1. Классификация с обучением

1.1. Классификация индивидов в случае двух популяций

Пусть имеется две нормально распределенные популяции W_1 и W_2 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 $X = (x_1, \ldots, x_p)^T$.

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

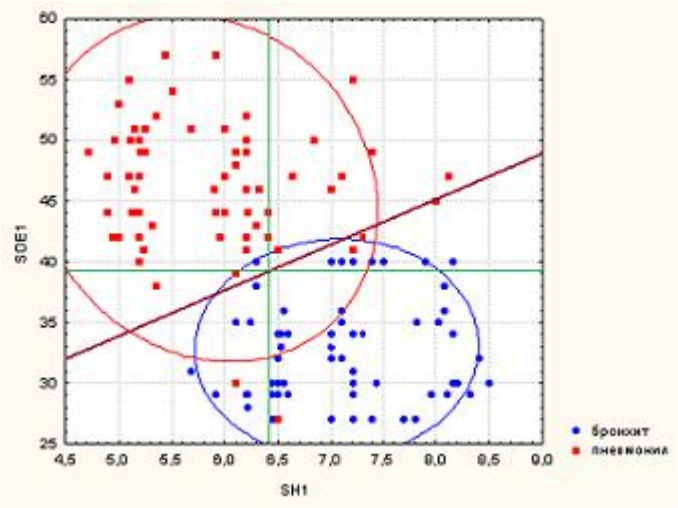

Рис. 1. Классифицирующая прямая $SOE1 = 3.7SH1 + 15.5$.

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 найти граничное значение c, так что если $z \geq c$, то наблюдение X относится к популяции W_1, и иначе в противоположном случае. Например, $z = -1.97x_1 + 0.53x_2$, $c = 8.22$ (рис.2), 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.

1.2. Расстояние Махаланобиса

Основная проблема заключается в выборе таких коэффициентов дискриминантной функции, чтобы ошибка классификации была как можно меньше.

Предложение. *Если наблюдение поступило из нормально распределенной популяции* $W \sim \mathcal{N}(\mu, \Sigma)$, *то* $z = \alpha^T X \sim \mathcal{N}(\zeta, \sigma)$, *где* $\zeta = \alpha^T \mu$, $\sigma^2 = \alpha^T \Sigma \alpha$.

Доказательство.

$$1. Ez = E\alpha^T X = \alpha^T EX = \alpha^T \mu,$$

$$2. \sigma^2 = E(\alpha^T X - \alpha^T \mu)^2 = E(\alpha^T (X - \mu))^2 =$$

$$= E(\alpha^T (x - \mu))(\alpha^T (X - \mu))^T = E\alpha^T (X - \mu)(X - \mu)^T \alpha =$$

$$= \alpha^T E(X - \mu)(X - \mu)^T \alpha = \alpha^T \Sigma \alpha.$$

Обозначим через $\zeta_i = \alpha^T \mu_i$ средние значения дискриминантной функции $z = \alpha^T X$ в случае популяции W_i, $i = 1, 2$. Имеет смысл выбрать такой вектор α, чтобы ζ_1 и ζ_2 были как можно дальше удалены друг от друга. Для этого введем расстояние Махаланобиса

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

Оно должно быть максимальным. В 1936 году Фишер показал, что *максимальное расстояние Махаланобиса доставляет такой вектор* α, *который*

является решением системы уравнений

$$\Sigma\alpha = \mu_1 - \mu_2. \tag{1}$$

Рассуждения, подтверждающие это, таковы. Обозначим $\mu_1 - \mu_2 = m$. Максимальное расстояние Махаланобиса доставляет вектор $\alpha = \Sigma^{-1}m$. Нетрудно убедиться, что $\Delta^2_{max} = \alpha^T\Sigma\alpha = m^T\Sigma^{-1}m$. Нужно показать, что расстояние Махаланобиса $\Delta^2 = \frac{(\alpha^T m)^2}{\alpha^T\Sigma\alpha}$ для любого другого вектора α будет меньше $\Delta^2_{max} = m^T\Sigma^{-1}m$. Действительно, пусть α является решением другого уравнения $\Sigma\alpha = y$. Следовательно, $\alpha = \Sigma^{-1}y, \alpha^T = y^T\Sigma^{-1}$. Тогда выражение $\Delta^2 \leq \Delta^2_{max}$ равносильно выражению

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

То есть справедливо неравенство Коши-Буняковского

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

Заметим, что при справедливости (1) имеет место

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

С другой стороны, $\Delta = \sigma$, так как

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

т.е. есть растояние Махаланобиса равно абсолюной величине разности между средними значениями дискриминантной функции.

