Классификация и кластерный анализ

Н.П.Алексеева

СПбГУ, мат-мех ф-т
для студентов факультета физической культуры и спорта

Задача кластеризации индивидов

- глаза: 1 - большие, круглые, 0 - узкие, мелкие;
- нос: 1 - большой, толстый, 0 - тонкий, маленький;
- рот: 1 - большой, 0 - узкий;
- уши: 1 - большие, музыкальные, 0 - мелкие.

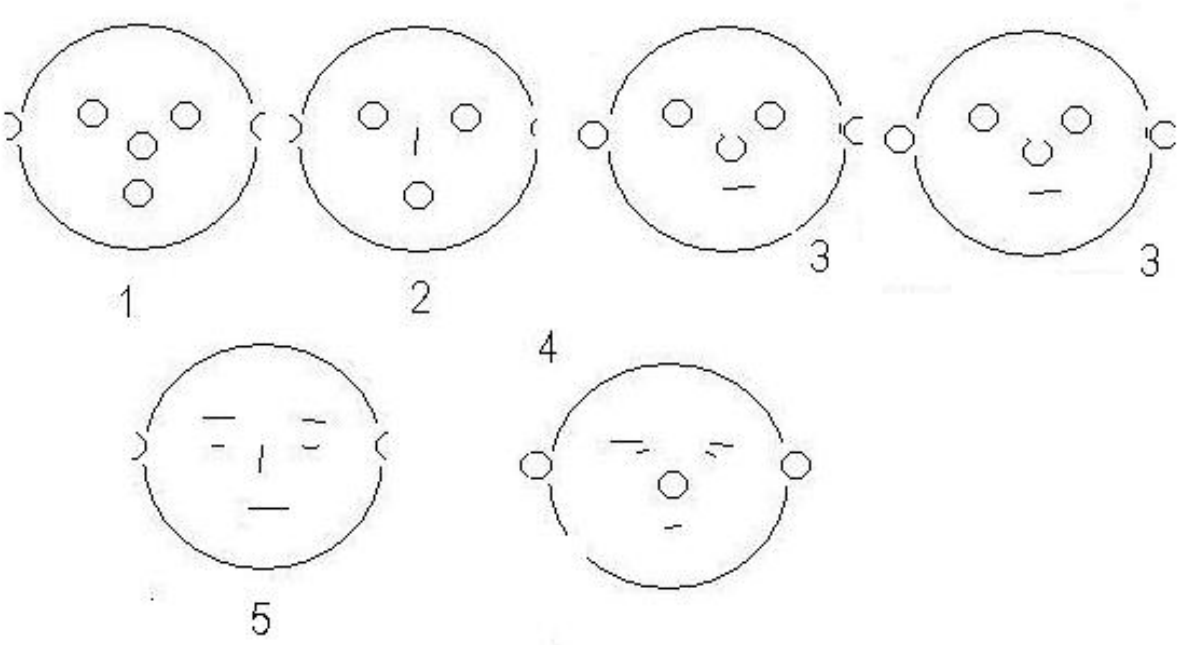


Figure: Данные для кластеризации.

Информационное разнообразие

$n = n_1 + n_2 + \ldots + n_m$ индивидов, отличающихся a_i, $i = 1, \ldots, m$, градациями признака X_1. Информационное разнообразие этой совокупности измеряется формулой:

$$I = I_1 = n \ln n - \sum_{i=1}^{m} n_i \ln n_i.$$

Если признаков несколько, то разнообразия складываются.

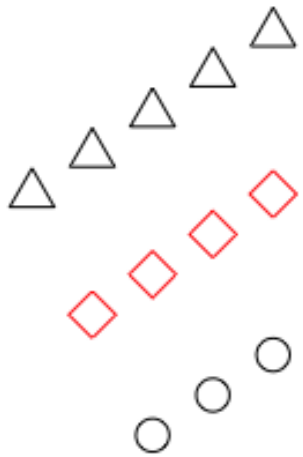
Информационный выигрыш от объединения от совокупностей A и B

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

Кластеризация осуществляется по принципу минимума информационного выигрыша

Примеры

- Разнообразие совокупности, содержащей одинаковые элементы равна нулю. $n \ln n - n \ln n = 0$.
- $I_{form} = 12 \ln 12 - 5 \ln 5 - 4 \ln 4 - 3 \log 3 = 12.93$
  $I_{color} = 12 \ln 12 - 4 \ln 4 - 8 \ln 8 = 7.64$
- Общее разнообразие равно сумме $I_{form} + I_{color} = 20.57$.



