  line  other |
| A16 | press | nominal | 821  802  813  824  815  816  827  828 |
| A18 | cylinder size | nominal | catalog  spiegel  tabloid |
| A20 | plating tank | nominal | 1910  1911  other |
| A21 | proof cut | numeric | 0-100 |
| A22 | viscosity | numeric | 0-100 |
| A23 | caliper | numeric | 0-1.0 |
| A37 | current density | numeric | 20-50 |
| A40 | band type | nominal | band  noband |

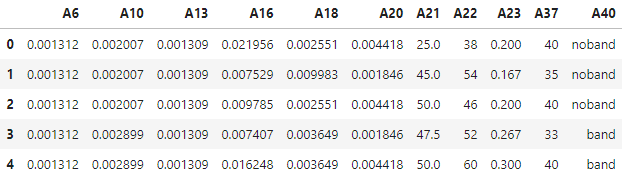
Данные до калибровки и дискретизации:



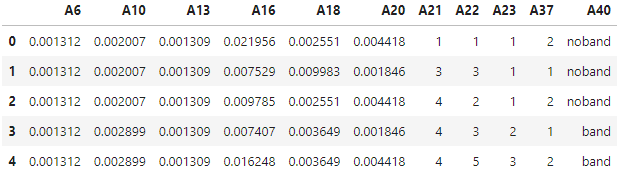
Откалиброванные с учетом априорного шанса с поправкой Лапласа значения:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Столбец/Переменная | c | n+ | n- | Поправка |
| A6, key | 500 | 198 | 302 | 0,001312 |
| A6, type | 0 | 0 | 0 | 1 |
| A10, coated | 194 | 54 | 140 | 0,002007 |
| A10, uncoated | 306 | 144 | 162 | 0,002899 |
| A10, super | 0 | 0 | 0 | 1 |
| A13, line | 483 | 187 | 296 | 0,001309 |
| A13, xylol | 14 | 11 | 3 | 0,176471 |
| A13, lactol | 0 | 0 | 0 | 1 |
| A13, naptha | 3 | 0 | 3 | 0,076923 |
| A13, other | 0 | 0 | 0 | 1 |
| A16, 821 | 70 | 43 | 27 | 0,021956 |
| A16, 802 | 67 | 22 | 45 | 0,007407 |
| A16, 813 | 41 | 7 | 34 | 0,005544 |
| A16, 824 | 54 | 7 | 47 | 0,003077 |
| A16, 815 | 90 | 54 | 36 | 0,016248 |
| A16, 816 | 100 | 43 | 57 | 0,007529 |
| A16, 827 | 46 | 14 | 32 | 0,009785 |
| A16, 828 | 32 | 8 | 24 | 0,011125 |
| A18, tabloid | 279 | 116 | 163 | 0,002551 |
| A18, spiegel | 51 | 17 | 34 | 0,009983 |
| A18, catalog | 170 | 65 | 105 | 0,003649 |
| A20, 1911 | 173 | 75 | 98 | 0,004418 |
| A20, 1910 | 327 | 123 | 204 | 0,001846 |
| A20, other | 0 | 0 | 0 | 1 |

Датафрейм после калибровки:



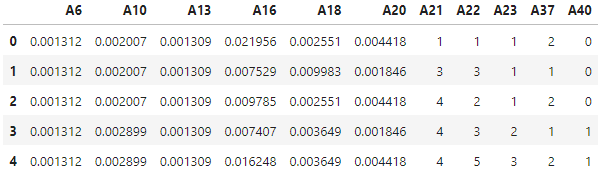
Датафрейм после дискретизации:



Промежутки:

|  |  |
| --- | --- |
| А21 | [25; 37.5]  (37.5; 40]  (40; 47]  (47; 50]  (50; 72.5] |
| А22 | [35; 43]  (43; 48]  (48; 54]  (54; 58]  (58; 72] |
| А23 | [0.133; 0.2]  (0.2; 0.267]  (0.267; 0.3]  (0.3; 0.333]  (0.333; 0.533] |
| А37 | [30; 35]  (35; 40]  (40; 45] |

Классовый признак:



**Задание 2.**

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, полученный остаток m от деления соответствует:

m=1 – показатель взаимной информации, коэффициент неопределенности, симметричный коэффициент.