1.3. 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 имеет место ошибка классификации $P(2|1)$.

Нужно, чтобы *ошибка классификации была минимальной*. Это достигается выбором постоянной, равноудаленной от средних

$$c = \frac{\zeta_1 + \zeta_2}{2}.$$

Покажем это. Пусть граница дискриминантной функции равна $c_\delta = c + \delta$. Вероятность $P_\delta(2|1)$ отнесения наблюдения к популяции W_2, когда оно принадлежит популяции W_1, определяется функцией $\Phi(x)$ стандартного нормального распределения

$$P_\delta(2|1) = P\{z < c + \delta | W_1\} = \Phi\left(\frac{c + \delta - \zeta_1}{\sigma}\right)$$

Учитывая, что $\Delta^2 = \sigma^2 = \zeta_1 - \zeta_2$ и подставляя значение c, имеем

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

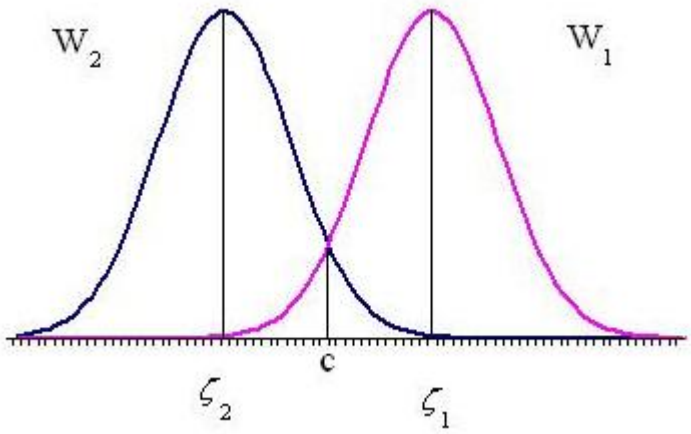

Рис. 2. Распределения дискриминантной функции в случае разных популяций.

Аналогично получаем

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

Минимальная вероятность ошибочной классификации

$$P_\delta = (P_\delta(2|1) + P_\delta(1|2))/2$$

достигается при $\delta = 0$ и равна $\Phi(-\frac{\Delta}{2})$. Действительно,

$$2P_\delta = P_\delta(2|1) + P_\delta(1|2) = \Phi\left(-\frac{\Delta}{2} + \frac{\delta}{\Delta}\right) + \Phi\left(-\frac{\Delta}{2} - \frac{\delta}{\Delta}\right) =$$
$$= \Phi\left(-\frac{\Delta}{2}\right) + P\left\{-\frac{\Delta}{2} < z < -\frac{\Delta}{2} + \frac{\delta}{\Delta}\right\} +$$
$$+\Phi\left(-\frac{\Delta}{2}\right) - P\left\{-\frac{\Delta}{2} - \frac{\delta}{\Delta} < z < -\frac{\Delta}{2}\right\} > 2\Phi\left(-\frac{\Delta}{2}\right),$$

так как $P\{-\frac{\Delta}{2} - \frac{\delta}{\Delta} < z < -\frac{\Delta}{2}\} < P\{-\frac{\Delta}{2} < z < -\frac{\Delta}{2} + \frac{\delta}{\Delta}\}$ вследствие того, что на интервале $(-\infty; 0)$ плотность стандартного нормального распределения является возрастающей функцией, а также $\zeta_1 > \zeta_2$ и $-\frac{\Delta}{2} < 0$.

Такая процедура классификации, не зависящая от априорной вероятности i-ой популяции, называется эвристической. В случае, когда априорные вероятности не одинаковы, требуется другой подход, связанный с байесовской процедурой классификации.