Информационный выигрыш от объединения двух одинаковых групп

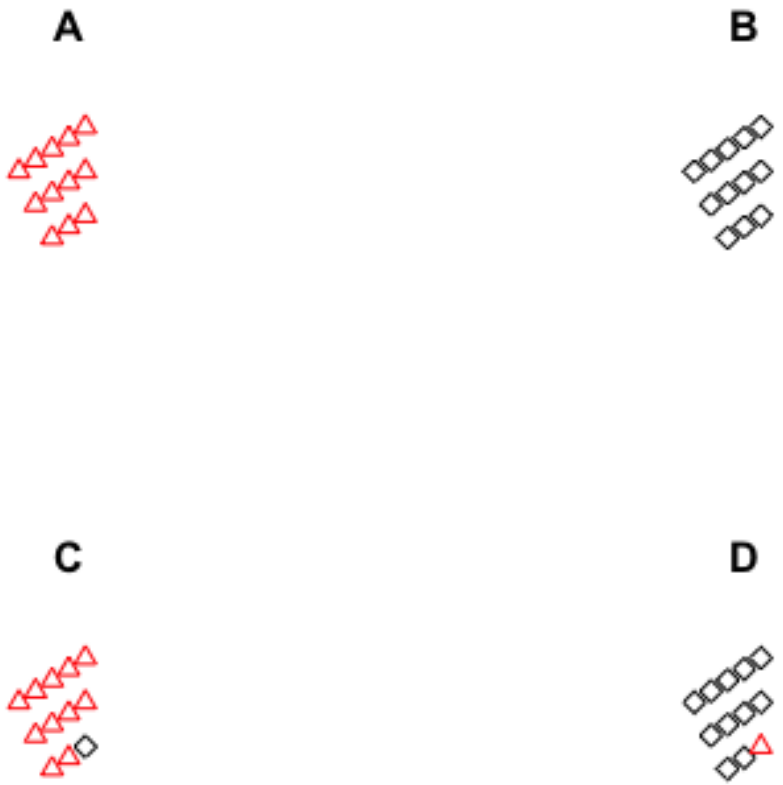
Пусть $I(A) = n \ln n - \sum_{i=1}^{k} n_i \ln n_i$, $I(B) = n \ln n - \sum_{i=1}^{k} n_i \ln n_i$,
$n_1 + \ldots + n_k = n$.

$$I(A+B) = 2n \ln(2n) - \sum_{i=1}^{k} (2n_i) \ln(2n_i) - 2(n \ln n - \sum_{i=1}^{k} n_i \ln n_i) =$$

$$= 2n \ln 2 + 2n \ln n - 2 \underbrace{\sum_{i}^{k} n_i}_{=n} \ln 2 - 2 \sum_{i}^{k} n_i \ln n_i - 2n \ln n + 2 \sum_{i=1}^{k} n_i \ln n_i = 0$$

Максимальный информационный выигрыш

При каком объединении информационный выигрыш максимальный?
1) A-B, 2) A-C, 3) A-D, 4) B-C, 5) B-D, 6) C-D.



Пример кластеризации наблюдений (вычисления)

- $\mathrm{I_A}[12] = \mathrm{I_B}[12] = 0$,
- $\mathrm{I_C}[11, 1] = \mathrm{I_D}[1, 11] = 12 \ln 12 - 11 \ln 11$
- $\mathrm{I_{A+B}}[12, 12] = \mathrm{I_{C+D}}[12, 12] = 24 \ln 24 - 12 \ln 12 - 12 \ln 12$
- $\Delta\mathrm{I_{A+B}}[12, 12] = \mathrm{I_{A+B}}[12, 12] - \mathrm{I_A}[12] - \mathrm{I_B}[12] = 24 \ln 24 - 24 \ln 12$
- $\Delta\mathrm{I_{C+D}}[12, 12] = \mathrm{I_{C+D}}[12, 12] - \mathrm{I_C}[12] - \mathrm{I_D}[12] =$
  $= (24 \ln 24 - 24 \ln 12) - 2(12 \ln 12 - 11 \ln 11) < \Delta\mathrm{I_{A+B}}[12, 12]$

Наибольший выигрыш от объединения при наибольшем отличии подгрупп с учетом их большей внутригрупповой однородности.

Замечание. Это свойство используется при решении обратной задачи разделения совокупности на наиболее отличающиеся однородные подгруппы.

Пример кластеризации наблюдений.