Формула показателя взаимной информации (через вероятности):

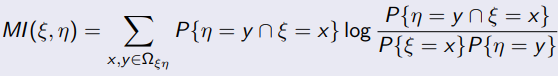


Таблица с показателями взаимной информации:



Формула коэффициента неопределенности:





Таблица с коэф-ми неопределенности:



Формула коэффициента симметричности:



Таблица с коэф-ми симметричности:



**Задание 3.**

Проверьте гипотезы о значимости взаимосвязей между переменными:

m=1 – для показателя взаимной информации с помощью перестановочного критерия.

Сделайте выводы.

Датафрейм 0,95-квантилей отсортированных массивов с показателями MI, полученный перестановочным критерием:



Если новое значение показателя MI меньше старого, то гипотеза отвергается (1), иначе не отвергается (0)



Проведем анализ результатов:

1. Гипотеза H0: между двумя признаками нет взаимосвязи (их зависимость случайна).

Гипотеза H1: между двумя признаками существует взаимосвязь (их зависимость не случайна).

1. Если MI для исходных данных больше, чем 0,95-квантиль для перестановок (1), то это означает, что наблюдаемая взаимосвязь между признаками значительно сильнее, чем можно было бы ожидать при случайном распределении данных. В этом случае отвергается H0 и делается вывод о наличии значимой связи между признаками.

Если MI для исходных данных меньше, чем 0,95 квантиль для перестановок (0), то это означает, что наблюдаемая взаимосвязь между признаками не отличается от случайной, и гипотеза H0 не отвергается.

Вывод: в наших данных не так много случайных взаимосвязей и их не так много в столбце с классовым признаком. В следующем задании будем брать только те признаки, на пересечении которых в датафрейме с 0 и 1 стоит 1 (то есть там, где связь есть, и она не случайная).

**Задание 4.**

Из подмножества признаков, значимо влияющих на переменную класса, выберите пару признаков, наиболее сильно взаимосвязанных друг с другом. Обозначим их X1 и X2. Проверьте наличие частной корреляции между признаком X1 (X2) и откликом при исключении влияния X2 (X1) с помощью:

m=0, 1, 4 – показателя частной взаимной информации;

Сделайте вывод о наличии или отсутствии ложной корреляции.

Взаимная информация и частная взаимная информация для качественных данных (Х1 – это признак А22, Х2 – это признак А37, Y – это А40):

MI(X1, Y) = MI(A22, A40) = 0.017370105214595363

I(X1, Y | X2) = I(A22, A40 | A37) = 0.039940001390988786

I > MI

MI(X2, Y) = MI(A37, A40) = 0.027744142542336366

I(X2, Y | X1) = I(A37, A40 | A22) = 0.050314038718729664

I > MI

Частная взаимная информация показала то, что зависимость X1 и Y больше, когда влияние X2 не учтено. Частная взаимная информация между X2 и Y при исключении X1 тоже больше, чем просто взаимная информация.

Вывод: X1 и X2 зависимы. То, что I(X1, Y | X2) > MI(X1, Y), говорит о том, что зависимость X1 и Y усиливается, когда X2 не учитывается. Аналогично, I(X2, Y | X1) > MI(X2, Y) показывает, что X2 и Y больше связаны, когда X1 не учитывается.

**Задание 5.**

Дайте интерпретацию полученным результатам исходя из практических соображений. Уделите особое внимание интерпретации взаимосвязей между объясняющими переменными и переменной класса.

Коэффициент неопределенности измеряет, насколько один признак уменьшает неопределенность в отношении другого. Коэффициент симметричности показывает, насколько сильно два признака связаны друг с другом. Будем ранжировать значимость признаков по второму коэф-ту.

От наименее к наиболее значимым для классового признака:

A6 (в наборе данных из двух значений встречалось только одно)

А18

А20

А23

А22

А21

А10

А13

А37

А16

Признак А6 можно убрать.