1.4. Байесовская процедура классификации

Определим априорные вероятности q_i того, что объект принадлежит популяции W_i, в частности, $q_1 = P(W_1), q_2 = P(W_2)$ $(q_1 + q_2 = 1)$, 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, а также $P(X|W_i)$ - условную вероятность получения вектора наблюдений X, если известно, что объект относится к популяции W_i.

Байесовская процедура классификации основана на соотношениями между апостериорными вероятностями: 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. 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, если известно, что объект принадлежит популяции W_i, может быть заменена на плотности $f_i(x)$:

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

Итак, 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,

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

Для нормально распределенных популяций с векторами средних μ_1 и μ_2 байесовская процедура классификации эквивалентна

$$ln\frac{q_1 f_1(x)}{q_2 f_2(x)} \geq 0 \Longleftrightarrow ln\frac{q_1}{q_2} + lnf_1(x) - lnf_2(x) \geq 0.$$

Используем формулу плотности многомерного нормального распределения

$$\begin{aligned} lnf_1(x) - lnf_2(x) = ln\left\{(2\pi)^{-p/2}|\Sigma|^{-1/2}\right\} - \frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - \\ -ln\left\{(2\pi)^{-p/2}|\Sigma|^{-1/2}\right\} + \frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2) = \\ = -\frac{1}{2}x^T\Sigma^{-1}x + \frac{1}{2}\mu_1^T\Sigma^{-1}x + \frac{1}{2}x^T\Sigma^{-1}\mu_1 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \\ +\frac{1}{2}x^T\Sigma^{-1}x - \frac{1}{2}\mu_2^T\Sigma^{-1}x - \frac{1}{2}x^T\Sigma^{-1}\mu_2 + \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 = \\ = \frac{1}{2}(\mu_1^T - \mu_2^T)\Sigma^{-1}x + \frac{1}{2}x^T\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}\mu_1\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2\Sigma^{-1}\mu_2 = \end{aligned}$$

Поскольку первые два слагаемые одинаковы, их можно объединить, кроме того, прибавим и отнимем одно и то же слагаемое $\mu_1^T\Sigma^{-1}\mu_2/2$, получаем с учетом (1) и определения $\zeta_i = \alpha^T\mu_i$

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

То есть байесовская процедура классификации для нормально распределенных популяций определяется неравенством

$$x^T\alpha \geq \frac{\zeta_1+\zeta_2}{2} + ln\frac{q_2}{q_1}.$$

Усовершенствование процедуры классификации заключается во включении в рассмотрение стоимости ошибочной классификации. Обозначим через $C(1|2)$ и $C(2|1)$ стоимости потерь при классификации популяции W_1, когда на самом деле популяция W_2 и наоборот. Обобщенная процедура классификации основана на неравенстве

$$x^T\alpha \geq \frac{\zeta_1+\zeta_2}{2} + ln\frac{q_2C(1|2)}{q_1C(2|1)}.$$

При выполнении этого неравенства объект относится к популяции W_1. Эта процедура минимизирует ожидаемую стоимость ошибочной классификации $q_1C(2|1)P(2|1) + q_2C(1|2)P(1|2)$. Вероятность ошибочной классификации определяется следующим образом. Пусть

$$K = ln\frac{q_2C(1|2)}{q_1C(2|1)},\ T = \frac{\zeta_1+\zeta_2}{2} + K.\ \text{Тогда}$$
$$P(2|1) = \Phi\left(\frac{T-\zeta_1}{\sigma}\right) = \Phi\left(\frac{\frac{\zeta_1+\zeta_2}{2}+K-\zeta_1}{\Delta}\right) =$$
$$= \Phi\left(\frac{\frac{\zeta_2-\zeta_1}{2}+K}{\Delta}\right) = \Phi\left(\frac{-\frac{\Delta^2}{2}+K}{\Delta}\right) = \Phi\left(-\frac{\Delta}{2}+\frac{K}{\Delta}\right),$$

так как $\sigma = \Delta$ и $\Delta^2 = \zeta_1 - \zeta_2$, через $\Phi(x)$ обозначена функция стандартного нормального распределения. Аналогично, используя $\Phi(-x) = 1 - \Phi(x)$,

