1) Рандомизированные и нерандомизированные решающие правила (РП).

Рассмотрим следующую статистическую задачу. Пусть имеется случайная выборка $X = (x_1^T, \dots, x_n^T)^T \in \mathcal{X} \subseteq \mathbb{R}^{nN}$ из некоторого N-мерного распределения вероятностей $\boldsymbol{P}_\theta(\cdot)$, заданного на измеримом пространстве $(\boldsymbol{\Omega}, \boldsymbol{\mathcal{F}})$, где $\theta \in \Theta \subseteq \mathbb{R}^m$ – неизвестное истинное значение векторного параметра; Θ – параметрическое пространство; $\mathcal{X} \subseteq \mathbb{R}^{nN}$ – выборочное пространство. Задано некоторое натуральное число $K \geqslant 2$ и определено некоторое разбиение параметрического пространства Θ на K областей:

Задача статистической проверки гипотез $H_0, \dots, H_{K-1}$ состоит в том, чтобы по наблюдаемой выборке X неким оптимальным образом оценить номер ν истинной гипотезы: $d = d(X) = k$ – выносим решение в пользу гипотезы H_k ($d = d(X)$ – статистическая оценка для ν). Возможно K решений ($k \in \{0, 1, \dots, K-1\}$). Множество возможных решений обозначим

$$D = \{0, 1, \dots, K-1\}, \ |D| = K,$$

и назовем *пространством решений*.

Определение 12.4. *Решающим правилом (решающей функцией, критерием, тестом) в вышесформулированной задаче статистической проверки гипотез называется функциональное отображение выборочного пространства $\mathcal{X}$ в пространство решений D:*

$$\mathcal{X} \xrightarrow{d(\cdot)} D. \tag{12.1}$$

Определение 12.5. *Нерандомизированным РП называется отображение* (12.1) *следующего вида:*

$$d = d(X) = \begin{cases} 0, & X \in \mathfrak{X}_0, \\ \vdots & \\ K-1, & X \in \mathfrak{X}_{K-1}, \end{cases}$$

где $\{\mathfrak{X}_0, \ldots, \mathfrak{X}_{K-1}\}$ – *некоторое детерминированное борелевское разбиение выборочного пространства:*

$$\mathfrak{X} = \bigcup_{k=0}^{K-1} \mathfrak{X}_k, \ \mathfrak{X}_k \cap \mathfrak{X}_l = \emptyset, \ k \neq l.$$

При этом, если выборка X *фиксирована, то решение* $d = d(X)$ *неслучайно.*

Определение 12.6. *Рандомизированным РП называется случайное отображение* (12.1) *следующего вида:*

$$d = d(X, \boldsymbol{\omega}), \ \boldsymbol{\omega} \in \boldsymbol{\Omega}, \ X \in \mathfrak{X}, \ d \in D,$$

причем если выборка X *фиксирована, то решение* $d = d(X, \boldsymbol{\omega})$ *является дискретной случайной величиной с множеством значений* D *и некоторым дискретным распределением вероятностей:*

$$\phi_i = \phi_i(X) = \boldsymbol{P}\{d = i | X\}, \ i \in D.$$

При этом борелевские функции $\phi_i = \phi_i(X)$, $i \in D$, *удовлетворяют следующим ограничениям:*

$$0 \leqslant \phi_i(X) \leqslant 1, \ i \in D; \quad \sum_{i \in D} \phi_i(X) = 1, \quad X \in \mathfrak{X},$$

и называются критическими функциями.

Укажем пошаговый алгоритм принятия решения с помощью рандомизированного решающего правила.

1. По выборке X вычисляем значения критических функций: $\phi_i = \phi_i(X)$, $i \in D$, и определяем дискретное распределение вероятностей $\{\phi_0(X), \phi_1(X), \ldots, \phi_{K-1}(X)\}$.

2. Проводим случайный эксперимент (жребий) со множеством исходов D и дискретным распределением вероятностей, найденным на шаге 1.

3. Регистрируем исход k этого жребия и принимаем решение $d = k$.

Нерандомизированное решающее правило есть частный случай рандомизированного решающего правила, если критические функции принимают одно из двух возможных значений:

$$\phi_i(X) \in \{0, 1\}, \ X \in \mathfrak{X}; \ \mathfrak{X}_i = \{X : \phi_i(X) = 1\}, \quad i \in D.$$

2) Риск как вероятность ошибочной классификации и его минимум, который достигается на байесовском РП (БРП). Нерандомизированный характер БРП и его запись через апостериорные вероятности классов.

