Пример

W_1 - больные бронхитом, W_2 - больные пневмонией,
SH1 серосодержащие белки в восстановительной форме,
SOE1 скорость оседания эритроцитов.
$SH1 = 7.1 \pm 0.7$, $SOE1 = 32.9 \pm 4.4$ в W_1
$SH1 = 5.8 \pm 0.8$, $SOE1 = 46.0 \pm 5.4$ в W_2.

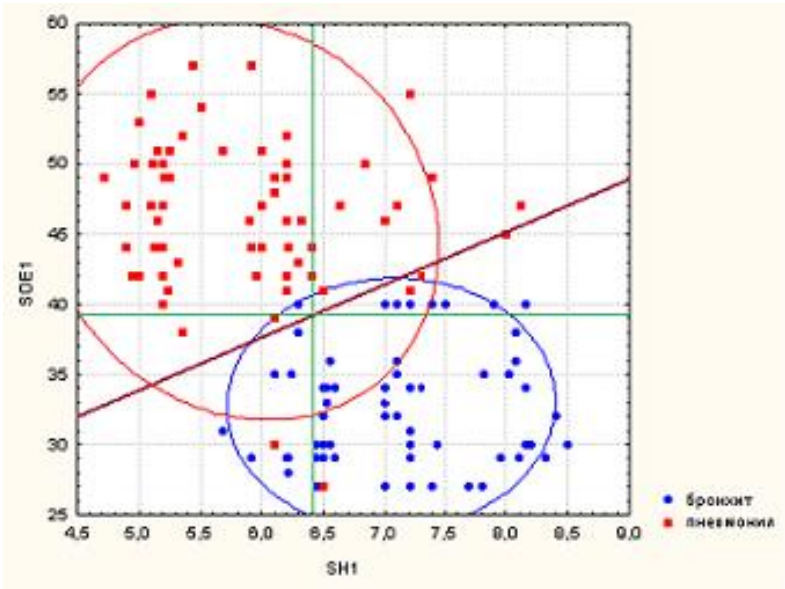

Figure: Классифицирующая прямая $SOE1 = 3.7SH1 + 15.5$.

Классификация индивидов в случае двух популяций

Пусть W_1 и W_2, которые характеризуются случайным вектором $(X_1, \ldots, X_p)$ с одинаковой известной ковариационной матрицей Σ и векторами средних μ_1 и μ_2.
Задача дискриминантного анализа состоит в том, чтобы решить, к какой из этих популяций отнести вектор наблюдений $(x_1, \ldots, x_p)^T$.
Нужно найти так называемую дискриминантную функцию 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 граничное значение c, так что если $z \geq c$, то наблюдение X относится к популяции W_1, и иначе в противоположном случае.

Например, $z(SH1, SOE1) = -1.97SH1 + 0.53SOE1$, $c = 8.22$.

$z(5.5, 40) = 10.365 > 8.22$ следовательно, скорее всего у больного с $(5.5, 40)$ имеет место пневмония.

Расстояние Махаланобиса

Если $W_i \sim \mathcal{N}(\mu_i, \Sigma)$, $i = 1, 2$, то $z = \alpha^T X \sim \mathcal{N}(\zeta_i, \sigma)$,
где $\zeta_i = \alpha^T \mu$, $\sigma^2 = \alpha^T \Sigma \alpha$.
Расстояние Махаланобиса

$$\Delta^2 = \frac{(\zeta_1 - \zeta_2)^2}{\sigma^2} = \frac{(\alpha^T(\mu_1 - \mu_2))^2}{\alpha^T \Sigma \alpha}.$$

Максимальное Δ^2 доставляет такой вектор α, который является решением системы уравнений $\Sigma\alpha = \mu_1 - \mu_2 = m$ [Фишер, 1936].

$$\Delta^2_{max} = m^T \Sigma^{-1} m > \Delta^2 = \frac{(\alpha^T m)^2}{\alpha^T \Sigma \alpha} \ \forall \alpha: \ \Sigma\alpha = y, \ \alpha = \Sigma^{-1} y$$

$$(m^T \Sigma^{-1} m)((y^T \Sigma^{-1})\Sigma(\Sigma^{-1} y)) > (y^T \Sigma^{-1} m)^2$$