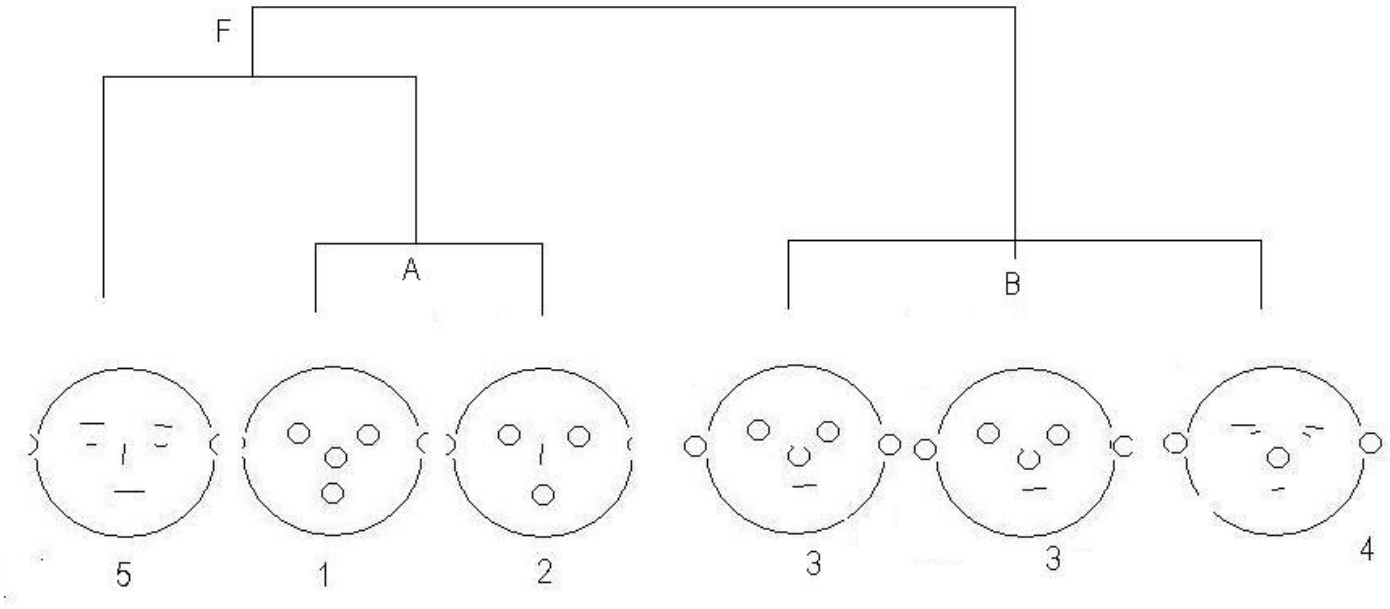


Figure: Дерево кластеризации.

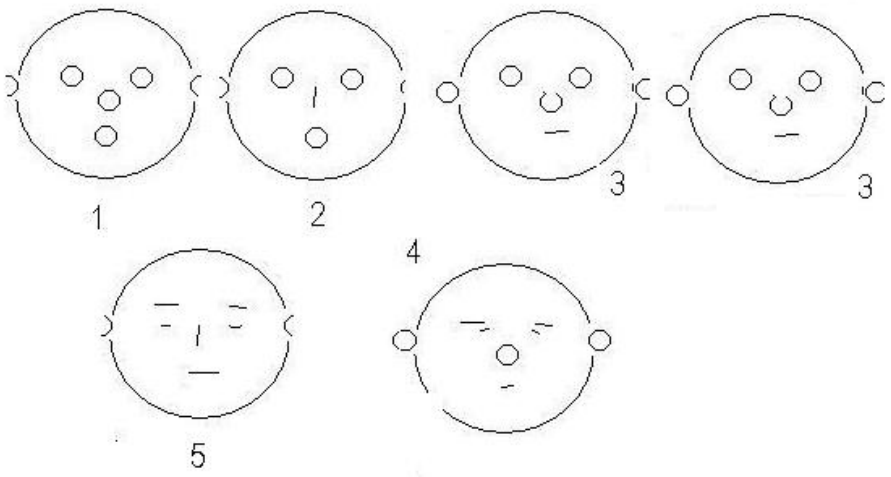
Количество признаков, по которым индивиды отличаются прпарно.

	1	2	3	4	5
1	–	1	2	3	3
2		–	3	4	2
3			–	1	3
4				–	2
5					–

$$I_{A(1,2)} = 2\ln 2 - 1\ln 1 - 1\ln 1 = 1.37,$$

$$I_{B(3,3,4)} = 3\ln 3 - 2\ln 2 - 1\ln 1 = 1.91,$$

$$I_C(5) = 0.$$



Вычисление информационного выигрыша

группа	индивиды	разнообразие	глаза	нос	рот	уши	Сумма
	1		1	1	1	0	
A	2		1	0	1	0	
		I_A	0	1.37	0	0	1.37
C	5		0	0	0	0	
		I_C	0	0	0	0	0
		I_{A+C}	1.91	1.91	1.91	0	5.73

$$\Delta I_{A+C} = I_{A+C} - I_A - I_C = 5.73 - 1.37 - 0 = 4.34\,.$$

группа	индивиды	разнообразие	глаза	нос	рот	уши	Сумма
	3		1	1	0	1	
	3		1	1	0	1	
B	4		0	1	0	1	
		I_B	1.91	0	0	0	1.91
C	5		0	0	0	0	
		I_C	0	0	0	0	0
		I_{B+C}	2.77	2.25	0	2.25	7.27

$$\Delta I_{B+C} = I_{B+C} - I_B - I_C = 7.27 - 1.91 - 0 = 5.36.$$

группа	индивиды	разнообразие	глаза	нос	рот	уши	Сумма
группа	индивиды	разнообразие	глаза	нос	рот	уши	Сумма
	1		1	1	1	0	
A	2		1	0	1	0	
		I_A	0	1.37	0	0	1.37
	3		1	1	0	1	
	3		1	1	0	1	
B	4		0	1	0	1	
		I_B	1.91	0	0	0	1.91
		I_{A+B}	2.5	2.5	3.37	3.37	11.73

$$\Delta I_{A+B} = I_{A+B} - I_A - I_B = 11.73 - 1.37 - 1.91 = 8.44,$$

Деревья по категориальным данным

Наименьший выигрыш достигается от объединения групп A$(1, 2)$ и C(5). Поэтому получаем два кластера: (1, 2, 5) и (3,3,4), которые отличаются ушками.