Предположим, что параметр θ – случайная величина, принимающая одно из двух возможных значений:

$$\theta \in \Theta = \{\theta_0, \theta_1\};\ \boldsymbol{P}\{\theta = \theta_i\} = \Pi_i,\ 0 < \Pi_i < 1,\ i = 0, 1;\ \Pi_0 + \Pi_1 = 1.$$

Наблюдается случайная выборка $X = (x_1^T, \ldots, x_n^T)^T \in \mathbb{R}^{nN}$ объема n из некоторого распределения вероятностей с условной плотностью $p(x|\theta)$, $x \in \mathbb{R}^N$, $\theta \in \Theta$.

Обозначим

$$p_i(X) = \prod_{j=1}^{n} p(x_j|\theta_i),\ i = 0, 1, \quad -$$

условная плотность распределения выборки X при условии, что $\theta = \theta_i$.

Истинное значение θ неизвестно, и определены две простые гипотезы:

$$H_i:\ \theta = \theta_i,\ i = 0, 1.$$

Задача заключается в построении теста для проверки H_0, H_1 по выборке X.

Построим рандомизированное решающее правило:

$$\begin{gathered} d = d(X,\omega) \in D = \{0,1\},\ X \in \mathbb{R}^{nN},\ \omega \in \mathbf{\Omega}; \\ \boldsymbol{P}\{d(X,\omega) = 1|X\} = \phi(X),\ \boldsymbol{P}\{d(X,\omega) = 0|X\} = 1 - \phi(X), \end{gathered} \tag{12.9}$$

где $\phi(X)$ – произвольная критическая функция $(0 \leqslant \phi(X) \leqslant 1)$.

Обозначим: $\nu = \nu(\omega) \in \{0,1\}$ – случайная величина Бернулли – номер истинной гипотезы H_ν. В силу случайности θ

$$\boldsymbol{P}(H_i) = \boldsymbol{P}\{\theta = \theta_i\} = \Pi_i,\ i = 0,1,$$

поэтому Π_i принято называть *априорной вероятностью* i-й гипотезы.

Определение 12.12. *Функцией потерь в рассматриваемой задаче проверки двух гипотез H_0, H_1 называется функция двух переменных:*

$$w = w(i,j) \geqslant 0,\ i,j \in D = \{0,1\},$$

где $w(i,j)$ – величина потерь, которые несет статистик в ситуации, когда на самом деле $\nu = i$ (верна H_i), а принято решение $d = j$ в пользу гипотезы H_j.

Определение 12.13. *Принято говорить, что имеет место $(0-1)$-функция потерь, если*

$$w(i,j) = 1 - \delta_{ij} = \begin{cases} 0, & i = j; \\ 1, & i \neq j. \end{cases}$$

Функцию потерь удобно задавать в виде матрицы потерь: $W = (w_{ij})$, $w_{ij} = w(i,j)$. В случае $(0-1)$-матрицы потерь имеем

$$W = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Определение 12.14. *Функционалом риска называется математическое ожидание случайных потерь (средние потери)*

$$r = r(\phi(\cdot)) = \boldsymbol{E}\{w(\nu, d(X,\omega))\} \geqslant 0. \tag{12.10}$$

Определение 12.15 (байесовский принцип оптимальности). *Критическую функцию $\phi(\cdot)$ в рандомизированном решающем правиле (12.9) надлежит выбирать таким образом, чтобы функционал риска (12.10) достигал минимального значения:*

$$r(\phi^*(\cdot)) = \inf_{\phi(\cdot)} r(\phi(\cdot)). \tag{12.11}$$

При этом критическая функция $\phi^*(\cdot)$, определяемая (12.11), называется байесовской критической функцией, а соответствующее решающее правило $d^*(X,\omega)$, определяемое (12.9), – байесовским решающим правилом (БРП).

Теорема 12.2. *Пусть в сформулированной выше задаче проверки простых гипотез H_0, H_1 функция потерь имеет следующий вид:*

$$w(i,j) = \begin{cases} 0, & i = j, \\ w_0, & i = 0,\ j = 1, \\ w_1, & i = 1,\ j = 0, \end{cases} \tag{12.12}$$

где $w_0 > 0$, $w_1 > 0$ – некоторые заданные величины.

