1. Классификация по расслоеннной выборке

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 $X = (x_1, \ldots, x_p)^T$.

Классификация осуществляется через построение линейной дискриминантной функции

$$f(X) = \alpha_1 x_1 + \ldots + \alpha_p x_p = \alpha^T X,$$

где $\alpha = (\alpha_1, \ldots, \alpha_p)^T$ вектор параметров, вычисление критического значения c и введение правила, согласно которому, если $f(X) > c$, то X относится к популяции W_1, иначе к W_2. Эта задача актуальна для диагностики и отдаленного прогнозирования. Если $f(X) \leq c$ для $X \in W_1$ или $f(X) \geq c$ для $X \in W_2$, то имеет место ошибочная классификация. Естественно необходимо, чтобы доля этой ошибки была как можно меньше. Для упорядочения признаков по степени их значимости для классификации используется пошаговая процедура. В случаях, когда линейные методы работают плохо, используют метод опорных векторов или нейронные сети. Но в этих методах остается скрытой структура классификации, а экспериментатору важно знать, за счет изменения каких признаков индивид относится к той или иной популяции.

Одной из распространенных причин неразделимости многомерных выборок при помощи гиперплоскостей является их неоднородность и связанное с нею различие в структуре значимых факторов для отдельных подвыборок. Оказалось, при достаточно большом объеме выборки можно снизить вероятность ошибочной классификации, если популяции расслоить и

построить отдельные линейные дискриминантные функции внутри слоев. Для выбора наилучшего расслаивающего фактора могут быть использованы корреляционные, симптомно-синдромальные или эвристические методы.

1.1. Вероятность ошибочной классификации

Если наблюдение из популяции $W \sim \mathcal{N}(\mu, \Sigma)$, то $z = \alpha^T X$ имеет нормальное распределение $\mathcal{N}(\zeta, \sigma)$, где $\zeta = \alpha^T \mu$, $\sigma^2 = \alpha^T \Sigma \alpha$.

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

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

В 1936 году Фишер показал, что *максимальное расстояние Махаланобиса доставляет вектор α, который является решением системы*

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

ПРИМЕР. Рассмотрим две популяции: W_1 - больные бронхитом ($n_1 = 66$), W_2 - больные пневмонией ($n_2 = 67$), которые характеризуются признаками: x_1 серосодержащие белки в восстановительной форме, x_2 скорость оседания эритроцитов. Средние значения с ошибкой равны $\mu_{11} = 7.1 \pm 0.7$ в W_1 и $\mu_{12} = 5.8 \pm 0.8$ в W_2 $\mu_{21} = 32.9 \pm 4.4$ в W_1 и $\mu_{22} = 46.0 \pm 5.4$ в W_2. Вычисляем оценку ковариационной матрицы:

$$\hat{\Sigma} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \begin{bmatrix} 0.57 & -0.18 \\ -0.18 & 24 \end{bmatrix},$$

где через S_1^2 и S_2^2 обозначены несмещенные оценки ковариационных матриц внутри популяций. Решая уравнение (1) относительно α, вычисляем коэффициенты α_1 и α_2 дискриминантной функции: $z = -1.97x_1 + 0.53x_2$. Среднее значение дискриминантной функции у больных бронхитом равно $\zeta_1 = \alpha^T \mu_1 = 3.45$, у больных с пневмонией $\zeta_2 = \alpha^T \mu_2 = 12.95$. Граничное значение равно $c = (\zeta_1 + \zeta_2)/2 = 8.22$. На рис. 1 уравнение классифицирующей прямой имеет вид: $x_2 = 3.7x_1 + 15.5$ Например, для больного

из W_2, у которого $x_1 = 5.5$, $x_2 = 40$, получаем $z = 10.37 > c = 8.22$. Этот больной классифицируется правильно, так как z находится ближе к $\zeta_2 = 12.95$, чем к $\zeta_1 = 3.45$.

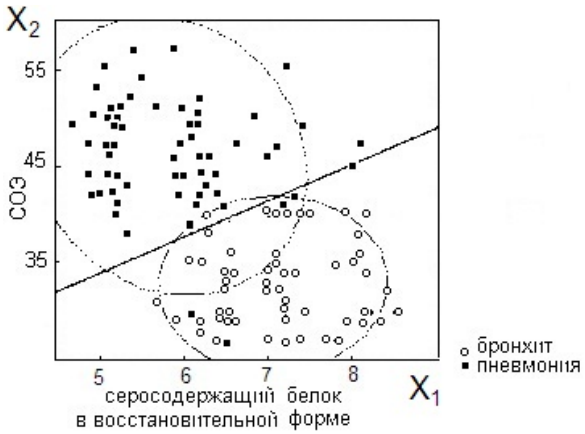

Рис. 1. Классификация больных по x_1 и x_2.

Для вычисления вероятности ошибочной классификации применяют формула полной вероятности:

$$\mathcal{P} = q_1 P(2|1) + q_2 P(1|2), \tag{2}$$

где $P(2|1)$ определяется как вероятность ошибочного отнесения индивида из W_1 к популяции W_2, $P(1|2)$ по аналогии (рис. 2), q_i — априорные вероятности того, что объект принадлежит популяции W_i: $q_1 = P(W_1), q_2 = P(W_2), q_1+q_2 = 1$. Через $P(W_1|X)$ и $P(W_2|X)$ обозначаются апостериорные вероятности того, что объект, характеризуемый вектором наблюдений X, принадлежит к W_i, а через $P(X|W_i)$ — условная вероятность получения вектора наблюдений X, если известно, что объект относится к популяции W_i.

