Методы многомерной статистики для студентов-заочников

Алексеева Н.П.

1 Кластерный анализ

1.1 Информация и способы ее измерения

Всем хорошо известно, что в любом живом существе происходит передача информации. Органы чувств собирают информацию о внешнем мире, нервная система передает ее в мозг, который ее перерабатывает и рассылает приказы на основе поступившей информации по нервным волокнам в органы и мышцы.

Аналогично передается информация в любом действующем учреждении: передача информации осуществляется в виде докладных записок, распоряжений, запросов и т.д.

Вычислительная техника по заданной программе, т.е. на основе заданной информации, с огромной скоростью перерабатывает большие количества информации.

Перечень примеров можно было бы продолжить, но ясно одно, что уровень достижения в большей степени определяется оперативностью работы с информацией, т.е. тем, с какой скоростью и с какой надежностью осуществляется передача, хранение (защита) и переработка информации.

Вопросы передачи и хранения информации тесно связаны со способом измерения информации и ее кодированием. Вопросы переработки информации опираются на многомерные статистические методы.

Измерять информацию, подобно расстоянию, массе, количеству теплоты и т.д., стали относительно недавно. Поясним способ измерения информации на основе игры "Бар-Кохба". Один из участников игры должен отгадать, что задумал другой, задавая ему вопросы, допускающие только два ответа: "да"или "нет". Обычно задумывают какие-либо предметы или строки стихотворений. Количество информации, необходимое для отгадывания, можно измерить количеством вопросов, которые требуется задать при наиболее рациональной тактике дознания.

Запишем полученные ответы, заменив единицами положительные ответы и нулями отрицательные. Получим последовательность нулей и единиц 001100011010..., заменяющую последовательность ответов. Процесс замены ответов нулями и единицами называется *кодированием*, а сама последовательность – *кодовым словом*.

Длина кодового слова может служить мерой количества информации.

Ответ на любой вопрос, допускающий два равно вероятных ответа "да"или "нет", содержит одну единицу информации, которую называют "бит".

Иными словами, количество информации, которое можно с одинаковой вероятностью закодировать нулем или единицей, будем считать единицей информации. Заметим, что при измерении количества информации мы сознательно оставляем без внимания содержание информации. Например, ответы на вопросы: "Любите ли вы сыр рокфор?"и "Успели ли вы сегодня выспаться?"содержат по одной единице информации.

Количество информации H, необходимое для выявления объекта, о котором только известно, что он принадлежит множеству из N объектов, вычисляется при помощи *формулы Хартли*: $H = log_2 N$. Например, мне известно, что у одного из 16 студентов сегодня день рождения. Эту неопределенность я могу устранить, задав немногословному старосте несколько вопросов, на которые он может отвечать только кивая головой в знак ответов "да"или "нет". Сколько вопросов мне нужно задать? – Ответ: $4 = log_2 16$. Для этого лучше всего разделить группу на две части по восемь человек и задать вопрос: "В первой половине именинник?". Восьмерку с именинником следует опять разделить пополам и задать аналогичный вопрос, и т.д. Четвертый вопрос устранит неопределенность полностью.

Но оказывается, что все эти вопросы можно задать все сразу. Для этого нужно уметь кодировать при помощи *двоичного кода*.

Кодовое слово $(a_0, a_1, \ldots, a_k)$, где $a_i = 0$ или 1, можно представить в виде натурального числа

$$n = a_0 \cdot 2^0 + a_1 \cdot 2^1 + a_2 \cdot 2^2 + \ldots + a_k \cdot 2^k.$$

Например, кодовое слово (0001) соответствует натуральному числу $8 = 0 \cdot 2^0 + 0 \cdot 2^1 + 0 \cdot 2^2 + 1 \cdot 2^3$. И наоборот, для любого натурального числа можно получить его двоичное представление: k является целой частью логарифма $log_2 n$, т.е. $k = \lfloor log_2 n \rfloor$. Затем число n представляется в виде степеней двойки, например, для натурального числа 21 имеем $k = \lfloor log_2 21 n \rfloor = 4$ и $21 = 16 + 5 = 16 + (4 + 1) = 2^4 + 2^2 + 2^0$. Отсюда $a_4 = 1, a_3 = 0, a_2 = 1, a_1 = 0, a_0 = 1$ и кодовое слово имеет вид (10101).

Вернемся к задаче отгадывания именинника. Закодируем номера в списке студентов в двоичной системе. Для удобства будем считать 16-й номер нулевым.

a_3	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
a_2	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
a_1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
a_0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
n	16	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Четыре вопроса, которые можно задавать все сразу, звучат так: "Верно, что первая, вторая, третья и четвертая цифра в закодированном номере студента в списке группы равна нулю?"Например, последовательность ответов "да", "нет", "да","да", соответствующая кодовому слову (1011), укажет на номер $1 \cdot 2^0 + 0 \cdot 2^1 + 1 \cdot 2^2 + 1 \cdot 2^3 = 1 + 4 + 8 = 13$.

В случае, когда количество N элементов множества не является степенью двойки, количество информации, вычисляемое по формуле Хартли, не является целым числом. Например, $log_2 12 = 3.58$. Это число равно среднему числу задаваемых вопросов при многократном отгадывании.

Если в множестве элементы не равновероятны, то для вычисления информации, необходимой для отгадывания элемента этого множества, используется *формула Шеннона.* Обозначим через

$$\xi : \begin{pmatrix} x_1 & \dots & x_n \\ p_1 & \dots & p_n \end{pmatrix}$$

случайный эксперимент, а через $x_1, \dots, x_n$ возможные исходы этого эксперимента, которые осуществляются с вероятностями $p_1, \dots, p_n$, $\sum\limits_{k=1}^{n} p_k = 1$. Количество информации $H(\xi)$, которую нужно получить, чтобы узнать, какой исход эксперимента осуществлен, вычисляется по формуле Шеннона:

$$H(\xi) = \sum_{k=1}^{n} p_k \log_2 \frac{1}{p_k}.$$

Например, из перемешанной колоды 36 карт вынимается одна карта. Значениями случайного эксперимента ξ_1 являются все 36 карт, вероятность появления каждой равна $1/36$:

$$\xi_1 : \begin{pmatrix} x_1 & \dots & x_{36} \\ \frac{1}{36} & \dots & \frac{1}{36} \end{pmatrix}.$$

В этом случае формула Шеннона совпадает с формулой Хартли и

$$H(\xi_1) = \sum_{k=1}^{36} \frac{1}{36} \log_2 36 = 5.17.$$

Пусть ξ_2 – некий случайный эксперимент, в результате которого из такой же колоды карт вынимается или нет туз: $\xi_2 : \begin{pmatrix} \text{'не туз'} & \text{'туз'} \\ \frac{32}{36} & \frac{4}{36} \end{pmatrix}$

$$H(\xi_2) = \frac{32}{36} \log_2 \frac{36}{32} + \frac{4}{36} \log_2 \frac{36}{4} = 0.352 + 0.151 = 0.503.$$

Необходимая для отгадывания результата случайного эксперимента ξ информация $H(\xi)$ равна неопределенности этого эксперимента, поэтому часто $H(\xi)$ называют *энтропией* распределения случайного эксперимента ξ.

1.2 Информационное разнообразие групп и информационный выигрыш

Качественными называются признаки, градации которых можно менять местами. Например, цветы могут быть красными, синими, белыми, желтыми и т.д. Пусть имеется группа n индивидов, характеризуемая признаком ξ со значениями $x_1, \ldots, x_m$, $n = a_1 + a_2 + \ldots + a_m$, где a_i – количество индивидов со значением x_i признака ξ. Информационное разнообразие I этой группы измеряется формулой:

$$I = n \ln n - \sum_{i=1}^{m} a_i \ln a_i.$$

Например, информационное разнообразие букета, состоящего из двух гвоздик, одного тюльпана и двух хризантем равно

$$I = 5 \ln 5 - 2 \ln 2 - 1 \ln 1 - 2 \ln 2 = 5.275.$$

В группе с одинаковыми значениями признака информационное разнообразие равно нулю, так как $I = n \ln n - 0 \cdot \ln 0 - n \ln n = 0$. Несложно показать, что информационное разнообразие группы пропорционально объему группы n и энтропии распределения $H(\xi)$ признака ξ и равно $I = nH(\xi)/\ln 2$.

Если группа индивидов характеризуется несколькими признаками, то разнообразие складывается из информационных разнообразий по каждому признаку.

Информационный выигрыш от слияния двух групп A и B с разнообразиями I_A и I_B равен

$$\Delta I = I_{A+B} - I_A - I_B.$$

Информационный выигрыш от слияния групп может быть нулевым или положительным. Например, если объединяются букет роз и букет ромашек, разнообразие которых по отдельности равно нулю, то выигрыш положительный, так как разнообразие смешанного букета положительно. Если к ромашке с колокольчиком присоединить еще одну пару ромашки с колокольчиком, то информационный выигрыш окажется нулевым, так как

$$\begin{aligned} I_A &= 2 \ln 2 - 1 \ln 1 - 1 \ln 1 = 2 \ln 2, \\ I_B &= 2 \ln 2, \\ I_{A+B} &= 4 \ln 4 - 2 \ln 2 - 2 \ln 2 = 4 \ln 4 - 4 \ln 2 = 4(\ln 4 - \ln 2) = 4 \ln 2, \\ \Delta I &= 4 \ln 2 - 2 \ln 2 - 2 \ln 2 = 0. \end{aligned}$$

1.3 Кластеризация индивидов по качественным признакам

Одной из задач кластерного анализа является агрегирование множества индивидов на однородные (одинаково распределенные) подмножества. В качестве иллюстративного примера рассмотрим данные международных сравнений из статистического справочника за 1999 год о смертности от разных причин (табл.1). Признаки: смертность (умершие на 100000 населения) от заболеваний системы кровообращения (ξ_1), злокачественных новообразований (ξ_2), инфекционных заболеваний (ξ_3) и несчастных случаев, убийств и других внешних воздействий (ξ_4) являются количественными признаками. Для устранения влияния случайностей и выявления наиболее общих тенденций преобразуем эти признаки в качественные:

$$\widetilde{\xi_1} = \begin{cases} 0, & \text{при } \xi_1 < 300, \\ 1 & \text{иначе} \end{cases}, \quad \widetilde{\xi_2} = \begin{cases} 0, & \text{при } \xi_2 < 189, \\ 1 & \text{иначе} \end{cases},$$

$$\widetilde{\xi_3} = \begin{cases} 0, & \text{при } \xi_3 < 10, \\ 1 & \text{иначе} \end{cases}, \quad \widetilde{\xi_4} = \begin{cases} 0, & \text{при } \xi_4 < 50, \\ 1 & \text{иначе} \end{cases}.$$

Вектор значений качественных признаков $\widetilde{\xi} = [\widetilde{\xi_1}, \widetilde{\xi_2}, \widetilde{\xi_3}, \widetilde{\xi_4}]^T$ удобнее закодировать в двоичной системе $\widehat{\xi} = \widetilde{\xi_1} + 2\widetilde{\xi_2} + 4\widetilde{\xi_3} + 8\widetilde{\xi_4}$. В табл.1 высокие показатели смертности отмечены звездочками. Например,

- Япония с низкими показателями смертности от заболеваний системы кровообращения, от злокачественных новообразований и от несчастных случаев и высокой смертностью от инфекционных заболеваний, соответствующими вектор-строке (0010), получает код, равный 4,

- Мексика с вектор-строкой (0011) – код 12 ,
- Италия, Нидерланды, Канада и Австралия с вектор-строкой (0100) – код 2,
- Франция и Бельгия с вектор-строкой (0101) – код 10,
- Финляндия, Болгария и Австрия с вектор-строкой (1001) – код 9,
- Румыния с вектор-строкой (1011) – код 13,
- Польша, Дания и Венгрия с вектор-строкой (1101) – код 11,
- США и Россия с вектор-строкой (1111) – код 15.