Тогда байесовская критическая функция задается соотношением ($X \in \mathbb{R}^{nN}$):

$$\phi^*(X) = \begin{cases} 0, & L(X) < C^*, \\ \varkappa^*, & L(X) = C^*, \\ 1, & L(X) > C^*, \end{cases} \tag{12.13}$$

где

$$L(X) = \frac{p_1(X)}{p_0(X)} \geqslant 0, \quad C^* = \frac{\Pi_0 w_0}{\Pi_1 w_1} \geqslant 0, \quad \varkappa^* \in [0,1]. \tag{12.14}$$

Следствие 12.2. *Среди байесовских решающих правил* (12.13) *существует нерандомизированное решающее правило:*

$$d = d^*(X) = \begin{cases} 0, & L(X) < C^*, \\ 1, & L(X) \geqslant C^*. \end{cases}$$

Доказательство. Для доказательства достаточно выбрать $\varkappa^* = 1$. □

Следствие 12.3. *Если имеет место $(0-1)$-функция потерь, т. е. $w_0 = w_1 = 1$, и гипотезы H_0, H_1 равновероятны $\Pi_0 = \Pi_1 = 1/2$, то БРП имеет вид*

$$d = d^*(X) = \begin{cases} 0, & p_0(X) > p_1(X), \\ 1, & p_1(X) \geqslant p_0(X). \end{cases} \tag{12.16}$$

Доказательство. Из формулы (12.14) имеем $C^* = 1, \varkappa^* = 1$. □

Заметим в заключение, что решающее правило (12.16) часто называют *тестом максимального правдоподобия.*

3) Модель Фишера и ее важность для практики.

Пусть теперь условные плотности $\{p_i(\cdot)\}_{i\in\mathcal{S}}$ из (16.14), описывающие классы $\{\boldsymbol{\Omega}_i\}_{i\in\mathcal{S}}$, – многомерные нормальные:

$$p_i(x) = n_N(x|\mu_i, \Sigma_i),\ x \in \mathbb{R}^N,\ i \in \mathcal{S}, \tag{16.19}$$

где наблюдения $x \in \mathbb{R}^N$ из класса $\boldsymbol{\Omega}_i$ ($d^o = i$) описываются условными: математическим ожиданием $\mu_i = \boldsymbol{E}\{x|\ d^o = i\}$ (так называемый центр i-го класса) и невырожденной ковариационной $(N \times N)$-матрицей $\Sigma_i = \boldsymbol{E}\{(x-\mu_i)(x-\mu_i)^T|\ d^o = i\}$ ($|\Sigma_i| \neq 0$).

В приложениях при решении реальных задач часто наблюдения, подлежащие классификации, адекватно определяются частным случаем модели (16.13), (16.19) – *моделью Фишера*:

$$p_i(x) = n_N(x|\mu_i, \Sigma),\ x \in \mathbb{R}^N,\ i \in \mathcal{S}, \tag{16.20}$$

с общей для всех классов невырожденной ковариационной $(N \times N)$-матрицей $\Sigma = \boldsymbol{E}\{(x-\mu_i)(x-\mu_i)^T|\ d^o = i\}$ ($i \in \mathcal{S}$, $|\Sigma| \neq 0$), описывающей статистический характер ошибок наблюдения: $x = \mu_{d^o} + \xi$, где распределение вероятностей N-вектора ошибок ξ не зависит от номера класса d^o, к которому принадлежит наблюдение x, и является N-мерным нормальным вектором с нулевым математическим ожиданим и ковариационной матрицей Σ ($\mathcal{L}\{\xi\} = N_N(\mathbf{0}_N, \Sigma)$).

Построим байесовское решающее правило.

Теорема 16.2. *Пусть классы $\{\boldsymbol{\Omega}_i\}_{i\in\mathcal{S}}$ определяются моделью* (16.13), (16.19) *с априорными вероятностями $\{\pi_i\}_{i\in\mathcal{S}}$ и невырожденными нормальными распределениями $\{N_N(\mu_i, \Sigma_i), |\Sigma_i| \neq 0\}_{i\in\mathcal{S}}$, тогда БРП* (16.15) *допускает представление* ($x \in \mathbb{R}^N$)*:*

$$d_o(x) = \arg\min_{i\in\mathcal{S}} \big((x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) + \ln|\Sigma_i| - 2\ln\pi_i\big), \tag{16.21}$$