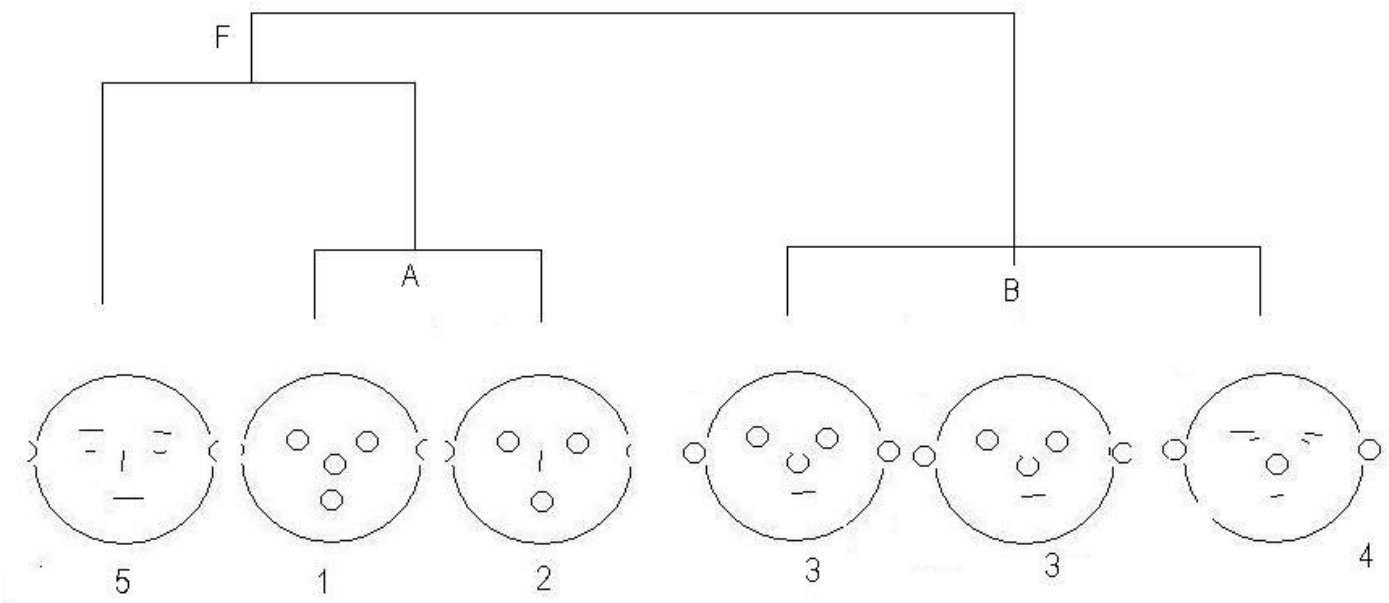


Figure: Дерево кластеризации.

Задание

Вычислить информационные выигрыши от объединения кластеров $\mathrm{F}(1,2,5)$ и $\mathrm{B}(3,3,4)$

$$\Delta \mathrm{I}_{\mathrm{F+D}} = \mathrm{I}_{\mathrm{F+D}} - \mathrm{I}_{\mathrm{F}} - \mathrm{I}_{\mathrm{D}}, \qquad \Delta \mathrm{I}_{\mathrm{B+D}} = \mathrm{I}_{\mathrm{B+D}} - \mathrm{I}_{\mathrm{B}} - \mathrm{I}_{\mathrm{D}},$$

и выяснить, к кому из них примкнет $\mathrm{D}(6)$ индивид (1001) шестого типа.

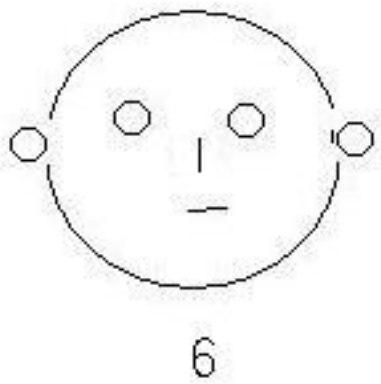


Figure: Индивид (1001) шестого типа.

группа	индивиды	разнообразие	глаза	нос	рот	уши	Сумма
	1		1	1	1	0	
	2		1	0	1	0	
F	5		0	0	0	0	
		I_F					
D	6		1	0	0	1	
		I_D					
		I_{F+D}					

Table: Форма для вычисления разнообразия объединений групп F(1, 2, 5)-D(6), B(3, 3, 4)-D(6).