Таблица 1

Страны	код	дв.код	Смертность от причин			
			заболев. сист. кровообр.(ξ_1) (300)	злокачеств. новообраз. (ξ_2) (189)	инфекц. заболев.(ξ_3) (10)	несчаст. сл. и др.внеш.возд.(ξ_4) (50)
Италия	2	0100	284.8	199.9*	2.8	40.8
Канада	2	0100	238.7	190*	5.5	43.6
Австралия	2	0100	287.9	189.8*	5.6	39.4
Нидерланды	2	0100	269.3	206.9*	6.3	29.8
Япония	4	0010	187.4	159.8	10.3*	43.3
Австрия	9	1001	369.3*	188.9	1.9	54.3*
Финляндия	9	1001	354*	163	6.3	79.1*
Болгария	9	1001	706.4*	163.6	9	66.8*
Бельгия	10	0101	267.7	214.5*	7.8	56.3*
Франция	10	0101	182.8	194.3*	8.8	64.5*
Венгрия	11	1101	603.8*	280.9*	7.3	102.4*
Дания	11	1101	316.5*	234.2*	9.9	55.9*
Польша	11	1101	532.2*	215.4*	7	77*
Мексика	12	0011	295.3	122	36.4*	81.8*
Румыния	13	1011	747.8*	162.5	15.2	80.8*
США	15	1111	307.2*	195*	11.4*	55.5*
Россия	15	1111	736.1*	194.4*	21.4 *	187.4*

Процедура кластеризации индивидов в случае качественных признаков начинается с построения таблицы информационных выигрышей от объединения всевозможных пар индивидов. В нашем примере индивидами являются группы стран с одинаковой структурой причин смертности, характеризуемой одинаковым кодом. Всего имеется 8 групп стран. Рассмотрим всевозможные $28 = C_8^2 = 8 \cdot 7/2$ попарных объединения этих групп и вычислим их информационные разнообразия с учетом четырех признаков. Так как информационное разнообразие стран с одинаковым кодом равно нулю, то информационное разнообразие попарных объединений равно информационным выигрышам от объединения.

Таблица 2

Таблица информационных выигрышей

коды	2	4	9	10	11	12	13	15
2	-	5.004	14.341	3.819	9.561	7.506	10.008	11.457
4		-	6.748	5.729	8.997	**1.386**	2.773	5.729
9			-	6.730	4.159	4.499	2.249	6.730
10				-	3.365	3.819	5.729	5.545
11					-	6.748	4.499	3.365
12						-	1.389	3.819
13							-	1.910
15								-

Например, для того чтобы получить информационный выигрыш от объединения группы стран, у которых код $2 = (0100)$ (Италия, Канада, Нидерланды, Австралия), с Японией (код $4 = (0010)$), нужно вычислить разнообразие смешанной группы по всем четырем признакам, представленным в следующей таблице:

Таблица 3

	ξ_1	ξ_2	ξ_3	ξ_4
Италия	0	1	0	0
Канада	0	1	0	0
Нидерланды	0	1	0	0
Австралия	0	1	0	0
Япония	0	0	1	0
Разнообразие $I(\xi)$	0	$5\ln 5 - 4\ln 4 - 1\ln 1 = 2.502$	$5\ln 5 - 1\ln 1 - 4\ln 4 = 2.502$	0

Разнообразие этой смешанной группы по всем признакам равно $I(\xi_1)+I(\xi_2)+I(\xi_3)+I(\xi_4) = 0+2.502+2.502+0 = 5.004$. информационный выигрыш равен 5.004. Аналогично вычисляются остальные выигрыши.

Затем из таблицы информационных выигрышей выберем пару кодов с наименьшим выигрышем (группы стран с кодами 4 и 12) и объединим их в группу $A(4,12)$. Таким образом, в первую наиболее однородную группу вошли Япония и Мексика. Найдем информационный выигрыш от объединения группы $A(4,12)$ со всеми оставшимися группами стран. Например, при объединении группы $A(4,12)$ с группой стран, код которых равен 9 (Австрия, Болгария, Финляндия), получаем группу из 5 индивидов. По признаку ξ_1 (смертность от заболеваний системы кровообращения) имеем две страны с низким показателем, и три с высоким, следовательно, информационной разнообразие равно

$$I_{A,2}(\xi_1) = 5\ln 5 - 2\ln 2 - 3\ln 3 = 3.365.$$

Все эти страны имеют невысокую смертность от злокачественных новообразований, следовательно, по признаку ξ_2 информационное разнообразие $I_{A,2}(\xi_2) = 0$. Среди этих пяти стран Япония и Мексика имеют высокие значения смертности от инфекционных заболеваний, а Австрия, Болгария, Финляндия, наоборот, невысокие. Следовательно, разнообразие равно

$$I_{A,2}(\xi_3) = 5\ln 5 - 2\ln 2 - 3\ln 3 = 3.365.$$

И, наконец, только Мексика имеет высокую смертность от несчастных случаев и убийств, и разнообразие равно

$$I_{A,2}(\xi_4) = 5\ln 5 - 1\ln 1 - 4\ln 4 = 2.502.$$

Складываем разнообразия по всем признакам и получаем

$$I_{A,2} = I_{A,2}(\xi_1) + I_{A,2}(\xi_2) + I_{A,2}(\xi_3) + I_{A,2}(\xi_4) = 3.365 + 0 + 3.365 + 2.502 = 9.232.$$

Поскольку $I_A = 1.386$, $I_9 = 0$, то информационный выигрыш от объединения равен

$$\Delta I(A,2) = I_{A,2} - I_A - I_2 = 9.232 - 1.386 = 7.846.$$

Таблица 4

	$A(4,12)$ $I_A = 1.386$	2 $I_2 = 0$	9 $I_9 = 0$	10 $I_{10} = 0$	11 $I_{11} = 0$	13 $I_{13} = 0$	15 $I_{15} = 0$
$A(4,12)$	-	8.955	7.846	6.408	9.181	2.433	6.408
2		-	14.341	3.819	9.561	10.008	11.457
9			-	6.730	4.159	2.249	6.730
10				-	3.365	5.729	5.545
11					-	4.499	3.365
13						-	**1.910**
15							-

Группу $B(13,15)$ образуют страны: Румыния и США, Россия.

Таблица 5

	$A(4,12)$ $I_A = 1.386$	$B(13,15)$ $I_B = 1.910$	2 $I_2 = 0$	9 $I_9 = 0$	10 $I_{10} = 0$	11 $I_{11} = 0$
$A(4,12)$	-	5.936	8.955	7.846	6.408	9.181
$B(13,15)$		-	15.302	6.068	7.323	3.958
2			-	14.341	3.819	9.561
9				-	6.730	4.159
10					-	**3.365**
11						-

Группу $C(10,11)$ образовали страны: Франция, Бельгия и Венгрия, Дания, Польша.

Таблица 6

	$A(4,12)$ $I_A = 1.386$	$B(13,15)$ $I_B = 1.910$	$C(10,11)$ $I_C = 3.365$	2 $I_2 = 0$	9 $I_9 = 0$
$A(4,12)$	-	**5.936**	9.409	8.955	7.846
$B(13,15)$		-	6.564	15.302	6.068
$C(10,11)$			-	6.679	17.105
2				-	14.341
9					-

Объединим в одну группу группы A и группу B. Информационное разнообразие

$$I_D = I_{A+B} = \Delta I + I_A + I_B = 5.936 + 1.386 + 1.910 = 9.232.$$

Эту группу образуют 5 стран: Япония, Мексика и Румыния, США, Россия.

Таблица 7

	$D(4,12,13,15)$ $I_D = 9.232$	$C(10,11)$ $I_C = 3.365$	2 $I_2 = 0$	9 $I_9 = 0$
$D(4,12,13,15)$	-	8.636	14.590	8.072
$C(10,11)$		-	**6.679**	17.105
2			-	14.341
9				-

На этом этапе объединению подлежат группы $C(10,11)$ и 2, получаем $E(10,11,2)$. Разнообразие

$$I_E = \Delta I(C,2) + I_C + I_2 = 6.679 + 3.365 + 0 = 10.044.$$

Страны с кодом 2 (Италия, Канада, Нидерланды и Австралия) присоединились к Франции, Бельгии, Венгрии, Дании, Польше.

Таблица 8

	$D(4,12,13,15)$ $I_D = 9.232$	$E(10,11,2)$ $I_E = 10.044$	9 $I_9 = 0$
$D(4,12,13,15)$	-	23.869	**8.072**
$E(10,11,2)$		-	11.191
9			-

На последнем этапе группа 9 (Австрия, Болгария и Финляндия) присоединяется к группе $D(4,12,13,15)$.

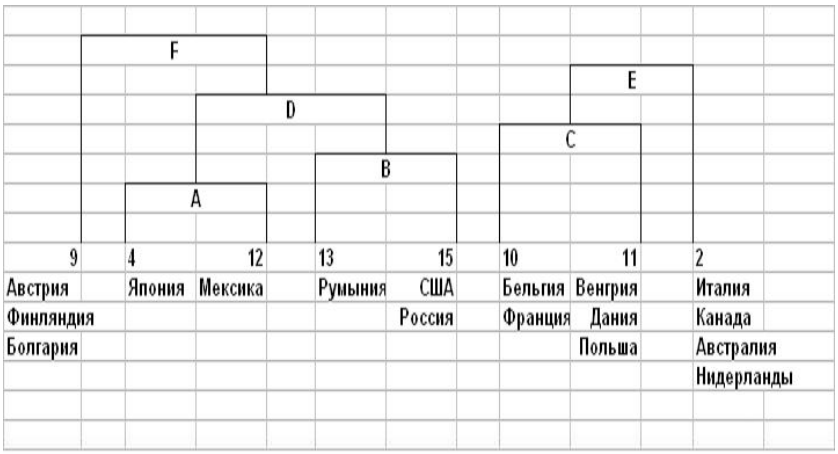

Рис. 1: Дерево дихотомической кластеризации.