и для модели Фишера (16.13), (16.20)*:*

$$d_o(x) = \arg\min_{i\in\mathcal{S}} \big((x-\mu_i)^T\Sigma^{-1}(x-\mu_i) - 2\ln\pi_i\big). \tag{16.22}$$

Доказательство. Очевидно и следует из соотношения

$$d_o(x) = \arg\max_{i\in\mathcal{S}}(\pi_i p_i(x)) = \arg\max_{i\in\mathcal{S}}\ln(\pi_i p_i(x)) = \arg\min_{i\in\mathcal{S}}(-2\ln(\pi_i p_i(x))),$$

и вида плотности многомерного нормального распределения (14.1). □

Следствие 16.2. *В условиях модели Фишера* (16.13), (16.20) *при равновероятных классах: $\pi_i = 1/L$, $i \in \mathcal{S}$, БРП имеет вид*

$$d_o(x) = \arg\min_{i\in\mathcal{S}} \rho(x, \mu_i),\ x \in \mathbb{R}^N, \tag{16.23}$$

где

$$\rho(x,y)=\sqrt{(x-y)^T\Sigma^{-1}(x-y)},\ x,y\in\mathbb{R}^N,\quad -\tag{16.24}$$

метрика Махаланобиса.

Следствие 16.3. *Для модели Фишера* (16.13), (16.20) *в случае двух классов* ($L=2$) *байесовский риск* r_o *из* (16.16) *может быть вычислен из соотношения*

$$r_o=\pi_1\Phi\left(-\frac{\Delta}{2}-\frac{h}{\Delta}\right)+\pi_2\Phi\left(-\frac{\Delta}{2}+\frac{h}{\Delta}\right),$$

$$h=\ln\frac{\pi_1}{\pi_2}=\ln\frac{\pi_1}{1-\pi_1},$$

и при равновероятных классах ($\pi_1=\pi_2=1/2$)

$$r_o=\Phi\left(-\frac{\Delta}{2}\right),$$

где $\Phi(\cdot)$ – *функция распределения вероятностей стандартного нормального закона* $N_1(0,1)$; $\Delta=\rho(\mu_1,\mu_2)=\sqrt{(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2)}$ – *межклассовое расстояние Махаланобиса* (*расстояние Махаланобиса между «центрами» классов*).

Доказательство. В случае двух классов ($L=2$) БРП (16.22) может быть записано в виде

$$d_o(x)=\begin{cases}1, & B(x)<0,\\ 2, & B(x)\geqslant 0,\end{cases}$$

где $B(x)=1/2\left((x-\mu_1)^T\Sigma^{-1}(x-\mu_1)-2\ln\pi_1-\left((x-\mu_2)^T\Sigma^{-1}(x-\mu_2)-2\ln\pi_2\right)\right)=b^Tx-H$ – линейная по $x\in\mathbb{R}^N$ функция с коэффициентами

$$b=\Sigma^{-1}(\mu_2-\mu_1),\quad H=\frac{1}{2}(\mu_1+\mu_2)^T\Sigma^{-1}(\mu_2-\mu_1)+h.$$

Для байесовского риска r_o из (16.16) имеем

$$r_o=\pi_1\boldsymbol{P}\{d_o(x)=2|d^o=1\}+\pi_2\boldsymbol{P}\{d_o(x)=1|d^o=2\}=$$

$$=\pi_1\boldsymbol{P}\{B(x)\geqslant 0|d^o=1\}+\pi_2\boldsymbol{P}\{B(x)<0|d^o=2\}.$$

Найдем условные распределения вероятностей случайной величины $B(x)$ при фиксированном номере класса $d^o=i$ для наблюдения x. Учтем, что при $d^o=i$ согласно модели Фишера случайный N-вектор-наблюдение $x\in\mathbb{R}^N$ имеет многомерное нормальное распределение $N_N(\mu_i,\Sigma)$, и по теореме 14.2 $B(x)=b^Tx-H\in\mathbb{R}^1$ также имеет условное нормальное распределение: $\mathcal{L}\{B(x)|d^o=i\}=N_1(m_i,\sigma_i^2)$, $i\in\mathcal{S}$. Найдем математическое ожидание m_i и дисперсию σ_i^2 ($i\in\mathcal{S}$, $\mathcal{S}=\{1,2\}$):

$$m_i=\boldsymbol{E}\{B(x)|d^o=i\}=b^T\mu_i-H=$$

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

$$=-\frac{1}{2}(\mu_1+\mu_2-2\mu_i)^T\Sigma^{-1}(\mu_2-\mu_1)-h=(-1)^i\frac{\Delta^2}{2}-h;$$

$$\sigma_i^2=\boldsymbol{D}\{B(x)|d^o=i\}=\operatorname{cov}\{b^Tx-H,b^Tx-H|d^o=i\}=$$

$$=b^T\operatorname{cov}\{x,x|d^o=i\}b=b^T\Sigma b=(\mu_2-\mu_1)^T\Sigma^{-1}\Sigma\Sigma^{-1}(\mu_2-\mu_1)=\Delta^2.$$

