**4. ПОИСК АССОЦИАЦИЙ**

1. **Описание задачи**

Во многих интернет магазинах требуется регистрация пользователей. Помимо всего прочего, данные регистрации могут быть использованы в следующих целях.

При входе на сайт магазина пользователю предоставляется возможность просмотреть каталоги товаров и цены на них. При этом количество товаров может быть настолько велико, что некоторым пользователям может просто не хватить терпения листать список в поисках нужного ему товара – и он просто уйдет с сайта, даже если нужный ему товар здесь есть и он самый дешевый и качественный.

 Чтобы этого не произошло осуществляется сужение области поиска, используя параметры пользователя.

Например.

- пользователь: девушка, 18 лет, студентка, не замужняя – с высокой вероятностью закажет косметику и/или кондитерские изделия;

- пользователь мужчина, 30 лет, оффисный работник, не женатый - с высокой вероятностью закажет алкогольные и/или табачные изделия;

и т.д.

Т.е. в сокращенный список товаров должны быть включены те товары, которые с наибольшей вероятностью купит пользователь, с характерными для него параметрами.

Отметим, что фирма MicroSoft решает подобные задачи с использованием нейросетевых технологий.

1. **Пример**

На сетевом диске имеется файл "Интернет магазин.xls". В нем приводятся сведения о покупаемой группе товаров пользователей с различным профилем свойств.

При этом рассматриваются только данные о пяти группах товаров:

|  |  |
| --- | --- |
| 1 | Детские товары |
| 2 | Одежда |
| 3 | Кондитерские изделия |
| 4 | Косметика |
| 5 | Алкоголь и табачные изделия |

Набор параметров пользователей ограничен тремя показателями:

|  |  |
| --- | --- |
|  | Пол |
| 1 | м |
| 2 | ж |
|  |  |
|  | Возраст |
| 1 | до 21 |
| 2 | 21-40 |
| 3 | более 40 |
|  |  |
|  | Профессия |
| 1 | Неработающие |
| 2 | Офисный |
| 3 | Рабочие |

Таким образом у нас имеется три входных параметра (пол, возраст и профессия), которые составят профиль покупателя. Количество возможных комбинаций значений входных параметров будет равно 2 \* 3 \* 3 = 18.

Представим эти комбинации в закодированном виде следующим образом (рекомендуется размещение данных по указанным в таблице адресам):

Таблица 1

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | F | G | H | I | J | K | L | M | N | O |
|  |  |  |  |  |  |  |  |  |  |  |
| 3 |  |  |  |  | Группа товаров |  |
| 4 |  | Пол | Возраст | Статус | 1 | 2 | 3 | 4 | 5 |  |
| 5 |  | 1 | 1 | 1 |  |  |  |  |  |  |
| 6 |  | 1 | 1 | 2 |  |  |  |  |  |  |
| 7 |  | 1 | 1 | 3 |  |  |  |  |  |  |
| 8 |  | 1 | 2 | 1 |  |  |  |  |  |  |
| 9 |  | 1 | 2 | 2 |  |  |  |  |  |  |
| 10 |  | 1 | 2 | 3 |  |  |  |  |  |  |
| 11 |  | 1 | 3 | 1 |  |  |  |  |  |  |
| 12 |  | 1 | 3 | 2 |  |  |  |  |  |  |
| 13 |  | 1 | 3 | 3 |  |  |  |  |  |  |
| 14 |  | 2 | 1 | 1 |  |  |  |  |  |  |
| 15 |  | 2 | 1 | 2 |  |  |  |  |  |  |
| 16 |  | 2 | 1 | 3 |  |  |  |  |  |  |
| 17 |  | 2 | 2 | 1 |  |  |  |  |  |  |
| 18 |  | 2 | 2 | 2 |  |  |  |  |  |  |
| 19 |  | 2 | 2 | 3 |  |  |  |  |  |  |
| 20 |  | 2 | 3 | 1 |  |  |  |  |  |  |
| 21 |  | 2 | 3 | 2 |  |  |  |  |  |  |
| 22 |  | 2 | 3 | 3 |  |  |  |  |  |  |
| 23 |  |  |  |  |  |  |  |  |  |  |
| 24 |  |  |  |  |  |  |  |  |  |  |

Подсчитаем сколько товаров разных групп купил каждый тип пользователя.

Для этого в ячейку J5 вводится формула:

=СЧЁТЕСЛИМН($B$4:$B$2745;$G5;$C$4:$C$2745;$H5;$D$4:$D$2745;$I5;$E$4:$E$2745;J$4) которая копируется до ячейки N22 (с форматом функции предлагается разобраться самостоятельно).