То есть справедливо неравенство Коши-Буняковского

$$(m^T \Sigma^{-1} m)(y^T \Sigma^{-1} y) > (y^T \Sigma^{-1} m)^2.$$

Заметим, что при $\Sigma\alpha = \mu_1 - \mu_2$ имеет место

$$\sigma^2 = \alpha^T \Sigma \alpha = \alpha^T(\mu_1 - \mu_2) = |\zeta_1 - \zeta_2|,$$

$$\Delta^2 = \frac{(\zeta_1 - \zeta_2)^2}{|\zeta_1 - \zeta_2|} = |\zeta_1 - \zeta_2|.$$

Вычисление вероятности ошибочной классификации

Если $X \in W_2$, но $z \geq c$, то имеет место ошибка классификации $P(1|2)$, если $X \in W_1$, но $z \leq c$, то $P(2|1)$. Пусть $c_\delta = c + \delta$, где $c = \frac{\zeta_1 + \zeta_2}{2}$. Тогда

$$P_\delta(2|1) = \Phi\left(\frac{c + \delta - \zeta_1}{\sigma}\right) = \Phi\left(\frac{\frac{\zeta_1+\zeta_2}{2} + \delta - \zeta_1}{\Delta}\right) = \Phi\left(-\frac{\Delta}{2} + \frac{\delta}{\Delta}\right),$$

$$P_\delta(1|2) = P\{z > c + \delta | W_2\} = 1 - \Phi\left(\frac{c + \delta - \zeta_2}{\sigma}\right) = \Phi\left(-\frac{\Delta}{2} - \frac{\delta}{\Delta}\right).$$

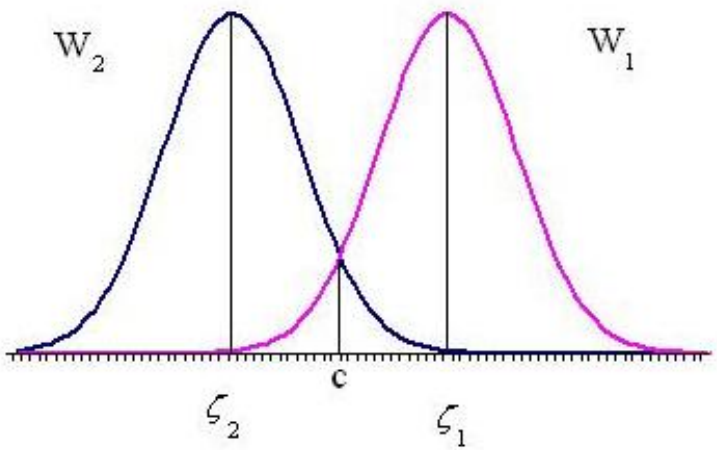

Figure: Распределения дискриминантной функции в случае двух популяций.

$$P_\delta(2|1) + P_\delta(1|2) = \Phi\left(-\frac{\Delta}{2} + \frac{\delta}{\Delta}\right) + \Phi\left(-\frac{\Delta}{2} - \frac{\delta}{\Delta}\right) =$$

$$= \Phi\left(-\frac{\Delta}{2}\right) + P\left\{-\frac{\Delta}{2} < z < -\frac{\Delta}{2} + \frac{\delta}{\Delta}\right\} +$$

$$+\Phi\left(-\frac{\Delta}{2}\right) - P\left\{-\frac{\Delta}{2} - \frac{\delta}{\Delta} < z < -\frac{\Delta}{2}\right\} > 2\Phi\left(-\frac{\Delta}{2}\right),$$

так как $P\{-\frac{\Delta}{2} - \frac{\delta}{\Delta} < z < -\frac{\Delta}{2}\} < P\{-\frac{\Delta}{2} < z < -\frac{\Delta}{2} + \frac{\delta}{\Delta}\}$.

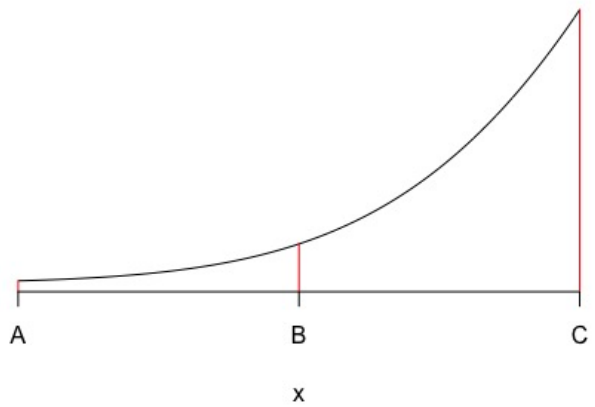

Figure: $A = -\frac{\Delta}{2} - \frac{\delta}{\Delta}$, $B = -\frac{\Delta}{2}$, $C = -\frac{\Delta}{2} + \frac{\delta}{\Delta}$.