Исходные данные

Переменные $X_1, X_2, \ldots, X_k$ и Y переменная типа класс.

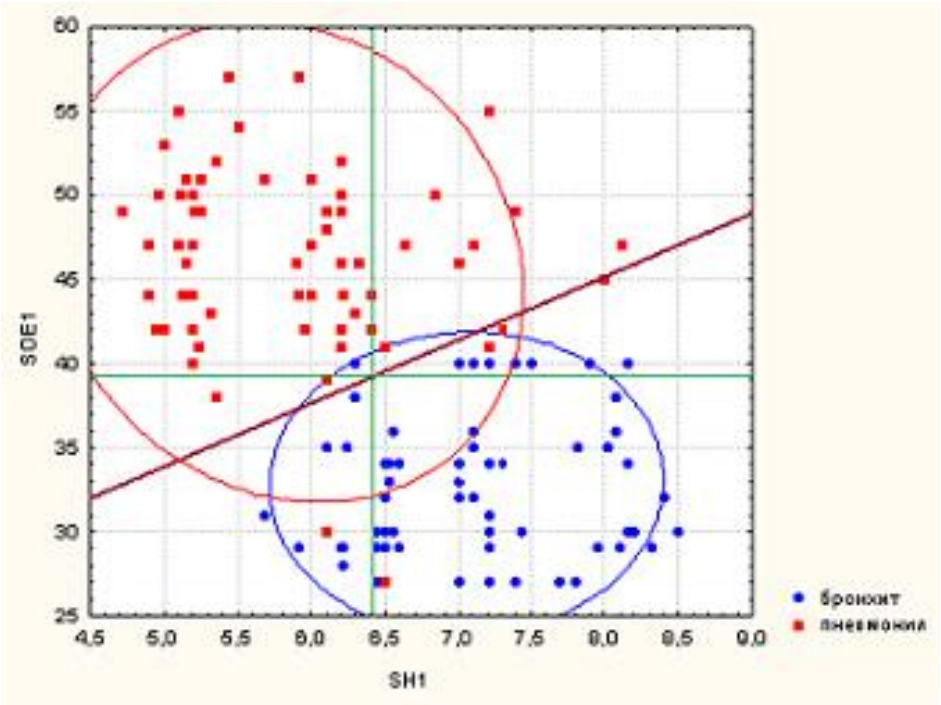


Figure: Классифицирующая прямая $\mathrm{SOE1} = 3.7\mathrm{SH1} + 15.5$.

ПРИМЕР

- W_1 - больные бронхитом, W_2 - больные пневмонией,
- x_1 серосодержащие белки в восстановительной форме (SH1), средние $SH1 = 7.1 \pm 0.7$ в W_1 и $SH1 = 5.8 \pm 0.8$ в W_2
- x_2 скорость оседания эритроцитов (SOE1) в начале заболевания, средние $SOE1 = 32.9 \pm 4.4$ в W_1 и $SOE1 = 46.0 \pm 5.4$ в W_2.

Пусть у некоторого больного $SH1 = 5.5$, $SOE1 = 40$. Можно ли сказать, какая форма заболевания органов дыхания у этого больного?

$3.7SH1 + 15.5 = 3.7 \cdot 5.5 + 15.5 = 35.85 < SOE = 40$, следовательно, у больного пневмония.