Продолжим преобразования риска r_o и получим

$$r_o = \pi_1 \boldsymbol{P}\left\{\frac{B(x)-m_1}{\Delta} \geqslant \frac{-m_1}{\Delta} \mid d^o = 1\right\} + \pi_2 \boldsymbol{P}\left\{\frac{B(x)-m_2}{\Delta} < \frac{-m_2}{\Delta} \mid d^o = 2\right\} =$$

$$= \pi_1\left(1 - \Phi\left(\frac{-m_1}{\Delta}\right)\right) + \pi_2 \Phi\left(\frac{-m_2}{\Delta}\right) = \pi_1 \Phi\left(\frac{m_1}{\Delta}\right) + \pi_2 \Phi\left(\frac{-m_2}{\Delta}\right) =$$

$$= \pi_1 \Phi\left(-\frac{\Delta}{2} - \frac{h}{\Delta}\right) + \pi_2 \Phi\left(-\frac{\Delta}{2} + \frac{h}{\Delta}\right),$$

где учтено свойство функции распределения стандартного нормального закона: $\Phi(z) = 1 - \Phi(-z)$, $z \in \mathbb{R}$. □

Замечание 16.3. Как видно из доказательства, в условиях модели Фишера БРП может быть выражено через функции, линейные по классифицируемому наблюдению, чего не скажешь о случае, когда ковариационные матрицы условных нормальных распределений, описывающих классы, различны. В связи с этим БРП для модели Фишера называется линейным, а при различных ковариационных матрицах – квадратичным.

4) Дискриминантный анализ в рамках модели Фишера.

Решим задачу дискриминантного анализа для модели (16.13), (16.19), когда классы $\{\Omega_i\}_{i\in\mathfrak{S}}$ описываются неизвестными априорными вероятностями $\{\pi_i\}_{i\in\mathfrak{S}}$ и невырожденными нормальными распределениями $\{N_N(\mu_i,\Sigma_i), |\Sigma_i|\neq 0\}_{i\in\mathfrak{S}}$ с неизвестными значениями параметров $\{\mu_i,\Sigma_i\}_{i\in\mathfrak{S}}$.

Наличие классифицированной обучающей выборки $X=\{x_1,\ldots,x_n\}$ объема n, для которой известен вектор истинной классификации $D^o=(d_1^o,\ldots,d_n^o)^T\in\mathfrak{S}^n$, позволяет построить несмещенные оценки неизвестных характеристик $\{\pi_i,\mu_i,\Sigma_i\}_{i\in\mathfrak{S}}$ классов $\{\boldsymbol{\Omega}_i\}_{i\in\mathfrak{S}}$ $(i\in\mathfrak{S})$:

$$\hat{\pi}_i=\frac{n_i}{n},\quad n_i=\sum_{t=1}^{n}\delta_{d_t^o,i}, \quad - \tag{16.25}$$

доля наблюдений, попавших, согласно вектору истинной классификации $D^o\in\mathfrak{S}^n$, в i-й класс;

$$\hat{\mu}_i=\overline{x}_{(i)}=\frac{1}{n_i}\sum_{t=1}^{n}\delta_{d_t^o,i}x_t \quad - \tag{16.26}$$

арифметическое среднее наблюдений из выборки X, попавших в i-й класс (оценка «центра» i-го класса);

$$\hat{\Sigma}_i=S_{(i)}=\frac{1}{n_i-1}\sum_{t=1}^{n}\delta_{d_t^o,i}(x_t-\hat{\mu}_i)(x_t-\hat{\mu}_i)^T \quad - \tag{16.27}$$

выборочная ковариационная матрица для i-го класса.