получаем

$$P(1|2) = 1 - \Phi\left(\frac{T-\zeta_2}{\sigma}\right) = 1 - \Phi\left(\frac{\Delta}{2} + \frac{K}{\Delta}\right) = \Phi\left(-\frac{\Delta}{2} - \frac{K}{\Delta}\right).$$

1.5. Процедура классификации в случае двух многомерных нормальных популяций при неизвестных параметрах

При неизвестных параметрах распределения можно воспользоваться выборочными оценками. Обобзначим p-мерный вектор, соответствующий значениям p признаков для j-го индивида из i-й популяции через X_j^i. Рассмотрим две независимые выборки: $X_1^1, \ldots, X_{n_1}^1$ объема n_1 из популяции W_1, другая $X_1^2, \ldots, X_{n_1}^2$ объема n_2 из популяции W_2. Выборочные оценки среднего и ковариационной матрицы для популяции W_i имеют вид

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

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

В качестве объединенной выборочной ковариационной матрицы используем выражение:

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

Для построения дискриминантной функции решается система уравнений $\Sigma\alpha = \mu_1 - \mu_2$, где μ_i заменяется на $\bar{X}^i$, а матрица Σ оценкой S. Полученные оценки a используются для построения дискриминантной функции $z = a^T X = \sum_{j=1}^{p} a_j x_j$. Граничное значение дискриминантной функции определяется в виде

$$z_k = \frac{\bar{z}_1 + \bar{z}_2}{2} + \ln\frac{q_2 C(1|2)}{q_2 C(2|1)},$$

где $\bar{z}_i = a^T \bar{X}_i$ среднее значение дискриминантной функции, если X относится к популяции W_i. Выборочная оценка s_z^2 дисперсии дискриминантной

функции z имеет вид $s_z^2 = a^T S a$. Выборочное расстояние Махаланобиса

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

является оценкой Δ^2. Априорная вероятность оценивается как $q_1 = \frac{n_1}{n_1+n_2}$, $q_2 = \frac{n_2}{n_1+n_2}$. 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: $P(2|1) = \frac{m_1}{n_1}, \quad P(1|2) = \frac{m_2}{n_2}$, где m_1 число индивидов из популяции W_1, которые на основе процедуры классификации относятся к популяции W_2; m_2 определяется аналогично.

1.6. Классификация в случае k популяций

Пусть $\mathbf{x} = (x_1, \ldots, x_p)^T$ некоторый вектор наблюдений. Задача состоит в том, чтобы отнести его к одной из популяций $W_1, \ldots, W_k$ с произвольным распределением. Обозначим через $q_i = P(W_i)$ априорную вероятность того, что случайный вектор $\mathbf{X} \in W_i$, $f_i(\mathbf{x})$ – плотность распределения популяции W_i, C_{ij} – стоимость отнесения к популяции W_i элемента из популяции W_j, P_{ij} – вероятность отнесения к популяции W_i элемента из популяции W_j. Будем считать, что все параметры известны. Апостериорная вероятность $P(W_j|\mathbf{x})$ выражается по формуле Байеса:

$$P(W_j|\mathbf{x}) = \frac{q_j f_j(\mathbf{x})}{\sum_{i=1}^{k} q_i f_i(\mathbf{x})}. \tag{3}$$

Условная потеря от отнесения $\mathbf{x}$ к популяции W_t выразится в виде

$$L_t = \sum_{j=1}^{k} C_{tj} P(W_j|\mathbf{x}) = \frac{\sum\limits_{j=1}^{k} C_{tj} q_j f_j(\mathbf{x})}{\sum\limits_{i=1}^{k} q_i f_i(\mathbf{x})}.$$

Минимальность потери эквивалентна максимальности дискриминанта S_t, равного по определению $S_t \stackrel{df}{=} -\sum\limits_{j=1}^{k} C_{tj} q_j f_j(\mathbf{x})$. В случае, когда стоимость ошибочной классификации не имеет значения, то есть $C_{tj} = 1,\ t \neq j,\ C_{tt} = 0$, максимизация дискриминанта

$$S_t = -\sum_{j=1}^{k} q_j f_j(\mathbf{x})|_{t \neq j} = c + q_t f_t(\mathbf{x}), \quad \text{где} \ \ c = -\sum_{j=1}^{k} q_j f_j(\mathbf{x}), \tag{4}$$

эквивалентна максимизации числителя апостериорной вероятности (3).