Байесовская процедура классификации

- априорные вероятности $q_1 = P(W_1), q_2 = P(W_2)$ $(q_1 + q_2 = 1)$,
- апостериорные вероятности $P(W_1|X)$ и $P(W_2|X)$, определяемые как вероятности того, что объект принадлежит к данной популяции при данном векторе наблюдений X,
- $P(X|W_i)$ - 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; иначе объект относится к W_2.
Формула Байеса

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

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

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

Байесовская процедура для нормально распределенных популяций

Используем плотность многомерного нормального распределения

$$\ln\frac{q_1 f_1(x)}{q_2 f_2(x)} \geq 0 \Longleftrightarrow \ln\frac{q_1}{q_2} + \ln f_1(x) - \ln f_2(x) \geq 0,$$

$$\ln f_1(x) - \ln f_2(x) = \ln\left\{(2\pi)^{-p/2}|\Sigma|^{-1/2}\right\} - \frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) -$$

$$-\ln\left\{(2\pi)^{-p/2}|\Sigma|^{-1/2}\right\} + \frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2) =$$

$$= -\frac{1}{2}x^T\Sigma^{-1}x + \frac{1}{2}\mu_1^T\Sigma^{-1}x + \frac{1}{2}x^T\Sigma^{-1}\mu_1 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 +$$

$$+\frac{1}{2}x^T\Sigma^{-1}x - \frac{1}{2}\mu_2^T\Sigma^{-1}x - \frac{1}{2}x^T\Sigma^{-1}\mu_2 + \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 =$$

$$= \frac{1}{2}(\mu_1^T - \mu_2^T)\Sigma^{-1}x + \frac{1}{2}x^T\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}\mu_1\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2\Sigma^{-1}\mu_2 =$$

Прибавим и отнимем е $\mu_1^T\Sigma^{-1}\mu_2/2$,

$$= x^T\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}\mu_1^T\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 - \mu_2)^T\Sigma^{-1}\mu_2 =$$

$$= x^T\alpha - \frac{1}{2}\mu_1^T\alpha - \frac{1}{2}\alpha^T\mu_2 = x^T\alpha - \frac{\zeta_1 + \zeta_2}{2}.$$

Процедура классификации в случае двух многомерных нормальных популяций при неизвестных параметрах

Две независимые выборки: $X_1^1, \ldots, X_{n_1}^1$ объема n_1 из W_1, другая $X_1^2, \ldots, X_{n_1}^2$ объема n_2 из W_2. Выборочные оценки среднего и ковариационной матрицы для популяции W_i имеют вид

$$\bar{X}^i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_j^i, \ i = 1, 2,$$

$$S_i = \left\{ s_{kj}^i \right\}_{k,j=1}^{p} = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_j^i - \bar{X}^i)(X_j^i - \bar{X}^i)^T.$$

Выборочная ковариационная матрица

$$S = \frac{1}{n_1 + n_2 - 2}((n_1 - 1)S_1 + (n_2 - 1)S_2).$$

Решается система уравнений $\hat{\Sigma}\alpha = \hat{\mu}_1 - \hat{\mu}_2$, a – оценки,

$z = a^T X = \sum_{j=1}^{p} a_j x_j$ ДФ, $\bar{z}_i = a^T \bar{X}_i$ среднее значение ДФ при $X \in W_i$,

выборочная оценка дисперсии $s_z^2 = a^T S a$. Граничное значение

$$z_k = \frac{\bar{z}_1 + \bar{z}_2}{2} + \ln \frac{q_2}{q_1},$$

Априорная вероятность оценивается как $q_1 = \frac{n_1}{n_1+n_2}$, $q_2 = \frac{n_2}{n_1+n_2}$.
Выборочное расстояние Махаланобиса

$$D^2 = \frac{(\bar{z}_1 - \bar{z}_2)^2}{s_z^2}$$

является оценкой Δ^2. Нулевая гипотеза вида $H_0 : \Delta^2 = 0$ эквивалентна $H_0 : \mu_1 = \mu_2$ и проверяется на основе статистики

$$F = \frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} \cdot \frac{n_1 n_2}{n_1 + n_2} D^2$$

которая имеет в случае справедливости нулевой гипотезы распределение Фишера $\mathcal{F}(p, n_1 + n_2 - p - 1)$. Вероятность ошибочной классификации оценивается по формулам: $P(2|1) = \frac{m_1}{n_1}$, $P(1|2) = \frac{m_2}{n_2}$, где m_1 число индивидов из популяции W_1, которые на основе процедуры классификации относятся к популяции W_2; m_2 определяется аналогично.