1.4 Кластеризация индивидов в случае метрических признаков

Пусть имеется n индивидов и p признаков, x_{ij} – наблюдение i-го признака и j-го индивида. В качестве мер сходства и различия используются следующие метрики:

1. евклидова метрика, равная геометрическому расстоянию между точками (например, индивидами 1 и 2) в p-мерном пространстве:

$$\sqrt{\sum_{k=1}^{p}(x_{i1}-x_{i2})^2}$$

2. квадрат евклидовой метрики:

$$\sum_{k=1}^{p}(x_{i1}-x_{i2})^2$$

3. манхеттенская метрика:

$$\sum_{k=1}^{p}|x_{i1}-x_{i2}|$$

4. Метрика Брея-Картиса:

$$\frac{\sum\limits_{k=1}^{p}|x_{i1}-x_{i2}|}{\sum\limits_{k=1}^{p}x_{i1}+\sum\limits_{k=1}^{p}x_{i2}}$$

5. Канберровская метрика:

$$\sum_{k=1}^{p}\left(\frac{|x_{i1}-x_{i2}|}{x_{i1}+x_{i2}}\right).$$

6. Метрика Адкинса:

$$\sum_{k=1}^{p}\left(\frac{|x_{i1}-x_{i2}|}{|x_{i1}|+|x_{i2}|}\right).$$

Последние три метрики используются часто при большом различии масштабов измерения. Например, если единицы измерения одного признака миллиметры, а другого километры, то ненормированные евклидовы и манхеттенские метрики будут измерять меру сходства и различия только по признаку с большими единицами измерения.

Кластеризация индивидов осуществляется на основе выбора специальной стратегии объединения. Сначала рассматриваются всевозможные пары индивидов и пары с минимальным различием объединяются в группы. Далее встает вопрос, как измерить расстояние между группами. Существует несколько стратегий измерения расстояния между группами. Все они по формальному признаку сводятся к выражению:

$$d_{kh}=\alpha_i d_{ih}+\alpha_j d_{jh}+\beta d_{ij}+\gamma|d_{ih}-d_{jh}|,$$

где через h обозначена группа, полученная в результате объединения индивидов i и j, $h=i\cup j$, через k – группа, расстояние до которой от группы h определяется в результате выбранной стратегии. Параметры α_i,α_j,β и γ называются параметрами стратегии.

Например, параметры вида $\alpha_i=\alpha_j=0.5$, $\beta=0$, $\gamma=-0.5$ определяют стратегию ближнего соседа, когда за расстояние между группами принимается расстояние между ближайшими соседями. Эта стратегия уменьшает расстояния между группами или, как говорят, сжимает пространство. Параметры вида $\alpha_i=\alpha_j=0.5$, $\beta=0$, $\gamma=0.5$ определяют стратегию дальнего соседа, когда за расстояние между группами принимается самое дальнее расстояние между элементами групп. Эта стратегия расширяет пространство. Параметры $\alpha_i=n_i/n_h,\alpha_j=n_j/n_h$, $\beta=0$, $\gamma=0$ определяют стратегию группового среднего, когда за расстояние между группами принимается расстояние между ближайшими соседями. Через n_i обозначено количество элементов в i-й группе. Недостаток этой стратегии – немонотонность, но она не изменяет пространство. Существует еще центроидная стратегия, когда расстояние между группами определяется как расстояние между центрами тяжести. Стратегия, удовлетворяющая условиям монотонности и сохранения метрики пространства, называется гибкой стратегией. Соответствующие ей параметры имеют вид: $\alpha_i=\alpha_j=0.625$, $\beta=-0.25$, $\gamma=0$.

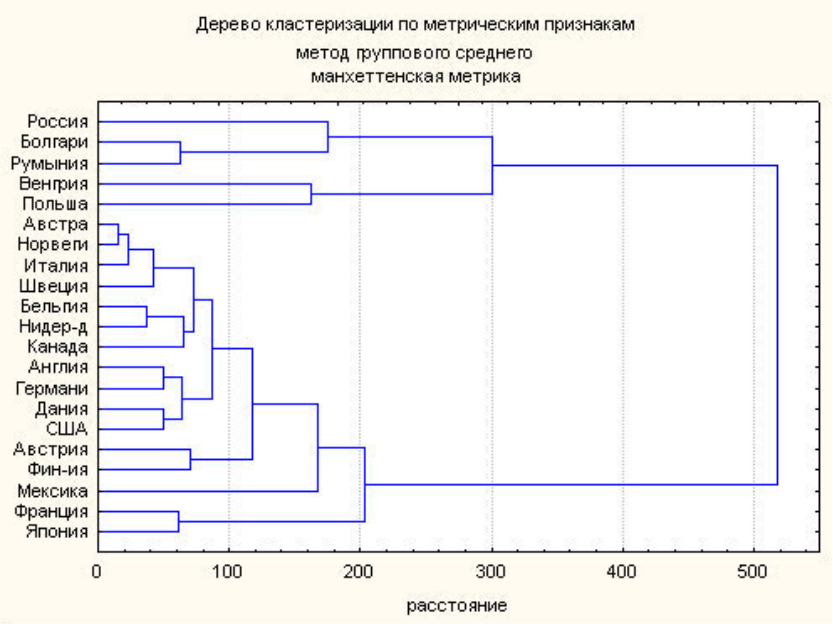

Рис. 2: Дерево кластеризации индивидов в случае исходных метрических признаков.

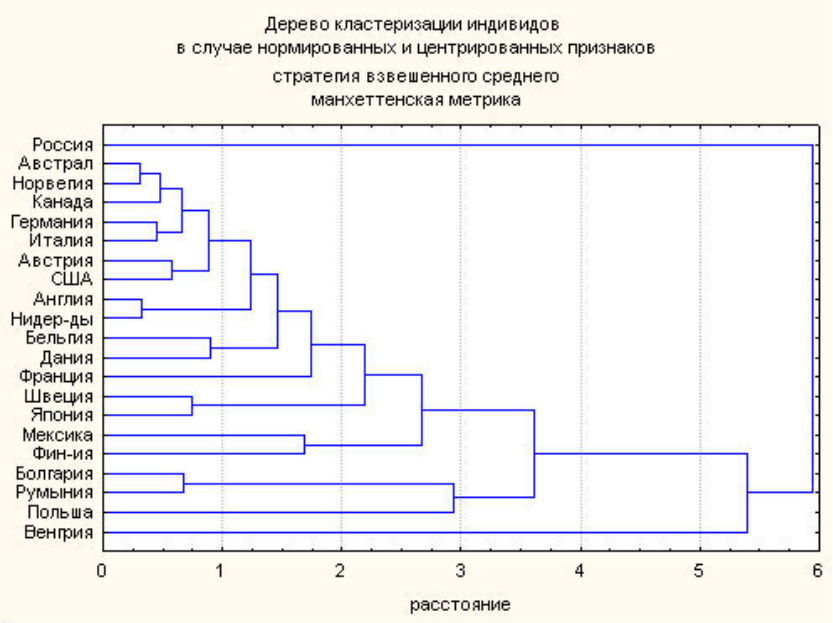

Рис. 3: Дерево кластеризации индивидов в случае нормированных и центрированных метрических признаков.

Для того чтобы все признаки участвовали в кластеризации индивидов, полезно исходные данные центрировать (вычитание математического ожидания) и нормировать (деление на стандартное отклонение). В этом случае кластеризация осуществляется с учетом вклада всех признаков, а не только признака с наибольшим масштабом измерения. Выбор манхеттенской метрики и стратегии группового среднего для нормированных и центрированных данных приводит дереву кластеризации, в котором минимальным расстоянием между собой отличаются страны: Австралия-Норвегия и Англия-Нидерланды.

1.5 Кластеризация признаков

Разбиение признаков на кластеры можно осуществлять также как индивиды. Для этого нужно только транспонировать матрицу наблюдений. Наряду с перечисленными метриками используются также метрики, связанные с коэффициентом корреляции. Напомним, что коэффициент корреляции является характеристикой меры линейной зависимости между признаками ξ и η:

$$\varrho = \frac{E(\xi - E\xi)(\eta - E\eta)}{\sqrt{E(\xi - E\xi)^2}\sqrt{E(\eta - E\eta)^2}}.$$

Если признаки не зависимы, то коэффициент корреляции равен нулю. Если признаки связаны линейной зависимостью вида $\eta = k\xi + b$, то коэффициент корреляции равен ± 1 в зависимости от знака коэффициента k.

При наличии матрицы наблюдений выборочный коэффициент корреляции между первым и вторым признаками вычисляется по формуле:

$$r = \frac{\sum\limits_{j=1}^{n} x_{1j}x_{2j} - n\bar{x_1}\bar{x_2}}{\sqrt{\sum\limits_{j=1}^{n} x_{1j}^2 - n\bar{x_1}^2}\sqrt{\sum\limits_{j=1}^{n} x_{2j}^2 - n\bar{x_2}^2}},$$

где также считаем, что имеется n индивидов и p признаков, x_{ij} – наблюдение i-го признака и j-го индивида, $\bar{x}_i$ – выборочное среднее i-го признака.

В качестве меры сходства между признаками выберем метрику, равную $1 - r^2$. Чем ближе коэффициент корреляции к нулю, тем больше метрика. Чем меньше метрика, тем больше связаны линейной зависимостью, т.е. можно сказать, тем меньше расстояние между признаками.

ПРИМЕР 1. Рассмотрим данные о жизни России за последние 30 лет.

Года	Продолж-ть жизни L	Кол-во (тыс.чел.) чиновников M	Водка (бут.в год) V	Доходы бедных P	Автомобили (шт.на 1000 чел.) A
1970	68.9	1060	25.3	7.8	5.5
1975	68.1	1101	28	9.5	15.3
1980	67.6	1147	30	10.1	30.2
1985	69.2	1204	23.5	10	44.5
1990	69.2	1602	18	9.8	58.6
1995	64.6	1893	38.4	5.5	93.3
1998	67	2777	29.6	6.2	122

Корреляционная матрица и матрица метрик вида $1 - r^2$ имеют вид:

r	L	M	V	P	A
L	1				
M	-0.5008	1			
V	-0.9341	0.2966	1		
P	0.7692	-0.7001	-0.6809	1	
A	-0.5953	0.9537	0.3701	-0.6692	1

$1 - r^2$	L	M	V	P	A
L	0				
M	0.749	0			
V	0.128	0.912	0		
P	0.408	0.510	0.536	0	
A	0.646	**0.091**	0.863	0.552	0

Находим минимальное расстояние между признаками. Оно равно 0.091 и соответствует признакам: количество чиновников и количество автомобилей. Объединяем эти признаки в один, называемый чиновники-автомобили (AM) и таким образом получаем не пять, а четыре признака: L, V, P, AM. Расстояния между первыми тремя известны. Вычислим расстояния между AM и остальными признаками на основе гибкой стратегии, согласно которой имеем:

$$d_{AM,L} = 0.625(d_{A,L} + d_{M,L}) - 0.25d_{A,M} = 0.625(0.646 + 0.749) - 0.25 \cdot 0.091 = 0.849,$$
$$d_{AM,V} = 0.625(d_{A,V} + d_{M,V}) - 0.25d_{A,M} = 0.625(0.863 + 0.912) - 0.25 \cdot 0.091 = 1.087,$$
$$d_{AM,P} = 0.625(d_{A,P} + d_{M,P}) - 0.25d_{A,M} = 0.625(0.552 + 0.510) - 0.25 \cdot 0.091 = 0.641.$$

d	L	V	P	AM
L	0			
V	**0.128**	0		
P	0.408	0.510	0	
A	0.849	1.087	0.641	0

Находим минимальное расстояние между признаками, это расстояние между продолжительностью жизни L и количеством продаваемой водки V (0.128). Получаем три признака: доходы бедных P, водка-продолжительность жизни VL и автомобили-чиновники AM. Если задача состоит в нахождении трех кластеров признаков, то все получено. Если нужно определить, к какому из кластеров: водка-продолжительность жизни VL или автомобили-чиновники AM относится признак "доходы бедных"P, то продолжаем. Расстояние $d_{AM,P} = 0.641$ из вычислений на предыдущем шаге.

$$d_{VL,P} = 0.625(d_{V,P} + d_{L,P}) - 0.25d_{V,L} = 0.625(0.536 + 0.408) - 0.25 \cdot 0.128 = 0.559$$
$$d_{VL,AM} = 0.625(d_{V,AM} + d_{L,AM}) - 0.25d_{V,L} = 0.625(1.087 + 0.849) - 0.25 \cdot 0.128 = 1.178.$$

Отсюда получаем, что доходы бедных лучше прогнозируются по продолжительности жизни и продаваемой водке, нежели по количеству автомобилей и чиновников.