Классификация в случае двух популяций W_1 и W_2 по переменным $x_1, \ldots, x_p$

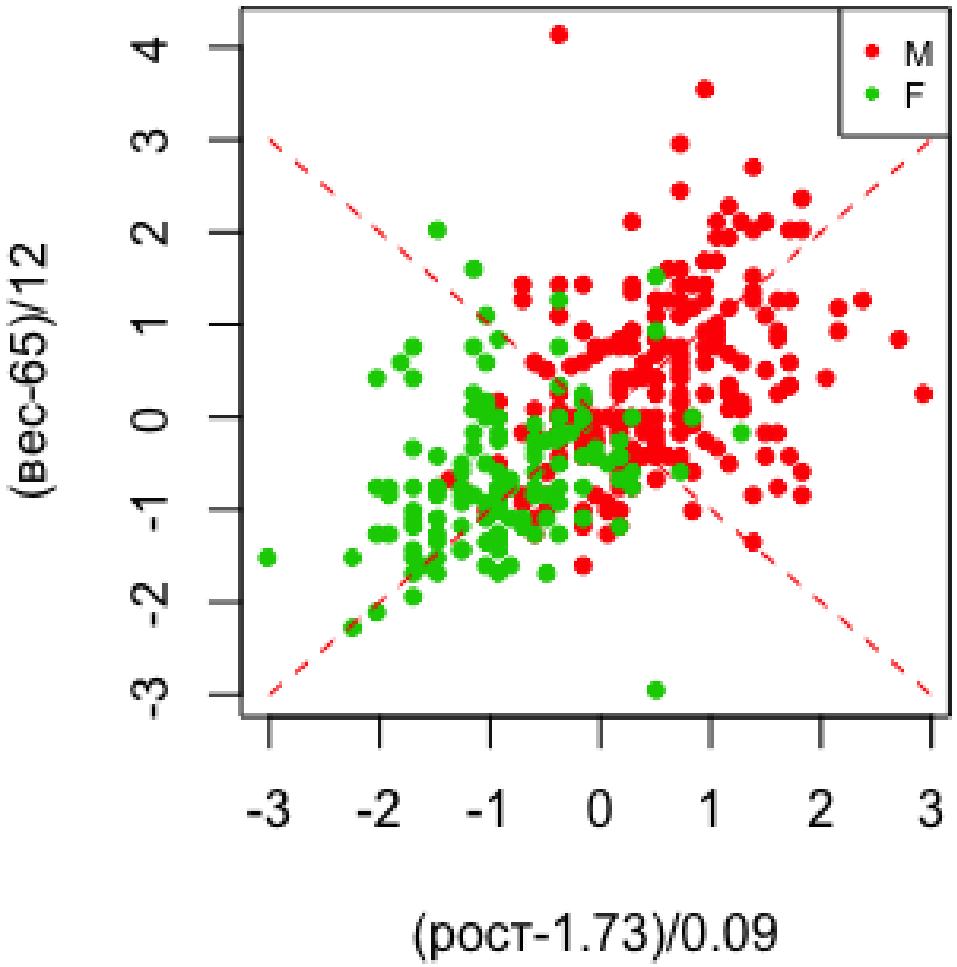
Ищется так называемая дискриминантная функция вида

$$z = \alpha_1 x_1 + \ldots + \alpha_p x_p = \alpha^T X$$

где $\alpha = (\alpha_1, \ldots, \alpha_p)^T$ – некоторый вектор констант, и нужно найти граничное значение c, так что если $z \geq c$, то наблюдение X относится к популяции W_1, и иначе в противоположном случае.

Например, $z = -1.97x_1 + 0.53x_2$, $c = 8.22$, отсюда $x_2 = 40 > 36 = 3.7x_1 + 15.5$ при $x_1 = 5.5$, - это область популяции W_2, следовательно, скорее всего у этого больного не бронхит, а пневмония.

Пример классификации по росту и весу



Вычисление коэффициентов дискриминантной функции

Пусть $X = (x_1, \ldots, x_p)^T$ вектор признаков и популяции W_1 и W_2 характеризуются векторами средних $\mu_1 = (\mu_{11}, \ldots, \mu_{1p})^T$ и $\mu_2 = (\mu_{21}, \ldots, \mu_{2p})^T$ соответственно, и имеют одинаковую ковариационную матрицу $\Sigma = \mathbb{E}(X - \mathbb{E}X)(X - \mathbb{E}X)^T = \{\sigma_{ij}\}_{i=1,j=1}^{p}$.

Оценка параметров

Объемы выборок: $n_1 = 235$(мужчины), $n_2 = 134$(женщины).

$$\Sigma = \frac{(n_1 - 1)\Sigma_1 + (n_2 - 1)\Sigma_2}{n_1 + n_2 - 2}$$

Ковариационные функции

Σ_1	Рост	Вес
Рост	48.3	27.7
Вес	27.7	123.3

Σ_2	Рост	Вес
Рост	41.8	15.4
Вес	15.4	89.1

Σ	Рост	Вес
Рост	46.0	23.3
Вес	23.3	110.9

Средние

	Рост	Вес
μ_1	178	69
μ_2	165	58
μ_1 - μ_2	13	11

Решение системы уравнений методом Крамера

$$\Sigma\alpha = \mu_1 - \mu_2$$

$$\begin{bmatrix} \sigma_{11} & \dots & \sigma_{1p} \\ \vdots & \dots & \vdots \\ \sigma_{p1} & \dots & \sigma_{pp} \end{bmatrix} \cdot \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} \mu_{11} - \mu_{21} \\ \vdots \\ \mu_{1p} - \mu_{2p} \end{bmatrix}$$

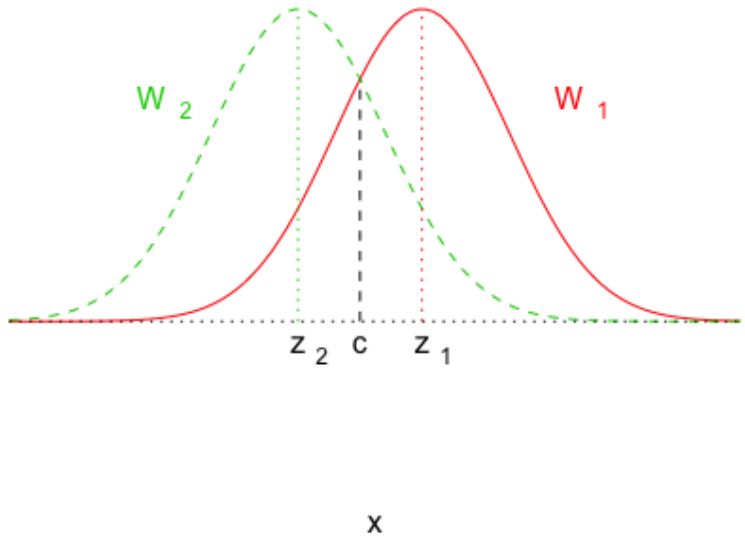
$$\begin{bmatrix} 46 & 23 \\ 23 & 111 \end{bmatrix} \cdot \begin{bmatrix} \alpha_1 \\ \alpha_p \end{bmatrix} = \begin{bmatrix} 178 - 165 \\ 69 - 58 \end{bmatrix} = \begin{bmatrix} 13 \\ 11 \end{bmatrix}$$