Результаты вычислений показаны в таб. 2

Таблица 2

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | F | G | H | I | J | K | L | M | N | O | P | Q |
| 2 |  |  |  |  |  |  |  |  |  |  |  |  |
| 3 |  |  |  |  | Группа товаров |  |  |  |
| 4 |  | Пол | Возраст | Статус | 1 | 2 | 3 | 4 | 5 |  |  |  |
| 5 |  | 1 | 1 | 1 | 11 | 41 | 12 | 3 | 78 | 145 | 5 | 2 |
| 6 |  | 1 | 1 | 2 | 1 | 34 | 44 | 27 | 62 | 168 | 5 | 3 |
| 7 |  | 1 | 1 | 3 | 1 | 34 | 13 | 1 | 110 | 159 | 5 | 2 |
| 8 |  | 1 | 2 | 1 | 8 | 25 | 29 | 10 | 83 | 155 | 5 | 3 |
| 9 |  | 1 | 2 | 2 | 22 | 54 | 31 | 11 | 39 | 157 | 2 | 5 |
| 10 |  | 1 | 2 | 3 | 18 | 53 | 21 | 19 | 37 | 148 | 2 | 5 |
| 11 |  | 1 | 3 | 1 | 12 | 20 | 33 | 31 | 48 | 144 | 5 | 3 |
| 12 |  | 1 | 3 | 2 | 24 | 69 | 21 | 18 | 20 | 152 | 2 | 1 |
| 13 |  | 1 | 3 | 3 | 6 | 36 | 8 | 27 | 45 | 122 | 5 | 2 |
| 14 |  | 2 | 1 | 1 | 2 | 34 | 57 | 61 | 3 | 157 | 4 | 2 |
| 15 |  | 2 | 1 | 2 | 9 | 42 | 42 | 69 | 5 | 167 | 4 | 3 |
| 16 |  | 2 | 1 | 3 | 16 | 33 | 45 | 54 | 17 | 165 | 4 | 3 |
| 17 |  | 2 | 2 | 1 | 8 | 43 | 27 | 50 | 30 | 158 | 4 | 2 |
| 18 |  | 2 | 2 | 2 | 46 | 43 | 19 | 36 | 9 | 153 | 1 | 2 |
| 19 |  | 2 | 2 | 3 | 65 | 33 | 7 | 23 | 15 | 143 | 1 | 2 |
| 20 |  | 2 | 3 | 1 | 30 | 78 | 12 | 10 | 25 | 155 | 2 | 1 |
| 21 |  | 2 | 3 | 2 | 37 | 34 | 19 | 35 | 13 | 138 | 1 | 4 |
| 22 |  | 2 | 3 | 3 | 36 | 49 | 18 | 30 | 23 | 156 | 2 | 1 |
| 23 |  |  |  |  | 352 | 755 | 458 | 515 | 662 | 2742 |  |  |
| 24 |  |  |  |  |  |  |  |  |  | 2742 |  |  |

Здесь в ячейках J5:N22 вычислено количество покупок соответствующей группы товаров для каждого типа покупателя.

В ячейках O5:O22 подсчитано общее количество покупок для каждого типа покупателя.

В ячейках J23:N23 подсчитано общее количество покупок для каждого типа товара.

В ячейках O23 и O24 подсчитано общее количество покупок для строки 23 и столбца O.

В столбцах P и Q показаны номера наиболее вероятных групп товаров, которые покупаются каждым типом покупателя (в порядке убывания).

Таким образом, для придания сайту интернет-магазина описанной функции достаточно организовать вычисления с итоговой таблицей, аналогичной табл. 2.

1. **Варианты заданий**

Получить частоты покупок для следующих профилей:

|  |  |
| --- | --- |
| №варианта | Профиль |
| 1 | Пол – Возраст |
| 2 | Пол – Статус |
| 3 | Возраст – Статус |
| 4 | Пол |
| 5 | Возраст |
| 6 | Статус |
| 7 | Пол – Возраст |
| 8 | Пол – Статус |
| 9 | Возраст – Статус |
| 10 | Пол |
| 11 | Возраст |
| 12 | Статус |

1. **Регрессионно-дисперсионный анализ**

***Основные понятия теории эксперимента***

*Фактор*– входной параметр, каким-то образом влияющий на объект исследования. Стандартное обозначение – X.

*Уровень фактора* – конкретное значение входного параметра. Отметим, что для выяснения влияния фактора на объект необходимо как минимум два уровня. Стандартное обозначение – Xi.

*Многофакторный эксперимент* – эксперимент, в котором исследуется влияние на выходной параметр сразу нескольких факторов (минимум двух).