Заменяя $P(X|W_i)$ на плотности $f_i(x)$, вычисляем апостериорные вероятности по формуле Байеса:

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

Наблюдение X относится к W_1, если $P(W_1|X) > P(W_2|X)$, что равносильно $q_1 f_1(x) \geq q_2 f_2(x)$, иначе к W_2. При этом, согласно [**?**] и [**?**], минимизируется (2). Если $W_i \sim \mathcal{N}(\mu_i, \Sigma)$, а вектор α есть решение системы (1) и $\zeta_i = \mu_i^T \alpha$, то классификация к W_1 означает выполнение неравенства:

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

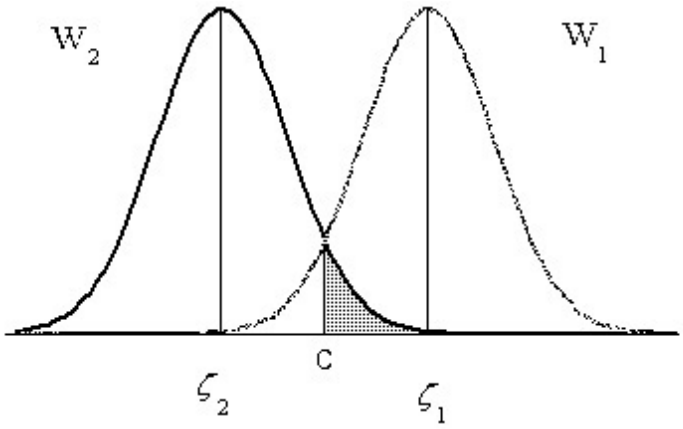

Рис. 2. Плотности дискриминантных функций.

Для определенности будем считать, что $\zeta_1 > \zeta_2$ [1]. Обозначим через $\Phi(x)$ функцию распределения $\mathcal{N}(0,1)$, а также

$$K = ln\frac{q_2}{q_1}, \quad T = \frac{\zeta_1 + \zeta_2}{2} + K. \tag{3}$$

Тогда, используя $\sigma = \Delta$ и $\Delta^2 = \zeta_1 - \zeta_2$, получим

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

[1] Иначе можно просто поменять знак у дискриминантной функции.

1.2. Расслоение популяции

Пусть $q = P(W_1)$ и $1 - q = P(W_2)$ — априорные вероятности. По некоторому дихотомическому признаку имеет место расслоение популяции на $S^1 = W_1^1 + W_2^1$ и $S^2 = W_1^2 + W_2^2$ с вероятностями $s = P(S^1)$, $1-s = P(S^2)$. Обозначим через $x = P(W_1^1)$, тогда

$$P(W_2^1) = s - x,\ P(W_1^2) = q - x,\ P(W_2^2) = 1 - s - q + x.$$

Предложение. *Если расстояния Махаланобиса $\Delta = \Delta_1 = \Delta_2$ инвариантны относительно расслоения популяции, то вероятность случайной классификации при расслоении не увеличится.*

Действительно, в соответствии с (3), по нерасслоенной ($i = 0$) и расслоенным ($i = 1, 2$) выборкам

$$u_i = \frac{K_i}{\Delta_i} - \frac{\Delta_i}{2},\ v_i = -\frac{K_i}{\Delta_i} - \frac{\Delta_i}{2},$$

где $K_0 = \ln\left(\frac{1}{q} - 1\right),\ K_1 = \ln\left(\frac{s}{x} - 1\right),\ K_2 = \ln\left(\frac{1-s}{q-x} - 1\right).$

Вычисляем вероятности ошибочной классификации $\mathcal{P}_i,\ i = 0, 1, 2,$:

$$\begin{aligned}
\mathcal{P}_0 &= q\Phi(u_0) + (1-q)\Phi(v_0), \\
s\mathcal{P}_1 &= x\Phi(u_1) + (s-x)\Phi(v_i), \\
(1-s)\mathcal{P}_2 &= (q-x)\Phi(u_2) + (1-s-q+x)\Phi(v_2),
\end{aligned}$$

Эффективность расслоения определяется через разность

$$\begin{aligned}
P(x) &= \mathcal{P}_0 - (s\mathcal{P}_1 + (1-s)\mathcal{P}_2) = \\
&= x(\Phi(u_0) - \Phi(u_1)) + (q-x)(\Phi(u_0) - \Phi(u_2)) + \\
&+(s-x)(\Phi(v_0) - \Phi(v_1)) + (1-s-q+x)(\Phi(v_0) - \Phi(v_2)).
\end{aligned}$$

Если признаки, определяющие расслоение общей популяции на W_1^1, W_2^1,W_1^2,W_2^2, независимы, то $x = qs$, следовательно, $K_0 = K_1 = K_2$, $u_0 = u_1 = u_2$,

$v_0 = v_1 = v_2$, $P(sq) = 0$. Дифференцируя по x, можно показать, что $P'(sq) = 0$, $P''(sq) > 0$, то есть в $x = qs$ функция $P(x)$ имеет минимум. Это означает, что улучшение классификации можно ожидать при большей зависимости признаков.

1.3. Задание для самостоятельной работы

1) Показать, что $P'(x) = 0$, $P''(xq) > 0$ при $x = qs$.

2) На данных о наркоманах сравнить результаты классификации фактора выполнение программы отказа от употребления наркотиков отдельно среди мужчин и женщин по индексам тяжести с результатами классификации по всей группе.