Классификация в случае k популяций

Задача состоит в том, чтобы отнести некоторый вектор наблюдений $x = (x_1, \ldots, x_p)^T$ к одной из популяций $W_1, \ldots, W_k$.

- $q_i = P(W_i)$ априорная 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.

Апостериорная вероятность $P(W_j|x)$ выражается по формуле Байеса:

$$P(W_j|x) = \frac{q_j f_j(x)}{\sum_{i=1}^{k} q_i f_i(x)}. \quad (1)$$

Наблюдение относится к популяции с наибольшей апостериорной вероятностью.

Классификация k нормально распределенных популяций

Пусть популяция $W_i \sim \mathcal{N}(\mu_i, \Sigma)$, Введем в рассмотрение дискриминант

$$\delta_i = \ln(q_i f_i(x)) - C_1, \quad \text{где} \quad C_1 = -\frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| + \frac{1}{2}x^T\Sigma^{-1}x \qquad (2)$$

постоянная, которая оказывается одинаковой для всех популяций W_i.

$$\delta_i = \ln(q_i f_i(x)) - C_1 = \ln(q_i(2\pi)^{-\frac{p}{2}}|\Sigma|^{-\frac{1}{2}})\exp\{-\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i)\} - C_1 =$$

$$= \ln q_i - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i) - C_1 =$$

$$= \ln q_i + {\mu_i}^T\Sigma^{-1}x - \frac{1}{2}{\mu_i}^T\Sigma^{-1}\mu_i = \alpha_i^T x + \gamma_i + \ln q_i,$$

где $\alpha_i = \Sigma^{-1}\mu_i$, $\gamma_i = -\frac{1}{2}{\mu_i}^T\Sigma^{-1}\mu_i$. Вектор x относится к популяции с максимальным значением дискриминанта.

Апостериорная вероятность

Апостериорная вероятность имеет вид

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

т. к. $C_1 = -\frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| + \frac{1}{2}x^T\Sigma^{-1}x$

$$\delta_i = \ln q_i f_i(x) - C_1 = \ln q_i f_i(x) - \ln e^{C_1} = \ln \frac{q_i f_i(x)}{e^{C_1}},$$

$$q_i f_i(x) = e^{C_1} e^{\delta_i}, \quad \text{то}$$

$$P(W_i|x) = \frac{q_i f_i(x)}{\sum\limits_{j=1}^{k} q_j f_j(x)} = \frac{e^{C_1} e^{\delta_i}}{\sum\limits_{j=1}^{k} e^{C_1} e^{\delta_j}} = \frac{e^{\delta_i}}{\sum\limits_{j=1}^{k} e^{\delta_j}}.$$

Случай k популяций и неизвестных параметров

- n_i объем выборки, относящейся к популяции W_i,
- $\bar{x}_i$ вектор средних используется в качестве оценки μ_i,
- S_i оценка ковариационной матрицы для популяции W_i.

В качестве оценки общей ковариационной матрицы берется

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

Тогда в качестве оценки дискриминантов рассматриваются

$$d_i = a^T x + g_i + \ln q_i,$$

где $a_i = S^{-1}\bar{x}_i$, $g_i = -\frac{1}{2}\bar{x}_i^T S^{-1}\bar{x}_i$. Вектор $x \in W_i$ в случае, когда d_i максимально.

ПРИМЕР (продолжение)

Матрица коэффициентов дискриминантных функций имеет вид:

	W_1	W_2
SH1	12.81	10.84
SOE1	1.46	1.99
Constant	−69.95	−78.15

получаем выражения для дискриминантов:
$\delta_1 = 12.81\text{SH1} + 1.46\text{SOE1} - 69.95$, $\delta_2 = 10.84\text{SH1} + 1.99\text{SOE1} - 78.15$.

	Процент	W_1	W_2
W_1	95.45	63	3
W_2	92.54	5	62
Всего	93.98	68	65

Трое больных бронхитом имеют настолько высокие показатели SH1 и SOE1, что процедура классификации относит их к больным пневмонией, а пять больных пневмонией по своим показателям более близки к больным бронхитом. Общий процент правильной классификации равен 93.98.