ПРИМЕР 2. Если в качестве метрики рассматривать характеристику $1-r$, показывающую степень удаленности от положительной корреляции, то результат получится несколько иной. Матрица расстояний имеет вид:

$1-r$	L	M	P	A	V
L	0				
M	1.50	0			
P	0.23	1.70	0		
A	1.60	**0.05**	1.67	0	
V	1.93	0.70	1.68	0.63	0

Также наиболее близкими являются признаки A -автомобили и M - чиновники.

d	AM	L	P	VM
AM	0			
L	1.92	0		
P	2.09	**0.23**	0	
V	0.82	1.93	1.68	0

$$d_{AM,L} = 0.625(d_{A,L} + d_{M,L}) - 0.25d_{A,M} = 0.625(1.6 + 1.5) - 0.25 \cdot 0.05 = 1.92,$$
$$d_{AM,V} = 0.625(d_{A,V} + d_{M,V}) - 0.25d_{A,M} = 0.625(0.63 + 0.7) - 0.25 \cdot 0.05 = 0.82,$$
$$d_{AM,P} = 0.625(d_{A,P} + d_{M,P}) - 0.25d_{A,M} = 0.625(1.67 + 1.7) - 0.25 \cdot 0.05 = 2.09.$$

Находим, что минимальное расстояние оказывается между продолжительностью жизни L и доходами бедных P.

Итак, имеется пара признаков AM и LP. К какой паре можно отнести количество продаваемой водки?

$$d_{V,LP} = 0.625(d_{V,L} + d_{V,P}) - 0.25d_{L,P} = 0.625(1.93 + 0.70) - 0.25 \cdot 0.23 = 2.2$$
$$d_{LP,AM} = 0.625(d_{L,AM} + d_{P,AM}) - 0.25d_{L,P} = 0.625(1.92 + 2.09) - 0.25 \cdot 0.23 = 2.45,$$
$$d_{V,AM} = 0.82.$$

Минимальное расстояние оказывается между признаками автомобили-чиновники и количеством продаваемой водки. Автоматически дерево кластеризации можно получить, используя стандартный пакет статистических программ StatSoft, установив в меню в качестве метрики коэффициент корреляции Пирсона, а в качестве стратегии, например, метод группового среднего.

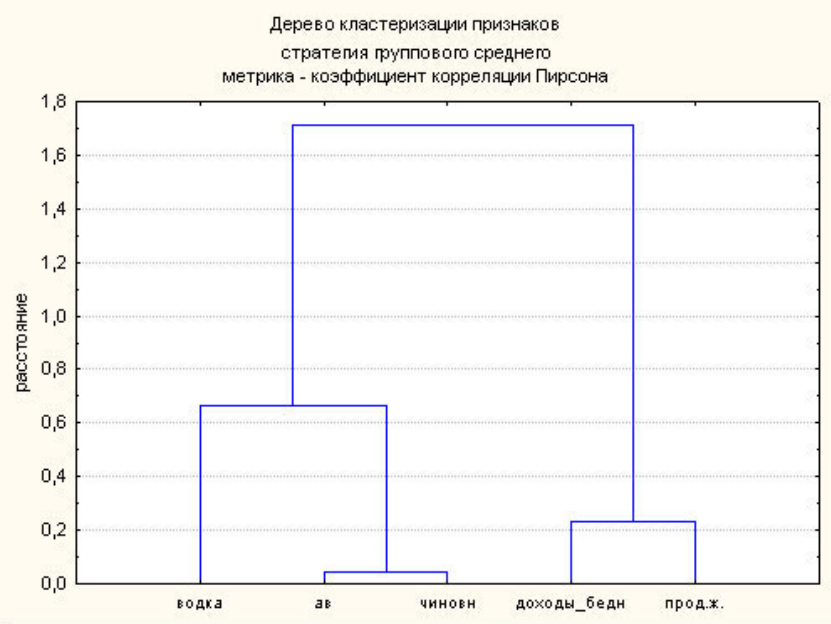

Рис. 4: Дерево кластеризации метрических признаков.

2 Исследование данных на предмет однородности

2.1 Случайные векторы и матрицы. Вектор средних значений и матрица ковариаций случайного вектора. Линейные преобразования случайных векторов

На практике исследователь, как правило, имеет дело с некоторым объектом генеральной совокупности, который характеризуется $p > 1$ признаками. Введем p случайных величин $\xi_1, \ldots, \xi_p$ так, чтобы они соответствовали этим признакам. Полезно рассматривать эти случайные величины как случайный вектор, то есть упорядоченный набор из p элементов в виде столбца:

$$\xi = \begin{bmatrix} \xi_1 \\ \xi_2 \\ . \\ . \\ \xi_p \end{bmatrix}$$

Каждый элемент ξ_i называется *компонентой* случайного вектора. Реализацию $X = (x_1, \ldots, x_p)^T$ случайного вектора ξ, где компоненты x_i являются реализациями случайных величин ξ_i, будем называть *многомерным наблюдением.*

Например, первая строка, описывающая затраты на рекламу авиалиний (6.38, 6.32, 3.37, 2.56) соответствует четырехмерному наблюдению вектора, компоненты которого описывают отдельные затраты на различные виды рекламы. Пусть у нас имеется n объектов (индивидов). У каждого из этих объектов мы измеряем p характеристик. Объединим эти n векторов в так называемую матрицу наблюдений

$$X(p,n) = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ . & . & . & . \\ x_{p1} & x_{p2} & \ldots & x_{pn} \end{bmatrix} = [X_1, \ldots, X_n].$$

Элемент x_{ij}, где $i = 1, \ldots, p$ – номер признака, а $j = 1, \ldots, n$ – номер индивида, есть реализация одномерной случайной величины ξ_{ij}.

Обозначим через $\mu_i = E\xi_i$, $i = 1, \ldots, p$, математические ожидания компонент. Их удобнее представлять в виде вектора средних $\mu = E\xi = (\mu_1, \ldots, \mu_p)^T$.

Аналогично дисперсии σ_i^2 величин ξ_i можно получить из частных распределений. Однако ковариацию σ_{ij} можно получить только из совместного распределения случайных величин ξ_i и ξ_j

$$\sigma_{ij} = cov(\xi_i, \xi_j) = E(\xi_i - \mu_i)(\xi_j - \mu_j),\ i, j = 1, 2, \ldots, p.$$

Заметим, что $\sigma_{ii} = \sigma_i^2,\ \sigma_{ij} = \sigma_{ji}$. Если $\sigma_{ij} = 0$, то случайные величины ξ_i и ξ_j называются некоррелированными. Дисперсии и ковариации вместе образуют ковариационную матрицу $\Sigma(p,p)$ (иначе она называется дисперсионной матрицей и обозначается по аналогии с одномерным анализом $D\xi$), которая является обобщением понятия дисперсия для одномерной случайной величины:

$$\Sigma(p,p) = D\xi = E(\xi - E\xi)(\xi - E\xi)^T = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2p} \\ . & . & \dots & . \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{bmatrix}.$$

Пусть ξ случайный вектор с нулевым математическим ожиданием и ковариационной матрицей $\Sigma_\xi = E\xi\xi^T$. Выясним, что произойдет с такими характеристиками многомерного распределения, как математическое ожидание и дисперсия, при линейном преобразовании случайного вектора. Рассмотрим новый случайный вектор

$$\eta = A\xi + \mu, \tag{1}$$

где матрица коэффициентов линейного преобразования имеет вид:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ . & . & . & . \\ a_{p1} & a_{p2} & \dots & a_{pp} \end{bmatrix}, \tag{2}$$

а μ неслучайный вектор. Очевидно, что вектор математических ожиданий совпадает с вектором μ, так как $E\eta = E(A\xi + \mu) = AE\xi + \mu = \mu$.

	Затраты на рекламу(*ln USD*)			
Рекламодатель	телевидение	пресса	наружная	радио
авиалинии	6.38	6.32	3.37	2.56
автомобили	6.94	8.05	5.89	6.34
безалк.нап.	8.94	5.37	6.86	6.16
безопасность	2.94	3.61	1.10	4.01
быт.тех.	9.16	3.83	3.78	3.61
канцтовары	0.00	5.15	1.10	1.39
компьютеры	6.09	7.61	4.16	2.56
конд.изд.	9.67	5.24	4.36	3.53
мебель	5.09	6.48	5.28	5.42
медицина	6.91	6.91	2.94	5.81
связь	3.43	6.68	5.22	6.65
сервис	2.77	4.42	2.20	4.47
туризм	5.38	6.68	2.83	6.42
фото.обор.	6.77	4.16	4.08	4.48
ювел.изд.	3.53	3.95	1.95	1.10
юрид.усл.	0.69	5.16	1.95	3.14

Таблица 1: Затраты на рекламу различных категорий товаров и услуг по все средствам массовой информации на август 1997

Лемма 1(о ковариации линейного преобразования). $D\eta = AD\xi A^T$. **Доказательство.**

$$\begin{aligned} D\eta = E(\eta - E\eta)(\eta - E\eta)^T = E(A\xi + m - m)(A\xi + m - m)^T = \\ = E(A\xi)(A\xi)^T = EA\xi\xi^T A^T = AE\xi\xi^T A = AD\xi A^T, \end{aligned}$$

так как неслучайную матрицу A можно вынести за знак математического ожидания.