Несмещенность оценок априорных вероятностей (16.25) очевидна:

$$\boldsymbol{E}\{\hat{\pi}_i\}=\boldsymbol{E}\left\{\frac{1}{n}\sum_{t=1}^{n}\delta_{d_t^o,i}\right\}=\frac{1}{n}n\boldsymbol{P}\{d_t^o=i\}=\pi_i,\ i\in\mathfrak{S}.$$

Оценки $\hat{\mu}_i$ и $\hat{\Sigma}_i$ из (16.26) и (16.27) при фиксированном $D^o\in\mathfrak{S}^n$ согласно теореме 14.5 являются условно несмещенными оценками параметров многомерного нормального распределения $N_N(\mu_i,\Sigma_i)$, построенными по подвыборке $X^{(i)}=\{x_t\in X:\ d_t^o=i\}$ объема n_i: $\boldsymbol{E}\{\hat{\mu}_i|D^o\}=\mu_i$, $\boldsymbol{E}\{\hat{\Sigma}_i|D^o\}=\Sigma_i$, где условные математические ожидания не зависят от D^o, что означает безусловную несмещенность оценок (16.26) (16.27).

Подстановочное БРП, соответствующее (16.21) и основанное на оценках (16.25) – (16.27) имеет вид ($x \in \mathbb{R}^N$)

$$\hat{d}_o(x) = \arg\min_{i \in \mathcal{S}} \Big((x - \hat{\mu}_i)^T \hat{\Sigma}_i^{-1} (x - \hat{\mu}_i) + \ln |\hat{\Sigma}_i| - 2 \ln \hat{\pi}_i \Big), \tag{16.28}$$

и определяет процедуру *квадратичного дискриминантного анализа* (применяется, если $|\hat{\Sigma}_i| \neq 0$, $i \in \mathcal{S}$).

В случае модели Фишера (16.13), (16.20) ($\Sigma_i = \Sigma$, $i \in \mathcal{S}$) можно продолжать пользоваться квадратичным РП (16.28), но с точки зрения точности оценивания лучше проводить *линейный дискриминантный анализ*, основанный на БРП (16.22) ($x \in \mathbb{R}^N$):

$$\hat{d}_o(x) = \arg\min_{i \in \mathcal{S}} \Big((x - \hat{\mu}_i)^T \hat{\Sigma}^{-1} (x - \hat{\mu}_i) - 2 \ln \hat{\pi}_i \Big), \tag{16.29}$$

где в (16.29) вместо оценок ковариационных матриц $\{\hat{\Sigma}_i\}_{i \in \mathcal{S}}$ используется несмещенная оценка общей для всех классов ковариационной матрицы Σ:

$$\hat{\Sigma} = \frac{1}{n - L} \sum_{t=1}^{n} (x_t - \hat{\mu}_{d_t^o})(x_t - \hat{\mu}_{d_t^o})^T, \tag{16.30}$$

вычисляемая по всем n наблюдениям из выборки X (применяется, если $|\hat{\Sigma}| \neq 0$). Несмещенность оценки (16.30) следует из того, что при фиксированном векторе истинной классификации D^o матрица $(n - L)\hat{\Sigma}$ является суммой L независимых случайных матриц Уишарта:

$$(n - L)\hat{\Sigma} = \sum_{i \in \mathcal{S}} A_i, \quad A_i = \sum_{x_t \in X^{(i)}} (x_t - \hat{\mu}_i)(x_t - \hat{\mu}_i)^T;$$

$$\mathcal{L}\{A_i | D^o\} = W_N(\Sigma, n_i - 1), \ i \in \mathcal{S},$$

а значит, по свойствам распределения Уишарта

$$\mathcal{L}\{(n - L)\hat{\Sigma} | D^o\} = W_N \left(\Sigma, \sum_{i \in \mathcal{S}} (n_i - 1) \right) = W_N(\Sigma, n - L)$$

и $\boldsymbol{E}\{\hat{\Sigma} | D^o\} = \Sigma$ не зависит от D^o, поэтому $\boldsymbol{E}\{\hat{\Sigma}\} = \Sigma$.

Замечание 16.4. Перед проведением дискриминантного анализа целесообразно проверить гипотезу однородности (см. пп. 16.4.3):

$$\mu_1 = \ldots = \mu_L, \quad \Sigma_1 = \ldots = \Sigma_L.$$

Если она принимается, то выборка считается однородной, и дискриминантный анализ не проводится. В противном случае проверяется гипотеза о совпадении ковариационных матриц (см. п. 16.4.2):

$$\Sigma_1 = \ldots = \Sigma_L.$$

Ее принятие или отклонение позволяет соответственно сделать выбор в пользу линейного (модель Фишера) или квадратичного дискриминантного анализа.