$$\Delta = \begin{vmatrix} 46 & 23 \\ 23 & 111 \end{vmatrix} = 4556$$

$$\alpha_1 = \frac{\begin{vmatrix} 13 & 23 \\ 11 & 111 \end{vmatrix}}{\Delta} \approx 0.25, \quad \alpha_2 = \frac{\begin{vmatrix} 46 & 13 \\ 23 & 11 \end{vmatrix}}{\Delta} \approx 0.05\,.$$

Распределение дискриминантной функции

Дискриминантная функция $z(x_1, \ldots, x_n) = \sum_{i=1}^{p} \alpha_i x_i$,
в частности, $z(x_1, x_2) = 0.25x_1 + 0.05x_2$.



Средние дискриминантной функции в популяциях: $z_1 = 48$, $z_2 = 44$, граничное значение: $c = \frac{48+44}{2} = 46$.

Прогнозирование и точность прогноза

Дискриминантная функция $z(x_1, x_2) = 0.25x_1 + 0.05x_2$,
граничное значение $c = 46$.

Задача

К какой популяции относится индивид со значениями $x_1 = 170, x_2 = 69$?

$z(170, 69) = 0.25 \cdot 170 + 0.05 \cdot 69 = 45.95 < 46$, следовательно, наблюдение относим к популяции W_2.

Матрица классификации

	к W_1	к W_2	всего
из W_1	193	42	235
из W_2	22	112	134

Точность прогнозирования ($P = 83\%$) считается как отношение правильно классифицируемых к общему числу наблюдений, умноженное на 100%. При классификации только по росту $P = 83\%$, только по весу ($P = 69\%$).

Классификация в случае двух популяций W_1 и W_2 разного объема

- априорные вероятности $q_1 = P(W_1), q_2 = P(W_2)$ $(q_1 + q_2 = 1)$,
- апостериорные вероятности $P(W_1|X)$ и $P(W_2|X)$ вероятность того, что объект при данном векторе наблюдений X принадлежит к популяции W_1 или W_2
- $P(X|W_i)$, $i = 1, 2$, - условную вероятность получения вектора наблюдений X, если известно, что объект относится к популяции W_i.

Байесовская процедура классификации

если $P(W_1|X) \geq P(W_2|X)$, то объект относится к популяции W_1; если $P(W_1|X) \leq P(W_2|X)$, то объект относится к популяции W_2.

Пусть $\alpha = \Sigma^{-1}(\mu_1 - \mu_2)$, z_1 и z_2 средние значения дискр.ф. $z = \alpha^T X$. Байесовская процедура классификации для нормально распределенных популяций определяется неравенством

$$x^T \alpha \geq \frac{z_1 + z_2}{2} + \ln \frac{q_2}{q_1}.$$

Апостериорные вероятности могут быть вычислены по формуле Байеса

$$P(W_i|X) = \frac{P(W_i)P(X|W_i)}{P(W_1)P(X|W_1) + P(W_2)P(X|W_2)}.$$

Вероятность $P(X|W_i)$ получения наблюдения X может быть заменена на плотности $f_i(x)$:

$$P(W_i|X) = \frac{q_i f_i(x)}{q_1 f_1(x) + q_2 f_2(x)}.$$

Байесовская процедура заключается в следующем: если

$$P(W_1|X) > P(W_2|X) \iff \frac{q_1 f_1(x)}{q_2 f_2(x)} \geq 1,$$

тогда наблюдение X относится к популяции W_1, в противном случае к W_2. В работах Рао (1965) и Андерсон (1958) показано, что эта процедура минимизирует вероятность ошибочной классификации, вычисляемую по формуле полной вероятности,

$$q_1 P(2|1) + q_2 P(1|2).$$

Обобщенная процедура классификации

Пусть $C(1|2)$ стоимость ошибочной классификации при отнесении наблюдения из популяции W_2 к популяции W_1. Соответственно $C(2|1)$ стоимость ошибочной классификации при отнесении наблюдения из популяции W_1 к популяции W_2.