Матрица, элементами которой являются коэффициенты корреляции $\varrho_{ij} = \frac{\sigma_{ij}}{\sqrt{D\xi_i}\sqrt{D\xi_j}}$ называется корреляционной:

$$\Upsilon(p,p) = \begin{bmatrix} \varrho_{11} & \varrho_{12} & \dots & \varrho_{1p} \\ \varrho_{21} & \varrho_{22} & \dots & \varrho_{2p} \\ . & . & \dots & . \\ \varrho_{p1} & \varrho_{p2} & \dots & \varrho_{pp} \end{bmatrix}.$$

В большинстве случаев параметры μ и Σ неизвестны, и вместо них рассматривают их выборочные аналоги. Пусть $X_1, \ldots, X_n$ – случайная выборка, имеющая многомерное нормальное распределение $\mathcal{N}(\mu, \Sigma)$. Тогда выборочные среднее и ковариационная матрица имеют вид:

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k,$$

$$\mathbf{S} = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})(X_k - \bar{X})^T.$$

Выборочная корреляционная матрица $\mathbf{R}$ получается из матрицы $\mathbf{S}$ путем деления элементов s_{ij} на соответствующие стандартные отклонения $\sqrt{s_{ii}}$ и $\sqrt{s_{jj}}$. Например, выборочная корреляционная матрица вектора признаков, компонентами которого являются затраты на рекламу на телевидении, в прессе, на стендах и на радио, имеет вид:

r	телевидение	пресса	наруж.реклама	радио
телевидение	1			
пресса	0.166	1		
наруж.реклама	0.644*	0.550*	1	
радио	0.332	0.434	0.711*	1

Значимые корреляции обозначены звездочкой. Значимость проверяется при помощи статистики

$$T = \frac{r}{1-r^2}\sqrt{n-2}, \tag{3}$$

которая имеет распределение Стьюдента при справедливости гипотезы $H_0 : \varrho = 0$. Для P-значений, соответствующих наблюдаемым значениям статистики Стьюдента T_* или значимостей α_* справедливы выражения: $\alpha_* = P\{|T| > |T_*|\} < 0.05$.

2.2 Анализ выбросов

Выборка считаются однородной, если все ее элементы одинаково распределены. Элемент с распределением, отличным от распределения остальных элементов выборки, называется *выбросом.*

Если одномерная случайная величина η имеет нормальное распределение $N(\mu, \sigma^2)$, где μ и σ среднее и дисперсия, то случайная величина $(\eta-\mu)^2/\sigma^2$ имеет распределение $\chi^2(1)$ хи-квадрат с одной степенью свободы. В многомерном случае можно показать, что если случайный вектор $\eta = (\eta_1, \ldots, \eta_p)^T$ с p компонентами имеет многомерное нормальное распределение $\mathcal{N}(\mu, \mathbf{\Sigma})$ с вектором средних $\mu = (\mu_1, \ldots, \mu_p)^T$ и матрицей ковариаций $\Sigma(p,p)$, то величина

$$\chi^2 = (\eta - \mu)^T \Sigma^{-1}(\eta - \mu)$$

имеет распределение $\chi^2(p)$. Действительно, если рассмотреть случайный вектор $\xi = A^{-1}(\eta - \mu)$, компоненты которого независимы, имеют нулевые математические ожидания и единичные дисперсии, и где матрица A такая, что $AA^T = \Sigma$, тогда

$$\chi^2 = (\eta - \mu)^T \Sigma^{-1}(\eta - \mu) = (\eta - \mu)^T (A^{-1})^T A^{-1}(\eta - \mu) = (A^{-1}(\eta - \mu))^T A^{-1}(\eta - \mu) = \xi^T \xi.$$

Как известно в статистике, сумма $\xi^T\xi = \sum_{i=1}^{p} \xi_i^2$ независимых случайных величин с нулевыми математическими ожиданиями и единичными дисперсиями имеют распределение хи-квадрат с p степенями свободы.

Если μ и Σ известны, то эта статистика может быть использована для проверки возможной аномальности наблюдаемого вектора $x = (x_1, \ldots, x_p)$, то есть наличия выбросов у его компонент. Для этого по реальным данным вычисляется значение статистики

$$\chi_*^2 = (x - \mu)^T \Sigma^{-1}(x - \mu) \tag{4}$$

и соответствующее ему $\alpha^* = \mathbf{P}\{\chi^2 >= \chi_*^2\}$, которое в разных источниках иногда называют $P-$значением. Затем полученное α^* сравниваются с некоторым заранее выбранным уровнем значимости α. Если $\alpha^* < \alpha$, то наблюдаемый вектор можно считать аномальным, и его координаты должны быть проверены на наличие ошибок. Таким образом могут быть проверены все векторы из случайной выборки.

ПРИМЕР. Мониторная система ведения больных, находящихся в критическом состоянии, предполагает ежеминутное измерение систолического и диастолического давлений, частоты дыхания, частоты сердечных сокращений и температуры. Для каждого вектора наблюдений по формуле (4) вычисляется статистика χ_*^2 и соответствующее значение α^*. Параметры μ и Σ известны из наблюдений над здоровыми людьми. При $\alpha^* > 0.2$ считается, что состояние больного находится в пределах нормы; если $0.05 \le \alpha^* \le 0.2$, то имеются некоторые отклонения от нормы; если $0.01 \le \alpha^* \le 0.05$, то зажигается световой сигнал тревоги; если $\alpha^* \ge 0.01$, то включается сирена. Это

помогает лечащему врачу и персоналу клиники определять изменения в состоянии больного или предупреждает о возможных неисправностях в оборудовании системы.

Если параметры распределения не известны, то используются их выборочные аналоги. Выборочным аналогом статистики χ^2 является выборочное расстояние Махаланобиса:

$$D^2 = (X - \bar{X})^T \mathbf{S}^{-1} (X - \bar{X}).$$

Можно показать, что величина

$$F = \frac{(n-p)n}{(n^2-1)p} D^2$$

имеет F-распределение с p и $n-p$ степенями свободы.

Исследуем на наличие выбросов данные о смертности в разных странах (табл.1). Проверим, являются ли выбросами 1) Россия и 2) Япония.

1) Вычислим по выборке из 18 стран-индивидов, исключив Россию, вектор средних $\bar{X} = (360.1, 191.4, 8.7, 56.2)^T$ и ковариационную матрицу

$$\mathbf{S} = \begin{bmatrix} 24127.53 & 523.14 & 82.074 & 1834.79 \\ 523.14 & 1037.44 & -111.84 & 60.89 \\ 82.074 & -111.84 & 49.18 & 60.09 \\ 1834.79 & 60.89 & 60.09 & 375.73 \end{bmatrix}.$$

Обратная матрица имеет вид:

$$\mathbf{S^{-1}} = \begin{bmatrix} 7.27 \cdot 10^{-5} & 4.71 \cdot 10^{-5} & 0.000533 & -0.00045 \\ 4.71 \cdot 10^{-5} & 0.001559 & 0.00504 & -0.00129 \\ 0.000533 & 0.00504 & 0.0436 & -0.0104 \\ -0.00045 & -0.00129 & -0.0104 & 0.0067 \end{bmatrix}.$$

Расстояние Махаланобиса между Россией и всей остальной выборкой равно 58.70. Значение статистики Фишера равно 11.45, значимость которой равна 0.000245. Значимость меньше 0.05, следовательно, гипотеза относительно принадлежности России всей выборке стран отвергается.

2) Вычислим по выборке из 18 стран-индивидов, исключив Японию, вектор средних $\bar{X} = (387.5, 193.2, 9.21, 63.4)^T$ и ковариационную матрицу

$$\mathbf{S} = \begin{bmatrix} 28952.94 & 258.13 & 320.80 & 3992.77 \\ 258.13 & 984.83 & -108.30 & 47.46 \\ 320.80 & -108.30 & 56.86 & 140.81 \\ 3992.77 & 47.46 & 140.81 & 1176.38 \end{bmatrix}.$$

Обратная матрица имеет вид:

$$\mathbf{S^{-1}} = \begin{bmatrix} 6.85 \cdot 10^{-5} & 3.65 \cdot 10^{-5} & 0.000373 & -0.00028 \\ 3.65 \cdot 10^{-5} & 0.001539 & 0.004527 & -0.00073 \\ 0.000373 & 0.004527 & 0.0393 & -0.00616 \\ -0.00028 & -0.00073 & -0.00616 & 0.002562 \end{bmatrix}.$$

Расстояние Махаланобиса между Японией и всей остальной выборкой равно 2.586. Значение статистики Фишера равно 0.504, значимость которой равна 0.733. Значимость больше 0.05, следовательно, гипотеза относительно принадлежности Японии всей выборке стран не отвергается.

2.3 Линейное преобразование, приводящее к некоррелированным компонентам

Пусть имеется нормированный и центрированный вектор $u = (u_1, u_2)^T$ ($Eu_1 = Eu_2 = 0, Du_1 = Du_2 = 1$) с ковариационной матрицей $\Sigma = \begin{bmatrix} 1 & \varrho \\ \varrho & 1 \end{bmatrix}$, которая ввиду нормированности вектора u является корреляционной матрицей. Нашей задачей является нахождение линейного преобразования $y = Bu$ с некоррелированными компонентами y_1 и y_2. Вычислим собственные числа и собственные вектора ковариационной матрицы Σ. По определению собственные числа λ и собственные вектора $\alpha = (\alpha_1, \alpha_2)^T$ являются решением матричного уравнения

$$(\Sigma - \lambda I)\alpha = \mathbf{0}. \tag{5}$$

Для нахождения собственных чисел используется характеристическое уравнение вида

$$|\Sigma - \lambda I| = \begin{vmatrix} 1-\lambda & \varrho \\ \varrho & 1-\lambda \end{vmatrix} = (1-\lambda)^2 - \varrho^2 = 0.$$

Корнями этого уравнения являются числа $\lambda_1 = 1+\varrho, \lambda_2 = 1-\varrho$. Чтобы найти собственный вектор, соответствующий собственному числу $1+\varrho$, подставим это значение в уравнение (5). Получим систему однородных уравнений

$$\left\{ \begin{array}{c} -\alpha_1\varrho + \alpha_2\varrho = 0 \\ \alpha_1\varrho - \alpha_2\varrho = 0 \end{array} \right. ,$$

решение которой определяется с точностью до умножения на константу $\alpha_1 = \alpha_2 = t$. Нормируем этот вектор, получаем $\alpha_1^2 + \alpha_2^2 = 1$ эквивалентно $2t^2 = 1$, следовательно, $\alpha = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$. Аналогично получает, что собственный вектор, соответствующий собственному числу $\lambda_2 = 1-\varrho$, имеет вид $\alpha = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}$.