Таким образом, обобщенная байесовская процедура относит наблюдение $\mathbf{x}$ к W_i, если дискриминант S_k является максимальным. При этом безусловная ожидаемая стоимость ошибочной классификации минимизируется. Покажем, что она имеет вид

$$M = \sum_{j=1}^{k} q_j (\sum_{t=1}^{k} P_{tj} C_{tj}) \tag{5}$$

Действительно, она вычисляется как

$$M = \sum_{t=1}^{k} M_t,$$

где $M_t = q_1 P_{t1} C_{t1} + \ldots + q_k P_{tk} C_{tk}$ есть ожидаемая стоимость ошибочной классификации при отнесении элемента к популяции W_t. Отсюда

$$M = q_1 \sum_{t=1}^{k} P_{t1} C_{t1} + \ldots + q_k \sum_{t=1}^{k} P_{tk} C_{tk} = \sum_{j=1}^{k} q_j \sum_{t=1}^{k} P_{tj} C_{tj}.$$

1.7. Классификация k нормально распределенных популяций

Пусть популяция W_i имеет нормальное распределение $N(\mu_\mathbf{i}, \mathbf{\Sigma})$, а стоимости ошибочной классификации не имеют значения: $C_{ij} = 1$ при $i \neq j$ и $C_{ii} = 0$. Введем в рассмотрение дискриминантную функцию вида

$$\delta_i = \ln(q_i f_i(\mathbf{x})) - C_1, \quad \text{где} \quad C_1 = -\frac{p}{2} ln(2\pi) - \frac{1}{2} ln|\mathbf{\Sigma}| + \frac{1}{2}\mathbf{x}^T \mathbf{\Sigma}^{-1}\mathbf{x} \qquad (6)$$

постоянная, которая оказывается одинаковой для всех популяций W_i.

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

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

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

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 $\mathbf{x}$ относится к популяции с максимальным значением дискриминантной функции.

1.8. Апостериорная вероятность

Апостериорная вероятность имеет вид

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

т. к. из (6) для выражения постоянной C_1 получаем

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

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

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

В случае неизвестных параметров μ и $\mathbf{\Sigma}$ можно воспользоваться их выборочными оценками. Пусть n_i объем выборки, относящейся к популяции W_i, $\bar{\mathbf{x}}_i$ 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

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

Тогда в качестве оценки дискриминантных функций рассматриваются

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

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

Если $k = 2$, то $\delta_1 = \alpha_1^T\mathbf{x} + \gamma_1 + \ln q_1$, $\delta_2 = \alpha_2^T\mathbf{x} + \gamma_2 + \ln q_2$. Наблюдение $\mathbf{x}$ относится к популяции W_i, если $\delta_1 \geq \delta_2$, что эквивалентно

$$\delta_1 - \delta_2 \geq 0 \iff (\alpha_1^T - \alpha_2^T)\mathbf{x} + (\gamma_1 - \gamma_2) + \ln\frac{q_1}{q_2} \geq 0.$$

Обозначим через $\alpha = \alpha_1 - \alpha_2 = \mathbf{\Sigma}^{-1}(\mu_\mathbf{1} - \mu_\mathbf{2})$.

$$\gamma_1 - \gamma_2 = -\frac{1}{2}\mu_\mathbf{2}{}^T\mathbf{\Sigma}^{-1}\mu_\mathbf{2} + \frac{1}{2}\mu_\mathbf{1}{}^T\mathbf{\Sigma}^{-1}\mu_\mathbf{1} + \frac{1}{2}\mu_\mathbf{1}{}^T\mathbf{\Sigma}^{-1}\mu_\mathbf{2} - \frac{1}{2}\mu_\mathbf{1}{}^T\mathbf{\Sigma}^{-1}\mu_\mathbf{2} =$$

$$\frac{1}{2}(\mu_\mathbf{1} - \mu_\mathbf{2})^T\mathbf{\Sigma}^{-1}\mu_\mathbf{2} + \frac{1}{2}\mu_\mathbf{1}{}^T\mathbf{\Sigma}^{-1}(\mu_\mathbf{1} - \mu_\mathbf{2}) = \frac{1}{2}\alpha^T\mu_\mathbf{2} + \frac{1}{2}\mu_\mathbf{1}{}^T\alpha = \frac{\zeta_1 + \zeta_2}{2}.$$