*Выходной параметр (отклик, функция отклика, параметр оптимизации)* – свойство исследуемого объекта, на который предположительно оказывает влияние фактор (факторы). Стандартное обозначение – Y.

*Взаимодействие факторов* – взаимное влияние факторов на выходной параметр, выражающееся в том, что зависимость выходного параметра от одного фактора зависит от значения уровня другого фактора.

Взаимодействие может быть положительным, отрицательным или полностью отсутствовать.

Положительное взаимодействие выражается в том, что увеличение значения уровня одного фактора при одновременном росте значений второго фактора приводит к непропорциональному росту значений выходного параметра.

Отрицательное взаимодействие выражается в том, что с увеличением значения уровня одного фактора при одновременном росте значений второго фактора происходит непропорционально высокое падение значений выходного параметра.

*Ортогональность плана эксперимента* – признак ортогональности плана рассмотрим на примере.

Предположим, что имеется следующий план фвухфакторного эксперимента:

|  |  |  |
| --- | --- | --- |
| X1(температура, 0С) | X2(давление, атм) | Y |
| 20 | 2 | Y1 |
| 40 | 4 | Y2 |
| 60 | 6 | Y3 |

Очевидно, что приведенный план является некорректным, поскольку в результате эксперимента не будет понятно, какой фактор на самом деле влияет на выходной параметр.

И в то же время план:

|  |  |  |
| --- | --- | --- |
| X1(температура, 0С) | X2(давление, атм) | Y |
| 20 | 2 | Y1 |
| 20 | 4 | Y2 |
| 20 | 6 | Y3 |
| 40 | 2 | Y4 |
| 40 | 4 | Y5 |
| 40 | 6 | Y6 |
| 60 | 2 | Y7 |
| 60 | 4 | Y8 |
| 60 | 6 | Y9 |

является корректным.

В терминах теории эксперимента такой план является полным факторным экспериментом (ПФЭ), поскольку реализованы все комбинации значений уровней факторов. При этом имеется возможность оценить влияние на выходной параметр каждого отдельного фактора и их взаимодействий.

В терминах планирования эксперимента полученная ранее таблица 2 является результатом полного факторного эксперимента, в котором имеется три входных параметра (пол, возраст, профессия) и пять выходных параметров (объем продаж пяти групп товаров). При этом факторы являются качественными, а параметры оптимизации – количественными.

Для обработки результатов подобных экспериментов применяется дисперсионный анализ. Однако в этом случае возможно применение специфической разновидности регрессионного анализа.

Суть стандартного регрессионного анализа заключается в получении параметров уравнения регрессии, которое является математической моделью какого-то процесса В простейшем случае такая модель представляет собой линейную зависимость следующего вида:

 (1).

В случае экспериментов с количественными факторами также можно испльзовать подобную модель. Но качественный характер факторов накладывает свою специфику на вид модели, метод расчета коэффициентов и на правила ее использования. Формально вид модели для объектов с качественными факторами аналогичен уравнению (1):

, (2)

где *y* – выходной параметр;

 *n* - количество факторов;

 *b0* – общее среднее эксперимента;

 *Fi* – эффект *i*–ого фактора.

Для рассматривемого примера уравнение (2) принимает вид:

, (3)

где P – величина эффекта фактора пола покупателя;

 V - величина эффекта фактора возраста покупателя;

 S - величина эффекта фактора профессии покупателя.

 Но поскольку все три фактора являются качественными, то уравнение (3) должно быть записано в виде:

, (4)

где *yijk* – значение выходного параметра при *i*–ом уровне первого фактора, *j*-ом уровне второго фактора и *k*–ом уровне третьего фактора;

 P, V, S – означают присутствие соответствующего уровня данного фактора в уравнении. Эти величины равны единице для i-ого уровня первого фактора, j-ого уровня второго фактора и k-ого уровня третьего фактора. Остальные их значения равны нулю.

 Например, для вычисления значения выходного параметра на втором уровне первого фактора, первом уровне второго фактора и третьем уровне третьего фактора уравнение сводится к виду:

 (4а)

 Кажущаяся сложность в использовании уравнения компенсируется простотой вычисления его параметров.

 Уравнение (4) можно переписать в следующем виде:

, (5)

 где m, n, k - количество уровней соответствующего фактора.

Для него составляется следующая система уравнений:

  (6)

где q - общее количество данных;

 m - количество уровней первого фактора;

n - количество уровней второго фактора;

p - количество уровней третьего фактора.

Полученная система является очень разреженной, благодаря свойству ортогональности плана эксперимента. Это свойство приводит к тому, что все суммы в уравнении (5) равны 0.