Матрица B образуется из элементов собственных векторов α_1 и α_2 и имеет вид

$$B = [\alpha_1|\alpha_2] \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}.$$

Таким образом имеется линейное преобразование $y = Bu$ вида

$$\left\{ \begin{array}{l} y_1 = \frac{1}{\sqrt{2}}u_1 + \frac{1}{\sqrt{2}}u_2 \\ y_2 = \frac{1}{\sqrt{2}}u_1 - \frac{1}{\sqrt{2}}u_2 \end{array} \right.$$

Новые случайные величины y_1 и y_2 называются *главными компонентами*. Как и u_1, u_2, главные компоненты имеют нулевые математические ожидания. Их дисперсии совпадают со значениями собственных чисел $1+\varrho$ и $1-\varrho$. Действительно, $Dy = Eyy^T = E(Bu)(Bu)^T = B\ Euu^T\ B^T = B\Sigma B^T =$

$$\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 1 & \varrho \\ \varrho & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{\sqrt{2}}(1+\varrho) & \frac{1}{\sqrt{2}}(1+\varrho) \\ \frac{1}{\sqrt{2}}(1-\varrho) & -\frac{1}{\sqrt{2}}(1-\varrho) \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}$$

$$= \begin{pmatrix} 1+\varrho & 0 \\ 0 & 1-\varrho \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}.$$

2.4 Факторный анализ

Идея факторного анализа состоит в том, чтобы заменить k-мерную случайную величину на $m < k$ линейных функций, теряя не слишком много информации. Как наилучшим образом выбрать эти m линейных функций? Эффективность любого выбора зависит от того, в какой степени эти m линейных функций дают возможность реконструировать k первоначальных величин. Один из методов реконструкции этой первоначальной случайной величины состоит в построении ее наилучшего предиктора на основе m линейных функций. Оказывается, что наилучшим выбором линейных функций является выбор первых m главных компонент.

Определим понятие главной компоненты. Пусть $U = (u_1, \ldots, u_k)^T$ – k-мерная случайная величина с $EU = 0$ и дисперсионной матрицей $\mathbf{\Sigma}$. Обозначим через $\lambda_1 \geq \ldots \geq \lambda_k$ собственные числа и $P_1, \ldots, P_k$ соответствующие собственные вектора ковариационной матрицы $\mathbf{\Sigma}$,

$$P_1 = \begin{bmatrix} p_{11} \\ \ldots \\ p_{k1} \end{bmatrix}, \ldots, P_k = \begin{bmatrix} p_{1k} \\ \ldots \\ p_{kk} \end{bmatrix}.$$

Это означает, что

$$\mathbf{\Sigma}P_i = \lambda_i P_i, \quad i = 1, \ldots, k. \tag{6}$$

Случайная величина $y_i = P_i^T U$ называется i-ой главной компонентой. Обозначим через $\mathbf{Y} = (y_1, \ldots, y_p)$ новый вектор случайных величин, $\mathbf{Y} = \mathbf{P}^T U$, где $\mathbf{P}$ матрица, составленная из собственных векторов

$$\mathbf{P} = (P_1|\ldots|P_p) = \begin{bmatrix} p_{11} & \ldots & p_{1k} \\ \ldots & \ldots & \ldots \\ p_{k1} & \ldots & p_{kk} \end{bmatrix}.$$

2.5 Свойства главных компонент. Факторные нагрузки. Интерпретация факторов

1. *Главные компоненты не коррелированы. Дисперсия i-ой главной компоненты $y_i = P_i^T U$ равна λ_i.*
Действительно, по (6)

$$Dy_i = Ey_i y_i = EP_i^T UU^T P_i = P_i^T \mathbf{\Sigma} P_i = P_i^T \lambda_i P_i = \lambda_i P_i^T P_i = \lambda_i,$$
$$Ey_i y_j = Cov(P_i^T U, P_j^T U) = EP_i^T UU^T P_j = P_i^T \mathbf{\Sigma} P_j = P_i^T \lambda_j P_j = \lambda_j P_i^T P_j = 0.$$

Здесь использовалось свойство ортогональности собственных векторов симметричных матриц, т.е. $P_i^T P_i = 1$, $P_i^T P_j = 0$. Таким образом, $Y = P^T U$ переводит коррелированное множество случайных величин в некоррелированное.

2. Нормированные главные компоненты $G_i = \lambda_i^{-1/2} y_i = \lambda_i^{-1/2} P_i^T U, i = 1, \ldots, r$, где r ранг дисперсионной матрицы, для $\lambda_i \neq 0$ называются *факторами.*

3. Корреляция между компонентой u_i и j-ой главной компонентой y_j называется факторной нагрузкой и определяется как

$$cor(u_i, y_j) = \frac{p_{ij}\sqrt{\lambda_j}}{\sigma_i}, \tag{7}$$

где p_{ij} – i-ый элемент j-го собственного вектора, σ_i – стандартное отклонение i-ой компоненты вектора U.

Так как $Y = P^T U$, то ввиду ортогональности матрицы собственных векторов P обратная матрица $P^{-1} = P^T$ и $U = PY$.

$$PY = (P_1|\ldots|P_p)Y = \begin{bmatrix} p_{11} & \ldots & p_{1k} \\ & \ldots & \\ p_{k1} & \ldots & p_{kk} \end{bmatrix} \begin{bmatrix} y_1 \\ \ldots \\ y_k \end{bmatrix} = \begin{bmatrix} u_1 \\ \ldots \\ u_k \end{bmatrix}$$

Следовательно, $\sum_{l=1}^{k} p_{il} y_l = u_i$. Это формула обращения. Отсюда

$$cov(u_i, y_j) = E \sum_{l=1}^{k} p_{il} y_l y_j = \sum_{l=1}^{k} p_{il} Ey_l y_j = \lambda_j p_{ij},$$

так как $Ey_l y_j = \lambda_j$ при $l = j$, иначе равно нулю. Используя то, что $Du_i = \sigma_i^2$, $Dy_j = \lambda_j$, получаем (7).

Для примера в качестве исходных рассмотрим данные смертности в разных странах (табл.1). Матрица факторных нагрузок имеет вид:

	Факторные нагрузки	
	1-го фактора	2-го фактора
См-ть от забол.сист.кров.	-0,738	-0,497
См-ть от злок.нов.	0,303	-0,867
См-ть от инф.забол.	-0,767	0,496
См-ть от несч.сл.	-0,885	-0,313
Дисперсия гл.комп.	2,007	1,343
Доля дисп.гл.комп.	0,502	0,336

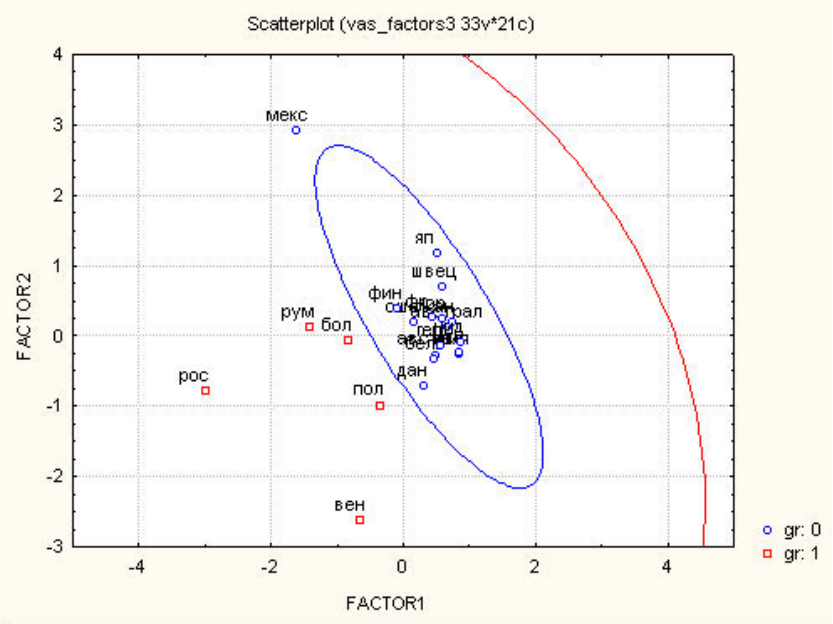

Рис. 5: Двумерная диаграмма первых двух факторов.

	Значения факторов	
	1 фактор	2 фактор
Россия	-2,78737	-0,97021
Австралия	0,69045	0,24838
австрия	0,51875	-0,17004
Бельгия	0,52804	-0,19983
Болгария	-0,77484	-0,10548
Англия	0,87657	-0,09888
Венгрия	-0,38952	-2,50084
Германия	0,59220	-0,01593
Дания	0,42262	-0,58187
Италия	0,86187	-0,08348
Канада	0,74328	0,31192
Мексика	-1,65770	2,73022
Нидерланды	0,89049	0,04762
Польша	-0,21647	-0,92525
Румыния	-1,33794	0,03267
США	0,19411	0,27568
Финляндия	-0,08943	0,41200
Франция	0,45648	0,36001
Япония	0,47843	1,23332

Отсюда видно, что первая главная компонента описывает почти 50,2% информации, а вторая 33,6%. Первая главная компонента ("успешное здравохранение") имеет выраженные отрицательные корреляции с показателями смертности от заболеваний системы кровообращения, инфекционных заболеваний и несчастных случаев, т.е.отражает внешнюю благоприятную среду, связанную в успешной работой министерства здравохранения: чем больше первая главная компонента, тем ниже указанные формы смертности. Вторая главная компонента имеет выраженную отрицательную корреляцию со смертностью от злокачественных новообразований и положительную со смертностью от инфекционных заболеваний: чем больше вторая главная компонента, тем меньше смертность от онкологических заболеваний и больше от инфекционных. Вторая главная компонента может быть интерпретирована как "бактериальный"фактор, с одной стороны, увеличивающий смертность от инфекционных заболеваний, а с другой стороны, не позволяющий увеличивать смертность от рака.

Отрицательные значения факторов свидетельствуют о неблагополучной обстановке в стране. Самой неблагополучной страной во всех отношениях является Россия, так как отрицательны оба фактора. Самой благополучной страной с точки зрения "бактериального"фактора, является Мексика. Противоположное место занимает Венгрия. Самой благополучной страной является Япония.

3 Задачи классификации и прогноза

3.1 Байесовская процедура классификации

Представим данные о смертности в 19 странах (табл.1) в виде двух популяций: к первой популяции W_1 отнесем страны, никогда не строившие социализм, а ко второй Россию, Болгарию, Венгрию, Румынию и Польшу. Поскольку Германия состоит из двух половин, одна из которых строила социализм, а другая нет, то не понятно, к какой популяции стоит ее отнести. Решение этой задачи можно доверить программе дискриминантного анализа, который на основе данных о популяциях разрабатывает процедуру классификации.

Пусть $X = (x_1, \ldots, x_p)^T$ некоторый вектор наблюдений p признаков. Задача состоит в том, чтобы отнести его к одной из k популяций $W_1, \ldots, W_k$. Обозначим через

$q_i = P(W_i)$ – априорную вероятность того, что случайный вектор $X \in W_i$,

$f_i(X)$ – плотность распределения популяции W_i,

P_{ij} – вероятность отнесения к популяции W_i элемента из популяции W_j.

