

Тема 7. Исследование взаимосвязей признаков

Практические задания для самостоятельного выполнения

Задания выполняются по вариантам. Формулировки заданий общие для всех вариантов; конкретные условия, указанные в общей формулировке, выбираются в соответствии с номером своего варианта.

Результаты выполнения заданий необходимо представить в виде двух файлов:

- 1) ноутбук в формате *ipynb*, содержащий программный код, результаты его выполнения, а также все необходимые пояснения, выводы и комментарии (в текстовых ячейках);
- 2) файл в формате *pdf*, полученный путем экспорта (или вывода на печать) ноутбука из п. 1).

Внимание: в названии файлов должна обязательно присутствовать фамилия автора. Безымянные работы проверяться не будут.

Исследование взаимосвязей количественных признаков

Задача 1 (максимум 0,5 балла).

Выполнить исследование имеющихся данных наблюдений на предмет наличия/отсутствия взаимосвязи между количественными признаками X и Y (см. таблицу 1). Данные наблюдений представлены в csv-файле (по вариантам). Имя файла: Вариант N.1, где N – номер варианта.

В ходе решения задачи

- импортировать данные наблюдений из файла;
- построить диаграмму рассеяния значений признаков;
- выполнить визуальную оценку диаграммы рассеяния и сформулировать (записать в текстовой ячейке) предположение о наличии или отсутствии взаимосвязи между признаками; в случае наличия – предположение о силе и направленности связи; пояснить, на чем основаны сделанные предположения;
- построить гистограммы распределения признаков и обосновать возможный выбор метода корреляционного анализа (свои соображения записать в текстовой ячейке);
- найти выборочный коэффициент корреляции Пирсона и выполнить оценку его значимости (уровень значимости положить равным $\alpha = 0,05$);
- дать интерпретацию полученным числовым значениям (записать в текстовой ячейке);
- найти ранговые коэффициенты корреляции Спирмена и Кендалла и выполнить оценку их значимости (уровень значимости положить равным $\alpha = 0,05$);
- сопоставить все полученные результаты, дать свои комментарии;
- сделать выводы (записать в текстовой ячейке) о наличии/отсутствии

взаимосвязи между анализируемыми признаками; в случае наличия – о направленности и силе связи.

Таблица 1. Описание признаков X и Y .

Вариант	Данные
1	X – стоимость основных производственных фондов (млн. руб.), Y – объем валовой продукции (по однотипным предприятиям)
2	X – стоимость основных средств предприятий (млн. руб.), Y – месячный выпуск продукции (по однотипным предприятиям)
3	X – фазовая проницаемость воды, Y – нефтенасыщенность породы
4	X – производственные средства завода (млн. руб.), Y – суточная выработка завода (по однотипным предприятиям)
5	X – величина товарооборота магазина (млн. руб.), Y – торговая площадь (m^2)
6	X – осевая статическая нагрузка на забой (тс), Y – удельный момент на долоте (кгс·м/тс) при бурении пород
7	X – пробег автомобиля (тыс. км), Y – стоимость ежемесячного технического обслуживания
8	X – энерговооруженность труда (кВт/час), Y – стоимость готовой продукции (тыс. руб.)
9	X – механическая скорость проходки (м/с), Y – количество израсходованных долот (шт.) при бурении скважин
10	X – нагрузка на долото (атм.), Y – скорость бурения (м/час) в твердых породах
11	X – выработка продукции (млн. руб.), Y – затраты топлива (по однотипным предприятиям)
12	X – производственные средства завода (млн. руб.), Y – суточная выработка завода (по однотипным предприятиям)
13	X – средний возраст техники, Y – коэффициент сменности техники по предприятию ПМК-7 объединения «Сибкомплектмонтаж»
14	X – реализация продукции (млн. руб.), Y – накладные расходы на реализацию (тыс. руб.)
15	X – содержание окиси железа (гр), Y – содержание закиси железа (гр) в пробах руды
16	X – вес детей школьного возраста (фунты), Y – рост детей школьного возраста (см)

Вариант	Данные
17	X – стаж (в годах) работы на Тюменском моторостроительном объединении в цехе резиново-технических и пластмассовых изделий на слесарном участке, Y – время на обработку одной детали (мин.)
18	X – веса детали (кг), Y – время, затрачиваемое на закрепление детали (с)
19	X – износ резца (мм), Y – диаметр вала (мм) при чистовом точении
20	X – среднегодовая стоимость основных производственных фондов (млн. руб.), Y – стоимость товарной продукции (млн. руб.)
21	X – вес детей школьного возраста (фунты), Y – рост детей школьного возраста (см)
22	X – содержание окиси железа (гр), Y – содержание закиси железа (гр) в пробах руды
23	X – реализация продукции (млн. руб.), Y – накладные расходы на реализацию (тыс. руб.)
24	X – величина товарооборота магазина (млн. руб.), Y – торговая площадь (m^2)
25	X – пробег автомобиля (тыс. км), Y – стоимость ежемесячного технического обслуживания
26	X – механическая скорость проходки (м/с), Y – количество израсходованных долот (шт.) при бурении скважин
27	X – выработка продукции (млн. руб.), Y – затраты топлива (по однотипным предприятиям)
28	X – среднегодовая стоимость основных производственных фондов (млн. руб.), Y – стоимость товарной продукции (млн. руб.)
29	X – производственные средства завода (млн. руб.), Y – суточная выработка завода (по однотипным предприятиям)
30	X – вес детей школьного возраста (фунты), Y – рост детей школьного возраста (см)

Задача 2 (максимум 0,5 балла).