5) Иерархический кластер-анализ и определение неизвестного числа классов для дискриминантного анализа.

Интернет:

Иерархический кластерный анализ - метод разбиения множества многомерных объектов на однородные группы, относящийся к классу агломеративных методов. Агломеративные методы последовательно объединяют отдельные объекты в кластеры.

Исходные данные:

- ***n***. Число объектов;
- ***m***. Число признаков;
- ***X***. Матрица объектов;
- ***Q***. Искомое число кластеров.

Пусть $X(n \times m)$ - матрица, описывающая n объектов в R^m. Алгоритм иерархической кластеризации заключается в последовательном слиянии кластеров с минимальным межкластерным расстоянием, начиная с n тривиальных кластеров по одному объекту в каждом и заканчивая на шаге $n - Q$ после построения ровно Q кластеров.

Изначально расстояния между однообъектными кластерами – это расстояния между соответствующими объектами, и на каждом шаге пересчету подлежит только расстояние от вновь образованного кластера посредством слияния до оставшихся кластеров.

Межкластерное расстояние и, соответственно, метод пересчета межкластерного расстояния между произвольным кластером i и кластером, образованным посредством объединения кластеров j, k может быть определен одним из способов:

- ближайший сосед:

$$d_{i,jk} = \min(d_{ij}, d_{ik})$$

- максимальное расстояние:

$$d_{i,jk} = \max(d_{ij}, d_{ik})$$

- групповое среднее:

$$d_{i,jk} = \frac{n_j}{n_j + n_k} d_{ij} + \frac{n_k}{n_j + n_k} d_{ik}$$

- центроид:

$$d_{i,jk} = \frac{n_j}{n_j + n_k} d_{ij} + \frac{n_k}{n_j + n_k} d_{ik} - \frac{n_j n_k}{(n_j + n_k)^2} d_{jk}$$

- медиана:

$$d_{i,jk} = \frac{1}{2}d_{ij} + \frac{1}{2}d_{ik} - \frac{1}{4}d_{jk}$$

- минимальная вариация:

$$d_{i,jk} = \frac{(n_i + n_j)d_{ij} + (n_i + n_k)d_{ik} - n_i d_{jk}}{n_i + n_i + n_k}$$

Где n_i, n_j, n_k - размеры кластеров.

Способы, используемые для определения первоначальных расстояний между объектами:

- сумма модулей:

$$\sum_i |x_{1,i} - x_{2,i}|$$

- евклидова норма:

$$\sum_i (x_{1,i} - x_{2,i})^2$$

- корень из евклидовой нормы:

$$\sqrt{\sum_i (x_{1,i} - x_{2,i})^2}$$

Далее выбирается метод кластеризации, на основе которого проводиться дальнейшее исследование. Также иерархическая кластеризация позволяет строить древовидную структуру (дендрограмму), последовательно объединяя или разделяя объекты.

6) Метод средних в кластер-анализе и его использование наряду с иерархическим кластер-анализом при определении неизвестного числа классов.

Однако на практике чаще используют упрощенный вариант приведенной выше процедуры кластер-анализа, основанный на модели Фишера и использовании решающего правила (16.23), в котором в качестве метрики наряду с метрикой Махаланобиса (16.24) может использоваться и любая другая метрика (например, евклидова). Данный подход известен как *метод L-средних*. Опишем его для метрик Махаланобиса и Евклида.

Шаг 0. Из выборки $X = \{x_1, \ldots, x_n\}$ выбираем какие-либо $L \geqslant 2$ наблюдений в качестве начальных приближений $\left\{\hat{\mu}_i^{(0)}\right\}_{i \in \mathcal{S}}$ для «центров» классов $\{\mu_i\}_{i \in \mathcal{S}}$. При использовании метрики Махаланобиса в качестве начального приближения для ковариационной матрицы Σ выбираем единичную матрицу: $\hat{\Sigma}^{(0)} = I_N$.