Таким образом получаем, что общая процедура классификации в случае $k = 2$ эквивалентна байесовской процедуре классификации вида

$$\alpha^T\mathbf{x} \geq= \frac{\zeta_1 + \zeta_2}{2} + \ln\frac{q_2}{q_1}.$$

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: $\delta_1 = 12.81SH1 + 1.46SOE1 - 69.95$, $\delta_2 = 10.84SH1 + 1.99SOE1 - 78.15$. Для больного из популяции больных пневмонией с показателями $SH1 = 5.5$ и $SOE1 = 40$ получаем значения дискриминанта $\delta_1 = 59 < \delta_2 = 61.2$, откуда следует, что процедура классификации для этого больного работает правильно. Однако правильная классификация осуществляется не всегда. Из матрицы классификации

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

узнаем, что 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. Процент 95.45 носит специальное название чувствительности (процент правильной идентификации более легких больных), а 92.54 специфичности (процент правильной идентификации более тяжелых больных). Общий процент правильной классификации равен 93.98.

1.9. Классификация в случае биномиальных распределений

Пусть имеются p дихотомических признаков $X_1, \ldots, X_p$, значения которых состоят только из нулей и единиц:

$$X_j = \begin{cases} 1, & "yes \\ 0, & "no" \end{cases}, \quad j = 1, \ldots, p,$$

например, ответы "да"и "нет"на ряд вопросов или наличие и отсутствие некоторых симптомов. Решается задача отнесения объекта к одной из k популяций $W_1, \ldots, W_k$ с биномиальными распределениями с известными априорными вероятностями $q_1, \ldots, q_k$. Положим

$$P(X_j = 1|W_i) = p_{ij} \quad \text{и} \quad P(X_j = 0|W_i) = 1 - p_{ij},$$

$i = 1, \ldots,\ k, j = 1, \ldots, p$. Тогда закон распределения X_j для популяции W_i имеет вид

$$f_i(x_j) = p_{ij}^{x_j}(1 - p_{ij})^{1-x_j}, \quad x_j \in \{0, 1\}.$$

При независимости признаков $X_1, \ldots, X_p$ совместный закон распределения имеет вид

$$f_i(\mathbf{x}) = f_i(x_1) \cdot \ldots \cdot f_i(x_p), \quad i = 1, \ldots, k, \quad \mathbf{x} = (x_1, \ldots, x_p)^T.$$

Вектор $\mathbf{x}$ относится к популяции W_i с максимальной величиной апостериорной вероятности, которая при одинаковой стоимости ошибочной классификации имеет вид:

$$P(W_i|\mathbf{x}) = \frac{q_i f_i(\mathbf{x})}{\sum\limits_{m=1}^{k} q_m f_m(\mathbf{x})}.$$

Пусть n_i число индивидов из популяции W_i, из них n_{ij} имеют j-ый симптом. Тогда оценка $\hat{p_{ij}} = n_{ij}/n_i$. Если априорные вероятности неизвестны, то их оценками являются $\hat{q_i} = n_i/n$, где $n_1 + \ldots + n_k = n$.

Поскольку в большинстве случаев предположение о независимости не выполняется, то когда все априорные вероятности равны, для классификации используется процедура, которая для любой комбинации симптомов вычисляет соответствующую долю объектов в каждой выборке. Новый объект относится к популяции, выборка из которой имеет наибольшую долю объектов с такой комбинацией. Например, $k = 2, p = 3$.

Комбинации	000	001	010	011	100	101	110	111
W_1	0.10	0.10	0.20	0.10	0.20	0.10	0.10	0.10
W_2	0.20	0.15	0.10	0.15	0.10	0.20	0.05	0.05

Поступающий объект с комбинацией симптомов 011 будет отнесен к популяции W_2, так как $0.15 > 0.10$. Неудобство этого метода заключается в том, что для каждой популяции надо вычислять 2^p параметров.