Рассчитаем параметры уравнения для первой группы товаров. Для удобства воспроизведем соответствующий фрагмент табл. 2.

Таблица 3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Пол | Возраст | Статус | Группа 1 | Расчетныезначения |
| 1 | 1 | 1 | 11 | -9,168 |
| 1 | 1 | 2 | 1 | 2,166 |
| 1 | 1 | 3 | 1 | 2,666 |
| 1 | 2 | 1 | 8 | 11,999 |
| 1 | 2 | 2 | 22 | 23,333 |
| 1 | 2 | 3 | 18 | 23,833 |
| 1 | 3 | 1 | 12 | 10,333 |
| 1 | 3 | 2 | 24 | 21,667 |
| 1 | 3 | 3 | 6 | 22,167 |
| 2 | 1 | 1 | 2 | 7,054 |
| 2 | 1 | 2 | 9 | 18,388 |
| 2 | 1 | 3 | 16 | 18,888 |
| 2 | 2 | 1 | 8 | 28,221 |
| 2 | 2 | 2 | 46 | 39,555 |
| 2 | 2 | 3 | 65 | 40,055 |
| 2 | 3 | 1 | 30 | 26,555 |
| 2 | 3 | 2 | 37 | 37,889 |
| 2 | 3 | 3 | 36 | 38,389 |
|  |  |  | 352 |  |

b0 = 352 / 18 = 19,556;

b11 = (11+1+1+8+22+18+12+24+6) / 9 – 19,556 = -8,111

b12 = (2+9+16+8+46+65+30+37+36)/9 – 19,556 = 8,111

b21= (11+1+1+2+9+16)/6 – 19,556 = -12,89

b22 = (8+22+18+8+46+65)/6-19,556 = 9,556

b23 = (12+24+6+30+37+36)/6-19,556 = 4,611

b31= (11+8+12+2+8+30)/6-19,556 = -7,723

b32 = (1+22+24+9+46+37)/6-19,556 = 3,611

b33 = (1+18+6+16+65+36)/6-19,556 = 4,111

**Проверка адекватности уравнения**

Проверка адекватности уравнения производится по схеме, применяемой для стандартного регрессионного анализа.

Для этого:

1. С помощью полученных коэффициентов рассчитываются теоретические значения частот покупки.

Например:

Y111 = 19,556 + (-8,111) + (-12,89) + (-7,723) = - 9,167

Y112 = 19,556 + (-8,111) + (-12,89) + (9,556) = 2,116

и т.д.

Результаты расчетов приведены в последней колонке табл. 3.

2. С помощью функции =ДИСПА(J5:J22)\*17 рассчитывается общая дисперсия исходных данных (равна 5178,444).

3. С помощью функции =СУММКВРАЗН(J5:J22;O5:O22) рассчитывается остаточная сумма квадратов данных, полученная в результате применения уравнения регрессии (равна 1945,333).

4. Рассчитывается сумма квадратов, приходящаяся на уравнение регрессии. Равна = 5178,444 – 1945,333 = 3233,111.

5. Рассчитываются дисперсии ошибки и уравнения.

Для ошибки =1945,333 / 12 = 162,111 (12 - число степеней свободы для ошибки).

Для уравнения = 3233,111 / 5 = 646,622 (5 – число степеней свободы для уравнения. Равно *f = k + 1,* где *k* – число используемых при расчетах количество коэффициентов уравнения).

6. Рассчитывается критерий Фишера = 646,622 / 162,111 = 3,99

7. Рассчитанный критерий Фишера сравнивается с критическим значением (определяется с помощью функции =F.ОБР.ПХ(0,05;5;12) = 3,11).

8. Если рассчитанное значение критерия больше критического, то уравнение признается адекватным. В противном случае уравнение не адекватно. В данном случае уравнение адекватно.

1. **Варианты заданий**

|  |  |  |
| --- | --- | --- |
| №варианта | Получить уравнение для | Профиль |
| 1 | 2 группы товаров | Пол – Возраст |
| 2 | 3 группы товаров | Пол – Статус |
| 3 | 4 группы товаров | Возраст – Статус |
| 4 | 5 группы товаров | Пол |
| 5 | 2 группы товаров | Возраст |
| 6 | 3 группы товаров | Статус |
| 7 | 4 группы товаров | Пол – Возраст |
| 8 | 5 группы товаров | Пол – Статус |
| 9 | 2 группы товаров | Возраст – Статус |
| 10 | 3 группы товаров | Пол |
| 11 | 4 группы товаров | Возраст |
| 12 | 5 группы товаров | Статус |