Выполнить исследование имеющихся данных наблюдений на предмет наличия/отсутствия взаимосвязи между количественными признаками X и Y . Данные наблюдений представлены в csv-файле (по вариантам). Имя файла: Вариант N.2, где N – номер варианта.

В ходе решения задачи

- импортировать данные наблюдений из файла;
- построить диаграмму рассеяния значений признаков;
- выполнить визуальную оценку диаграммы рассеяния и сформулировать (записать в текстовой ячейке) предположение о наличии или отсутствии

взаимосвязи между признаками; в случае наличия – предположение о силе и направленности связи; пояснить, на чем основаны сделанные предположения;

- построить гистограммы распределения признаков и обосновать возможный выбор метода корреляционного анализа (свои соображения записать в текстовой ячейке);
- найти ранговые коэффициенты корреляции Спирмена и Кендалла и выполнить оценку их значимости (уровень значимости положить равным $\alpha = 0,05$);
- дать интерпретацию полученным числовым значениям (записать в текстовой ячейке);
- сделать выводы (записать в текстовой ячейке) о наличии/отсутствии взаимосвязи между анализируемыми признаками; в случае наличия – о направленности и силе связи.

Исследование взаимосвязей категориальных признаков

Задача 3 (максимум 0,5 балла).

В ходе эксперимента испытуемым был предложен тест, в котором первый вопрос был направлен на изучение признака X , второй – на изучение признака Y . Выполнить исследование на наличие/отсутствие взаимосвязи между признаками X и Y на основании полученных экспериментальных данных (ответов испытуемых на вопросы теста). Данные для анализа представлены в csv-файле (по вариантам). Имя файла: Вариант N.3, где N – номер варианта.

В ходе решения задачи

- импортировать данные наблюдений из файла;
- построить таблицу сопряженности признаков;
- проверить правомерность применения критерия «хи-квадрат» к данной выборке; полученный вывод (с обоснованием) записать в текстовой ячейке;
- в случае правомерности применения критерия выполнить соответствующий расчет и получить вывод о наличии или отсутствии связи между признаками (уровень значимости положить равным $\alpha = 0,05$);
- дать интерпретацию всем полученным числовым значениям (записать в текстовой ячейке);
- в случае наличия связи между признаками оценить силу связи с помощью коэффициента Крамера;
- выводы о наличии/отсутствии, а также силе связи (при ее наличии) записать в текстовой ячейке (с объяснением, как получены эти выводы).

Исследование набора данных клиентов банка: дополнительные задания для тех, кто успешно справился с задачами 1 – 3.

Решение задач 4 и 5 будет оцениваться только у тех студентов, которые

представили полное решение задач 1 – 3 со всеми указанными пояснениями и обоснованиями.

В условиях задач 4 и 5 нет точного описания всех шагов решения. Автору решения предлагается самостоятельно продумать все шаги, которые, по его мнению, должны привести его к правильным выводам. Оценка выполнения этих заданий будет включать в том числе и оценку логики выполняемых действий, с учетом представленных автором обоснований того или иного действия.

Задача 4 (максимум 0,5 балла).

Используя набор данных клиентов банка (файл Данные_клиентов.txt), проанализировать, существует ли взаимосвязь между возрастом клиента и продолжительностью его разговора с сотрудником банка.

Все расчеты и выводы, полученные на основании этих расчетов, пояснить. Подробные пояснения записать в текстовых ячейках.

Наличие иллюстраций (графиков, диаграмм) вместе с наблюдениями и предположениями/выводами, которые могут быть сделаны по этим иллюстрациям, очень приветствуется.

Задача 5 (максимум 1 балл).

Как было отмечено в учебном ноутбуке, наборы данных, подобные представленному в файле Данные_клиентов.txt, используются для построения предсказательных моделей Data Mining, в частности, для прогнозирования отклика клиентов на предложение банка. В задаче прогнозирования отклика одноименный признак (столбец “Отклик”) является выходным (целевым, прогнозируемым), а остальные признаки могут рассматриваться в качестве входных (на основе которых будущая модель будет строить свои предсказания).

Однако, среди потенциальных входных признаков не все могут быть одинаково важными (значимыми) для построения прогноза. Некоторые признаки могут вообще быть «шумовыми». Такие признаки, фактически не влияют на прогнозируемую величину, а только создают «шум», который может заметно затруднять обучение модели. В связи с этим имеет смысл анализировать значимость входных признаков, и принимать решение о включении того или иного признака в предсказательную модель с учетом результатов этого анализа.

В данном задании предлагается проанализировать влияние имеющихся в наборе категориальных признаков на целевой признак.

Замечание 1. Строго говоря, в данном случае целевой признак является бинарным («да»/«нет» может кодироваться как 1/0). Для исследования связей бинарных признаков с другими признаками существуют более специализированные методы (за пределами данного учебного курса). В этом

задании предлагается рассматривать целевой признак как категориальный (с двумя категориями).

Замечание 2. При выполнении данного задания не обязательно выполнять расчеты для всех имеющихся категориальных признаков. Некоторые выводы можно сделать на основе первичного изучения набора (без выполнения расчетов).

В результате выполнения данного задания должны быть отобраны наиболее значимые категориальные признаки (1-2-3 – как покажут результаты анализа).

Все выполняемые действия и полученные выводы должны быть прокомментированы и обоснованы.