ПРИМЕР 1. Рассмотрим популяции (табл.1) W_1 (слова, означающие число "1 на $n_1 = 9$ языках) и W_2 (слова, означающие число "3 на тех же $n_2 = 9$ языках). В качестве дихотомических признаков X_j, $j = 1, 2, \ldots, 12$, будем использовать факт наличия в слове некоторых однотипных звуков. Имеем $k = 2$ число популяций, $p = 12$ число признаков. Пусть n_{ij} – число слов из популяции W_i, содержащих звук X_j, оценка p_{ij} имеет вид:

$$\widehat{p_{ij}} = n_{ij}/n_i, \quad i = 1, 2; \quad j = 1, 2, \ldots, 12.$$

Выражение $n_{14} = 3$ означает, что в популяции слов, означающих число "один три слова содержат буквы "к,х,г"(вахид, йек, нынг).

Оценками априорных вероятностей являются $\widehat{q_i} = n_i/n, \quad i = 1, 2, \ldots, k$, где $n = \sum_{l=1}^{k} n_l$ (табл.2). В нашем случае априорные вероятности равны $\widehat{q_1} = 9/18 = 0.5; \quad \widehat{q_2} = 9/18 = 0.5$.

Например, слову "вахид"соответствует вектор $X = (101110010000)$. Вероятность его появления при условии W_1 равна

$$\widehat{f_1}(X) = \frac{2}{9} \cdot \frac{8}{9} \cdot \frac{1}{9} \cdot \frac{3}{9} \cdot \frac{7}{9} \cdot \frac{6}{9} \cdot \frac{8}{9} \cdot \frac{8}{9} \cdot \frac{6}{9} \cdot \frac{9}{9} \cdot \frac{8}{9} \cdot \frac{9}{9} = 0.002.$$

язык	один	три
арабский	вахид	саляса
персидский	йек	сэ
турецкий	бит	юч
японский	ити	сан
амхарский	анд	сост
тайский	нынг	сам
грузинский	эрти	сами
китайский	и	сань
аккадский	иштен	шалаш

Таблица 1. Названия цифр 1 и 3 в восточных языках.

Если компонента $X_j = 1$, то соответствующий сомножитель равен $\widehat{p}_{1j}$, если $X_j = 0$, то сомножитель равен $1 - \widehat{p}_{1j}$. Аналогично вероятность появления вектора $X = (101110010000)$ при условии W_2 равна

$$\widehat{f_2}(X) = \frac{6}{9} \cdot \frac{9}{9} \cdot \frac{0}{9} \cdot \frac{1}{9} \cdot \frac{1}{9} \cdot \frac{8}{9} \cdot \frac{0}{9} \cdot \frac{1}{9} \cdot \frac{3}{9} \cdot \frac{8}{9} \cdot \frac{9}{9} \cdot \frac{8}{9} = 0.$$

Вероятность получения вектора $X = (101110010000)$ равна

$$P(X) = \widehat{q}_1 \cdot \widehat{f}_1(X) + \widehat{q}_2\widehat{f}_2(X) = 0.5 \cdot 0.002 + 0.5 \cdot 0 = 0.001.$$

Апостериорные вероятности равны

$$P(W_1|X) = \frac{\widehat{q}_1 \cdot \widehat{f}_1(X)}{P(X)} = \frac{0.5 \cdot 0.002}{0.001} = 1, \quad P(W_2|X) = \frac{\widehat{q}_2 \cdot \widehat{f}_2(X)}{P(X)} = \frac{0}{0.001} = 0.$$

Следовательно, слово "вахид"правильно классифицируется как слово из популяции W_1 единиц.

Результаты 100%-правильной классификации остальных слов сводятся в таблицу:

	к популяции W_1	к популяции W_2
из популяции W_1	9	0
из популяции W_2	0	9

ПРИМЕР 2. Рассмотрим, например, две популяции.