Будем считать, что все параметры известны. Апостериорная вероятность $P(W_j|X)$ выражается по формуле Байеса

$$P(W_j|X) = \frac{q_j f_j(X)}{\sum\limits_{i=1}^{k} q_i f_i(X)} \tag{8}$$

и означает условную вероятность того, что вектор наблюдений относится к популяции W_j. Байесовская процедура относит наблюдение X к W_i в случае, когда апостериорная вероятность $P(W_j|X)$ максимальна. Поскольку в выражениях для апостриорных вероятностей знаменатель один и тот же, то максимальным должен быть только числитель вида $q_j f_j(X)$,который в случае Байесовской процедуры классификации называют дискриминантной функцией.

Если распределения популяций являются нормальными $N(\mu_\mathbf{i}, \mathbf{\Sigma})$ с одной и той же ковариационной матрицей $\mathbf{\Sigma}$, то дискриминантная функция имеет вид:

$$\delta_i = {\mu_\mathbf{i}}^T \mathbf{\Sigma}^{-1} X - \frac{1}{2} {\mu_\mathbf{i}}^T \mathbf{\Sigma}^{-1} \mu_\mathbf{i} + \ln q_i = \alpha_i^T X + \gamma_i + \ln q_i,$$

где $\alpha_i = \mathbf{\Sigma}^{-1}\mu_\mathbf{i}$, $\gamma_i = -\frac{1}{2}{\mu_\mathbf{i}}^T \mathbf{\Sigma}^{-1}\mu_\mathbf{i}$. Вектор X относится к той популяции, для которой значение дискриминантной функции максимально.

В случае неизвестных параметров μ и $\mathbf{\Sigma}$ можно воспользоваться их выборочными оценками. Пусть n_i объем выборки, относящейся к популяции W_i, $\bar{X}_i$ вектор средних используется в качестве оценки $\mu_\mathbf{i}$, S_i – оценка ковариационной матрицы для популяции W_i. В качестве оценки общей ковариационной матрицы берется оценка

$$S = \frac{\sum\limits_{i=1}^{k} (n_i - 1) S_i}{\sum\limits_{i=1}^{k} n_i - k}.$$

Тогда в качестве оценки дискриминантных функций рассматриваются функции

$$d_i = a_i^T X + g_i + \ln q_i = a_i^T X + c_i, \tag{9}$$

где $a_i = S^{-1}\bar{X}_i$, $g_i = -\frac{1}{2} X_i^T S^{-1} X_i$, $c_i = g_i + \ln q_i$. Вектор $X \in W_i$, если d_i максимально.

Построим процедуру классификации стран на $k = 2$ популяции по четырем признакам (табл.1). Объем первой выборки W_1 равен $n_1 = 13$, Объем второй W_2 равен $n_2 = 5$, $n = n_1 + n_2 = 13 + 5 = 18$. Оценим априорные вероятности: $\hat{q}_1 = n_1/n = 13/18$, $\hat{q}_2 = n_2/n = 5/18$. Программа дискриминантного анализа выдает следующие коэффициенты дискриминантной функции.

	коэффициенты	W_1	W_2
см-ть от забол.сист.кр.	a_{i1}	0.1089	0.204
см-ть от зл.новообр.	a_{i2}	0.3463	0.429
см-ть от инф.забол.	a_{i3}	1.0321	1.161
см-ть от несч.сл.	a_{i4}	−0.1151	−0.121
	c_i	−50.2567	−113.593

Вектор наблюдений, соответствующий Германии, не участвующей в построении дискриминантной функции, имеет вид: $X = (343.8, 196.7, 5.5, 41.2)^T$. Получим значение дискриминантной функции (9) в случае принадлежности

вектора X популяции W_1:

$$a^T X + c = [0.1089, 0.3463, 1.0321, -0.0051] \begin{bmatrix} 343.8 \\ 196.7 \\ 5.5 \\ 41.2 \end{bmatrix} - 50.2567 = 56.2243$$

и в случае принадлежности вектора X популяции W_2:

$$a^T X + c = [0.204, 0.429, 1.161, -0.121] \begin{bmatrix} 343.8 \\ 196.7 \\ 5.5 \\ 41.2 \end{bmatrix} - 113.593 = 42.3965.$$

Значение дискриминантной функции случае принадлежности вектора X популяции W_1 выше, следовательно, по показателям смертности Германия относится к странам, никогда не строившим социализм.

Указанную процедуру можно использовать для классификации всех стран, участвующих в построении этой процедуры. Результаты этой проверки обычно заносятся в таблицу, которая называется матрицей ошибочной классификации.

	относится к W_1	относится к W_2	n_i
принадлежит W_1	13	0	18
принадлежит W_2	0	5	5
Всего	13	5	18

В нашем примере процедура классификации безукоризнена, все элементы из популяции W_i к ней и относятся в результате работы программы дискриминантного анализа.

3.2 Классификация в случае биномиальных распределений

Пусть исследуется задача отнесения объекта к одной из k популяций с биномиальными распределениями, то есть имеются данные, значения которых состоят только из нулей и единиц (например, ответы "да"и "нет"на ряд вопросов или наличие или отсутствие некоторых симптомов). В общем случае требуется классифицировать объект на основе наличия ("да") или отсутствия ("нет") p событий. Определим для каждого i-го события $j = 1, \ldots, p$, случайную величину

$$X_j = \begin{cases} 1, & \text{'да'}, \\ 0, & \text{'нет'} \end{cases}$$

Положим $P(X_j = 1|W_i) = p_{ij}$ и $P(X_j = 0|W_i) = 1 - p_{ij}$ для $i = 1, \ldots, k$ популяций. Тогда закон распределения X_j для популяции W_i при значениях p признаков $x_j = 0$ или $x_j = 1$ имеет вид

$$f_i(x_j) = p_{ij}^{x_j}(1 - p_{ij})^{1-x_j} = \begin{cases} p_{ij}, & \text{при} \quad x_j = 1 \\ 1 - p_{ij}, & \text{при} \quad x_j = 0 \end{cases}.$$

Рассмотрим, например, популяцию W_1 (13 стран, никогда не строивших социализм, $n_1 = 13$) и популяцию W_2 (5 стран, некогда строивших социализм, $n_2 = 5$, $n_1 + n_2 = n = 18$). Число популяций $k = 2$, число признаков $p = 4$ (X_1 – низкая (0) или высокая (1) смертность от заболеваний системы кровообращения, X_2 – низкая (0) или высокая (1) смертность от злокачественных новообразований, X_3 – низкая (0) или высокая (1) смертность от инфекционных заболеваний, X_4 – низкая (0) или высокая (1) смертность от несчастных случаев и других внешних воздействий.

Обозначим через n_{ij} $(i = 1, \ldots, k;\ j = 1, \ldots, p)$ количество наблюдений из популяции W_i со значениями j-го признака, равными 1. Например, в популяции W_2 во всех странах высокая смертность от заболеваний системы кровообращения, т.е. у всех индивидов все значения признака X_1 равны 1, поэтому $n_{21} = 5$ и оценка вероятности высокой смертности от заболеваний системы кровообращения в странах, когда-то строивших социализм, равна $\hat{p}_{21} = n_{21}/n_2 = 5/5 = 1$. В популяции W_1 в девяти странах ($n_{12} = 9$) из тринадцати ($n_1 = 13$) имеем значения второго признака, равные 1, и оценка вероятности p_{12} высокой смертности от злокачественных новообразований в странах, никогда не строивших социализм, равна $\hat{p}_{12} = n_{12}/n_1 = 9/13 = 0.692$.

n_{ij}	X_1	X_2	X_3	X_4
W_1	5	9	3	7
W_2	5	3	2	5

$\hat{p}_{ij}$	X_1	X_2	X_3	X_4
$W1$	0, 385	0, 692	0, 230	0.538
$W2$	1	0.6	0.4	1

Если предположить независимость признаков $X_1, \ldots, X_p$, то совместный закон распределения

$$f_i(x_1, \ldots, x_p) = f_i(x_1) \cdot \ldots \cdot f_i(x_p), \quad i = 1, \ldots, k.$$

Полагая, что известны априорные вероятности $q_1, \ldots, q_k$ и одинаковы стоимости ошибочной классификации, вычислим апостериорные вероятности:

$$P(W_i|x_1, \ldots, x_p) = \frac{q_i f_i(x_1, \ldots, x_p)}{\sum\limits_{m=1}^{k} q_m f_m(x_1, \ldots, x_p)}.$$

Случайный вектор X со значениями $(x_1, \ldots, x_p)^T$ относится к популяции W_i, для которой величина апостериорной вероятности максимальна.

Выясним, к какой популяции W_1 или W_2 будет отнесена Германия со значениями признаков $X = (1100)^T$ результате этой процедуры классификации. В качестве оценок априорных вероятностей рассмотрим $\hat{q}_1 = n_1/n = 13/18 = 0.722$ и $\hat{q}_2 = n_2/n = 5/18 = 0.278$. Вероятность получить значение вектора $X = (1100)^T$ при условии популяции W_1 равна

$$\begin{aligned} \hat{f}_1(1,1,0,0) = \hat{f}_{11}(1) \cdot \hat{f}_{12}(1) \cdot \hat{f}_{13}(0) \cdot \hat{f}_{14}(0) = \\ = \hat{p}_{11}\hat{p}_{12}(1-\hat{p}_{13})(1-\hat{p}_{14}) = \\ = 0,385 \cdot 0,692 \cdot (1-0,230) \cdot (1-0.538) = 0.0945. \end{aligned}$$

Отсюда получаем вероятность получить значение вектора $X = (1100)^T$ из популяции W_1, равную

$$\hat{q}_1\hat{f}_1(1,1,0,0) = 0.722 \cdot 0.0945 = 0.0683.$$

Вероятность получить значение вектора $X = (1100)^T$ при условии популяции W_2 равна

$$\begin{aligned} \hat{f}_2(1,1,0,0) = \hat{f}_{21}(1) \cdot \hat{f}_{22}(1) \cdot \hat{f}_{23}(0) \cdot \hat{f}_{24}(0) = \\ = \hat{p}_{21}\hat{p}_{22}(1-\hat{p}_{23})(1-\hat{p}_{24}) = 1 \cdot 0.6 \cdot (1-0.4) \cdot (1-1) = 0. \end{aligned}$$

Вероятность получить значение вектора $X = (1100)^T$ из популяции W_2 равна

$$\hat{q}_2\hat{f}_2(1,1,0,0) = 0.278 \cdot 0 = 0.$$

Оценки апостериорных вероятностей имеют вид:

$$P(W_1|1,1,0,0) = \frac{\hat{q}_1\hat{f}_1(1,1,0,0)}{\hat{q}_1\hat{f}_1(1,1,0,0) + \hat{q}_2\hat{f}_2(1,1,0,0)} = \frac{0.0683}{0.0683+0} = 1,$$

$$P(W_2|1,1,0,0) = \frac{\hat{q}_2\hat{f}_2(1,1,0,0)}{\hat{q}_1\hat{f}_1(1,1,0,0) + \hat{q}_2\hat{f}_2(1,1,0,0)} = \frac{0}{0.0683+0} = 0.$$

Наибольшая апостериорная вероятность достигается в случае популяции W_1, т.е. как и в случае метрических признаков Германия относится к популяции, никогда не строившей социализм.