Обобщенная процедура классификации основана на неравенстве

$$x^T\alpha \geq \frac{\zeta_1 + \zeta_2}{2} + \ln\frac{q_2 C(1|2)}{q_1 C(2|1)}.$$

При выполнении этого неравенства объект относится к популяции W_1. Эта процедура минимизирует ожидаемую стоимость ошибочной классификации $q_1 C(2|1)P(2|1) + q_2 C(1|2)P(1|2)$.

Классификация в случае более двух популяций

Задача состоит в том, чтобы отнести вектор наблюдений $x = (x_1, \ldots, x_p)^T$ с ковариационной матрицей Σ к одной из популяций $W_1, \ldots, W_k$ с векторами средних $\mu_1, \ldots, \mu_k$.

- $q_i = P(W_i)$ априорная вероятность того, что $x \in W_i$,
- $f_i(x)$ плотность распределения популяции W_i,
- апостериорная вероятность имеет вид

$$P(W_j|x) = \frac{q_j f_j(x)}{\sum\limits_{i=1}^{k} q_i f_i(x)} = \frac{e^{\delta_i}}{\sum\limits_{j=1}^{k} e^{\delta_j}},$$

где $\delta_i = \alpha_i^T x + \gamma_i + \log q_i$, $\alpha_i = \Sigma^{-1}\mu_i$, $\gamma_i = -\frac{1}{2}\mu_i{}^T\Sigma^{-1}\mu_i$.

Наблюдение относится к той популяции, для которой $P(W_j|x)$ наибольшее.

Вариант 1. Возраст менее 21. Найти коэф. дискр.ф-ии и проверить правильность классификации для пяти наблюдений, c = 45

	Рост	Вес
Рост	85	22
Вес	22	93

	Рост	Вес
Рост	48	30
Вес	30	36

Table: Ков.матр 1 и 2

	Рост	Вес	n
m	182	71	11
f	165	54	19

	Рост	Вес	Пол
1	153	38	2
2	166	46	2
3	165	53	2
4	178	61	1
5	163	55	2

Вариант 2. Возраст от 21 до 28. Найти коэф. дискр.ф-ии и проверить правильность классификации для пяти наблюдений, c = 41

	Рост	Вес
Рост	55	33
Вес	33	100

	Рост	Вес
Рост	52	32
Вес	32	78

Table: Ков.матр 1 и 2

	Рост	Вес	n
m	180	68	45
f	166	56	46

	Рост	Вес	Пол
1	155	40	2
2	170	74	2
3	170	69	2
4	165	75	2
5	181	64	1

Вариант 3. Возраст от 29 до 33. Найти коэф. дискр.ф-ии и проверить правильность классификации для пяти наблюдений, c = 39

	Рост	Вес
Рост	45	36
Вес	36	103

	Рост	Вес
Рост	43	1
Вес	1	111

Table: Ков.матр 1 и 2

	Рост	Вес	n
m	178	67	48
f	168	59	25

	Рост	Вес	Пол
1	167	54	1
2	189	89	1
3	185	66	1
4	178	63	1
5	172	53	1

Вариант 4. Возраст от 34 до 39. Найти коэф. дискр.ф-ии и проверить правильность классификации для пяти наблюдений, c = 59

	Рост	Вес
Рост	45	42
Вес	42	158

	Рост	Вес
Рост	32	20
Вес	20	45

Table: Ков.матр 1 и 2

	Рост	Вес	n
m	178	72	46
f	164	60	16

	Рост	Вес	Пол
1	190	93	1
2	180	65	1
3	198	75	1
4	169	53	1
5	165	59	1

Вариант 5. Возраст от 40 до 46. Найти коэф. дискр.ф-ии и проверить правильность классификации для пяти наблюдений, c = 84

	Рост	Вес
Рост	30	30
Вес	30	124

	Рост	Вес
Рост	23	8
Вес	8	70

Table: Ков.матр 1 и 2

	Рост	Вес	n
m	177	69	31
f	163	54	10

	Рост	Вес	Пол
1	175	51	1
2	181	79	1
3	180	70	1
4	176	76	1
5	165	59	1

Вариант 6. Возраст от 47 до 54. Найти коэф. дискр.ф-ии и проверить правильность классификации для пяти наблюдений, c = 66

	Рост	Вес
Рост	44	2
Вес	2	164

	Рост	Вес
Рост	5	12
Вес	12	39

Table: Ков.матр 1 и 2

	Рост	Вес	n
m	176	71	34
f	162	61	6

	Рост	Вес	Пол
1	178	60	1
2	172	60	1
3	176	76	1
4	188	56	1
5	192	70	1

Вариант 7. Возраст от 55. Найти коэф. дискр.ф-ии и проверить правильность классификации для пяти наблюдений, c = 60

	Рост	Вес
Рост	48	10
Вес	10	80

	Рост	Вес
Рост	10	-8
Вес	-8	170

Table: Ков.матр 1 и 2

	Рост	Вес	n
m	175	67	20
f	162	67	12

	Рост	Вес	Пол
1	165	54	2
2	155	56	2
3	164	49	2
4	165	65	1
5	167	80	1