W_1	W_2
$n_1 = 5$	$n_2 = 13$
Россия	Италия
Болгария	Канада
Венгрия	Австралия
Румыния	Австрия
Польша	Бельгия
	Дания
	Мексика
	Нидерланды
	США
	Финляндия
	Франция
	Швеция
	Япония

Четыре ($p = 4$) дихотомических признака

- X_1: 0 – низкая (< 300) или 1 – высокая смертность от заболеваний системы кровообращения;

j	1	2	3	4	5	6	7	8	9	10	11	12
X_j	а я	б п	в ф	г к х	д т	е э	с з ш ч	й ы и	л м н	о	р	у ю
вахид	1	0	1	1	1	0	0	1	0	0	0	0
йек	0	0	0	1	0	1	0	1	0	0	0	0
бит	0	1	0	0	1	0	0	1	0	0	0	0
ити	0	0	0	0	1	0	0	1	0	0	0	0
анд	1	0	0	0	1	0	0	0	1	0	0	0
нынг	0	0	0	1	0	0	0	1	1	0	0	0
эрти	0	0	0	0	1	1	0	1	0	0	1	0
ити	0	0	0	0	1	0	0	1	0	0	0	0
иштен	0	0	0	0	1	1	1	1	1	0	0	0
саляса	1	0	0	0	0	0	1	0	1	0	0	0
сэ	0	0	0	0	0	1	1	0	0	0	0	0
юч	0	0	0	1	0	0	1	0	0	0	0	1
сан	1	0	0	0	0	0	1	0	1	0	0	0
сост	0	0	0	0	1	0	1	0	0	1	0	0
сам	1	0	0	0	0	0	1	0	1	0	0	0
сами	1	0	0	0	0	0	1	1	1	0	0	0
сань	1	0	0	0	0	0	1	0	1	0	0	0
шалаш	1	0	0	0	0	0	1	0	1	0	0	0
n_{1j}	2	1	1	3	7	3	1	8	3	0	1	0
n_{2j}	6	0	0	1	1	1	9	1	6	1	0	1

Таблица 2. Вычисление частот появления звуков в слове.

- X_2: 0 – низкая (< 189) или 1 – высокая смертность от злокачественных новообразований;
- X_3: 0 – низкая (< 10) или 1 – высокая смертность от инфекционных заболеваний;
- X_4: 0 – низкая (< 50) или 1 – высокая смертность от несчастных случаев и других внешних воздействий.

n_{ij} – количество стран из популяции W_i с высокими показателями смертности X_j. $\hat{p}_{ij} = \frac{n_{ij}}{n_i}$ – оценка вероятности высокого показателя смертности X_j в популяции W_i.

X_j	1	2	3	4
n_{1j}	5	3	2	5
n_{2j}	5	9	3	7

X_j	1	2	3	4
$\hat{p}_{1j}$	1	0.6	0.4	1
$\hat{p}_{2j}$	0.385	0.692	0.230	0.538

Выясним, к какой популяции W_1 или W_2 будет отнесена Германия со значениями признаков $X = (1100)^T$ результате 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 W_1 равна

$$\hat{f}_1(1,1,0,0) = \hat{p}_{11}\hat{p}_{12}(1-\hat{p}_{13})(1-\hat{p}_{14}) = 1 \cdot 0.6 \cdot (1-0.4) \cdot (1-1) = 0.$$

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

$$\hat{f}_2(1,1,0,0) = \hat{p}_{21}\hat{p}_{22}(1-\hat{p}_{23})(1-\hat{p}_{24}) = 0.385 \cdot 0.692 \cdot (1-0.230) \cdot (1-0.538) = 0.0945.$$

Оценки апостериорных вероятностей имеют вид:

$$P(W_1|1,1,0,0) = \frac{\hat{q}_1\hat{f}_1(1,1,0,0)}{\hat{q}_1\hat{f}_1(1,1,0,0) + \hat{q}_2\hat{f}_2(1,1,0,0)} = \frac{0}{0.0683+0} = 0,$$
$$P(W_2|1,1,0,0) = \frac{\hat{q}_2\hat{f}_2(1,1,0,0)}{\hat{q}_1\hat{f}_1(1,1,0,0) + \hat{q}_2\hat{f}_2(1,1,0,0)} = \frac{0.0683}{0.0683+0} = 1.$$

Наибольшая апостериорная вероятность достигается в случае популяции W_2, т.е. процедура классификации по показателям смертности характеризует Германию, в свое время объединившую свои восточную и западную части, как ее западную часть.