Поскольку в большинстве случаев предположение о независимости не выполняется, то когда все априорные вероятности равны, для классификации используется другая процедура, которая для любой комбинации значений признаков вычисляет соответствующую долю объектов в каждой выборке. Новый объект, задаваемый комбинацией значений признаков, относится к популяции, выборка из которой имеет наибольшую долю объектов с такой комбинацией.

Пусть, например, $k = 2, p = 4$. По двум выборкам объема $n_1 = 13$ и $n_2 = 5$ составляется следующая таблица вероятностей появления набора признаков:

Код	значения	частота кода		вероятности кода	
k	признаков	в W_1	в W_2	в W_1	в W_2
2	0100	4	0	1	0
3	1100	1	0	1	0
4	0010	1	0	1	0
9	1001	2	1	0.667	0.333
10	0101	2	0	1	0
11	1101	1	2	0.333	0.667
12	0011	1	0	1	0
13	1011	0	1	0	1
15	1111	1	1	0.5	0.5

Германия с комбинацией признаков $X = (1100)^T$ будет отнесена к популяции W_1, так как вероятность такого набора значений признаков в популяции W_1 равна 1, а в популяции W_2 равна 0.

3.3 Множественная регрессия и задачи прогнозирования

При статистическом анализе данных часто бывает необходимо исследовать влияние на некоторый признак (его называют "зависимой"переменной) ряда других признаков (их называют "независимые"переменные). Если это влияние существенно, то появляется возможность спрогнозировать поведение "зависимой"переменной при известных значениях "независимых"переменных.

Математическая постановка задачи заключается в следующем: пусть имеются p "независимых"переменных $X_1, \ldots, X_p$ и "зависимая"переменная $Y = X_0$. Матрица наблюдений имеет вид $\{x_{ij}\}$, $i = 0, \ldots, p$, $j = 1, \ldots, n$, где x_{ij} – значение i-го признака j-го индивида, n – количество индивидов.

Регрессией называют функцию, являющуюся условным математическим ожиданием

$$E(Y|X_1, \ldots, X_p)$$

переменной Y в зависимости от значений переменных $X_1, \ldots, X_p$. Модель множественной регрессии представляет собой уравнение

$$Y = E(Y|X_1, \ldots, X_p) + \varepsilon,$$

где ε – случайная не систематическая ошибка с нулевым математическим ожиданием и дисперсией σ^2. Для практических целей удобной оказывается линейная аппроксимация множественной регрессии в виде

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon.$$

Коэффициенты $\beta_0, \beta_1, \ldots, \beta_p$ называются частными коэффициентами регрессии.

Итак, нужно найти коэффициенты $\beta_0, \beta_1, \ldots, \beta_p$, по которым наилучшим образом можно предсказать Y, т.е. построить наилучший линейный предиктор $\hat{Y}(X_1, \ldots, X_p)$ вида

$$\hat{Y}(X_1, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

в смысле минимальности математического ожидания квадрата разности отклонения от переменной Y ($E(Y - \hat{Y}(X_1, \ldots, X_p))^2 \to min$.

На практике оценки $b_k = \hat{\beta}_k$ коэффициентов множественной линейной регрессии получают из условия минимизации выражения

$$\Sigma = \sum_{j=1}^{n} (y_j - \beta_0 - \beta_1 x_{1j} - \ldots - \beta_p x_{pj})^2.$$

Разность $Y - \hat{Y}$ называется *остатком*. Важным свойством остатка является его не коррелированность ни с одной из переменных $X_1, \ldots, X_p$, входящих в предиктор $\hat{Y}(X_1, \ldots, X_p)$.

Рассмотрим наилучшие линейные предсказания $\hat{Y}(X_2, \ldots, X_p)$ и $\hat{X}_1(X_2, \ldots, X_p)$ переменных Y и X_1 по переменным $X_2, \ldots, X_p$. Остатки $Y - \hat{Y}(X_2, \ldots, X_p)$ и $X_1 - \hat{X}_1(X_2, \ldots, X_p)$ не коррелированы ни с одной из переменных $X_2, \ldots, X_p$. По сути дела это те же переменные Y и X_1, из которых устранено влияние переменных $X_2, \ldots, X_p$. Коэффициент корреляции ρ_{Yx_1} между остатками $Y - \hat{Y}(X_2, \ldots, X_p)$ и $X_1 - \hat{X}_1(X_2, \ldots, X_p)$ называется *частным коэффициентом корреляции*.

Для того чтобы проверить достоверность прогноза, введем некоторые понятия, связанные с источниками вариации переменной Y. Основным источником дисперсии является сумма квадратов отклонений от среднего значения $\bar{y} = \frac{1}{n} \sum_{j=1}^{n} y_j$ переменной Y вида:

$$SS_T = \sum_{j=1}^{n} (y_j - \bar{y})^2.$$

Можно показать, что SS_T раскладывается на две суммы квадратов $SS_T = SS_R + SS_E$, где

$$SS_R = \sum_{j=1}^{n} (\hat{y}_j - \bar{y})^2$$

является источником вариации, обусловленным регрессией, а

$$SS_E = \sum_{j=1}^{n} (y_j - \hat{y}_j)^2$$

является источником вариации, обусловленным ошибкой. Через $\hat{y}_j$ обозначены значения предиктора $\hat{y}_j = \beta_0 + \beta_1 x_{1j} + \ldots + \beta_p x_{pj}$.

Доля дисперсии, объясняемая регрессией, равная отношению

$$R^2 = \frac{SS_R}{SS_T}$$

называется *коэффициентом детерминации*. Чем ближе коэффициент детерминации к единице, тем ближе точки наблюдений к предиктору, тем более достоверен прогноз. Корень из коэффициента детерминации называется *множественным коэффициентом корреляции*. Он равен корреляции между переменной Y и ее наилучшим предсказанием $\hat{Y}(X_1, \ldots, X_p)$.

В качестве оценки дисперсии σ^2 рассматривают $\hat{\sigma}^2 = \frac{SS_E}{n-p-1}$. Значимость коэффициента детерминации, т.е. гипотезу относительно равенства нулю R^2 определяется при помощи статистики Фишера вида:

$$F = \frac{SS_R/p}{SS_E/(n-p-1)},$$

которая при равенстве $R^2 = 0$ имеет распределение Фишера с p и $n-p-1$ степенью свободы. P-значение или значимость $\alpha_* = P\{F > F_*\}$ указывает, насколько можно доверять прогнозу. Если $\alpha_* < 0.05$, то гипотеза относительно равенства нулю R^2 отвергается , и прогнозу можно доверять с вероятностью, большей 0.95.

Исследуем влияние на продолжительность жизни национального богатства и количеством продаваемой водки.

Года	Продолж-ть жизни Y	Нац.богатсво (млрд.дол.) X_1	Водка (бут.в год) X_2
1970	68.9	2372*	25.3
1975	68.1	2372*	28
1980	67.6	2489	30
1985	69.2	3379	23.5
1990	69.2	4130	18
1995	64.6	1171	38.4
1998	67	689	29.6

В этой таблице звездочками отмечены артефакты или неполные наблюдения. Поскольку данные о национальном богатстве в 1970 и 1975 годах отсутствуют, вместо прочерков указаны средние значения национального богатства. Имеем $n = 7$, $p = 2$.

Построим корреляционную матрицу по трем указанным признакам и таблицу соответствующих значимостей коэффициента корреляции согласно статистике Стьюдента (3).

r	Y	X_1	X_2
Y	1		
X_1	0.72	1	
X_2	-0.93	-0.8	1

α_*	Y	X_1	X_2
Y			
X_1	0.068		
X_2	0.0024	0.0307	

Уравнение множественной линейной регрессии имеет вид:

$$Y = 75.05 - 0.00009x_1 - 0.2557x_2 + \varepsilon, \tag{10}$$

где ε – случайная ошибка с нулевым математическим ожиданием и дисперсией σ^2. Обратим внимание на то, что корреляция между продолжительностью жизни Y и национальным богатством X_1 значима ($\alpha_* = 0.068$) и положительна, т.е. чем больше национальное богатство, тем продолжительность жизни. Однако в уравнении регрессии видно, что следствием увеличения национального богатства является уменьшение продолжительности жизни. Этот парадокс может быть объяснен при помощи частных коэффициентов корреляции.

Рассмотрим предикторы $\hat{Y}(X_2)$ и $\hat{X}_1(x_2)$ переменных Y и X_1 по переменной X_2, т.е. линейные прогнозы продолжительности жизни и национального богатства в зависимости от объемов продаваемой водки:

$$\hat{Y}(X_2) = 74.46 - 0.242X_2,$$
$$\hat{X}_1(X_2) = 6429 - 151.159X_2.$$

Если бы водки не продавалось, то продолжительность жизни равнялась бы 74.46, а национальное богатство равнялось бы 6429.

Составим таблицу, в которой будет отражена динамика продолжительности жизни и национального богатства при условии, если бы водка не продавалась вообще.

Года	Усл.продолж-ть жизни $Y - \hat{Y}(X_2) + 74.46$	Усл.нац.богатсво (млрд.дол.) $X_1 - \hat{X}_1(X_2) + 151.159$
1970	75	6196,3*
1975	74,9	6604,5*
1980	74,9	7023,8
1985	74,9	6931,2
1990	73,6	6850,9
1995	73,9	6975,5
1998	74,2	5163,3

Коэффициент корреляции между $Y - \hat{Y}(X_2) + 74.46$ и $X_1 - \hat{X}_1(X_2) + 151.159$ равен -0.11 ($\alpha_* = 0.4072$). Значимость 0.4072 указывает на отсутствие линейной зависимости между продолжительностью жизни и национальным богатством при условии, если бы объемы продаваемой водки во все годы были одинаковы.

Итак, между продолжительностью жизни и национальным богатством корреляции не обнаруживается. Но при уменьшении национального богатства увеличивается объем продаваемой водки, что, в свою очередь, приводит к уменьшению продолжительности жизни. Так объясняется возникновение корреляции продолжительностью жизни и национальным богатством.

Проверим значимость модели, т.е. проверим гипотезу о равенстве нулю коэффициента детерминации. $SS_T = 16.14,\ SS_R = 14.11,\ SS_E = 2.03,\ R^2 = 14.11/16.14 = 0.874$. Значение статистики Фишера равно

$$F = \frac{14.11/2}{2.03/(7-2-1)} = 13.88.$$

В случае справедливости гипотезы о равенстве $R^2 = 0$ статистика F имеет распределение Фишера с 2 и $7-2-1 = 4$ степенями свободы. Значимость равна $\alpha_* = P\{F > 13.88\} = 0,0159$. Это означает, что с вероятностью, большей 0.9841 можно доверять прогнозу (10). Оценка дисперсии равна $\hat{\sigma}^2 = SS_E/(n-p-1) = 2.03/4 = 0.5075$.