Шаг k $(k = 1, 2, \ldots)$. Классифицируем наблюдения из выборки по «близости» к «центрам» классов:

$$\hat{d}_t^{(k)} = \arg\min_{i \in \mathcal{S}} \rho\left(x_t, \hat{\mu}_i^{(k-1)}\right), \ t = 1, \ldots, n,$$

где либо

$$\rho\left(x_t, \hat{\mu}_i^{(k-1)}\right) = \sqrt{\left(x_t - \hat{\mu}_i^{(k-1)}\right)^T \left(\hat{\Sigma}^{(k-1)}\right)^{-1} \left(x_t - \hat{\mu}_i^{(k-1)}\right)} -$$

метрика Махаланобиса, либо

$$\rho\left(x_t, \hat{\mu}_i^{(k-1)}\right) = \left|x_t - \hat{\mu}_i^{(k-1)}\right| = \sqrt{\left(x_t - \hat{\mu}_i^{(k-1)}\right)^T \left(x_t - \hat{\mu}_i^{(k-1)}\right)} \quad -$$

метрика Евклида.

Получаем $\hat{D}^{(k)} = \left(\hat{d}_1^{(k)}, \ldots, \hat{d}_n^{(k)}\right)^T \in \mathcal{S}^n$ – оценка вектора истинной классификации $D^o \in \mathcal{S}^n$ на k-м шаге.

Уточняем оценки «центров» классов:

$$\hat{\mu}_i^{(k)} = \frac{1}{n_i^{(k)}} \sum_{t=1}^{n} \delta_{\hat{d}_t^{(k)}, i} \, x_t, \quad n_i^{(k)} = \sum_{t=1}^{n} \delta_{\hat{d}_t^{(k)}, i}, \ i \in \mathcal{S},$$

и ковариационной матрицы, если используется метрика Махаланобиса:

$$\hat{\Sigma}^{(k)} = \frac{1}{n - L} \sum_{t=1}^{n} \left(x_t - \hat{\mu}_{\hat{d}_t^{(k)}}^{(k)}\right) \left(x_t - \hat{\mu}_{\hat{d}_t^{(k)}}^{(k)}\right)^T.$$

Шаг-остановка. При $\hat{D}^{(k)} = \hat{D}^{(k-1)}$ $(k \geqslant 2)$ итерационный процесс останавливаем и полагаем: $\hat{\mu}_i := \hat{\mu}_i^{(k)}$, $i \in \mathcal{S}$, – оценки «центров» классов $\{\mu_i\}_{i \in \mathcal{S}}$; $\hat{\Sigma} := \hat{\Sigma}^{(k)}$ – оценка ковариационной матрицы Σ при использовании метрики Махаланобиса; $\hat{D} := \hat{D}^{(k)}$ – оценка вектора истинной классификации $D^o \in \mathcal{S}^n$.

Замечание 16.5. Эффективность всех методов кластер-анализа существенно зависит от межклассовых расстояний $\rho(\mu_i, \mu_j)$, $i \neq j \in \mathcal{S}$: чем они больше, тем меньше доля ошибочных решений (16.18) и ниже «чувствительность» метода к выбору начальных приближений для «центров» классов и метрики (в методе L-средних).

С методом L-средних связана еще одна характеристика – *псевдо-F-статистика Фишера*:

$$\mathrm{PFS}(L) = \frac{\dfrac{1}{L-1}\displaystyle\sum_{i\in\mathcal{S}} n_i |\hat{\mu}_i - \overline{x}|^2}{\dfrac{1}{n-L}\displaystyle\sum_{t=1}^{n} |x_t - \hat{\mu}_{\hat{d}_t}|^2};$$

$$\hat{\mu}_i = \frac{1}{n_i}\sum_{t=1}^{n} \delta_{\hat{d}_t,i}\, x_t,\ \ n_i = \sum_{t=1}^{n} \delta_{\hat{d}_t,i},\ \ i \in \mathcal{S}; \quad \overline{x} = \sum_{i\in\mathcal{S}} \frac{n_i}{n}\hat{\mu}_i = \frac{1}{n}\sum_{t=1}^{n} x_t,$$

которая при истинном числе классов L и $\hat{D} := D^o$ имеет F-распределение Фишера с $L-1$ и $n-L$ степенями свободы. Но более важно другое ее свойство: она имеет глобальный максимум на истинном числе классов. Это позволяет использовать ее для оценивания неизвестного числа классов L, задавая верхнюю границу L_+ и проводя кластер-анализ при каждом значении $L = 2, \ldots, L_+$:

$$\hat{L} = \arg \max_{2\leqslant L\leqslant L_+} \mathrm{PFS}(L).$$

GPT

Иерархическая кластеризация позволяет строить дендрограмму — древовидную структуру, которая отражает, как группы объектов объединяются на каждом этапе. Она особенно полезна, если число кластеров заранее неизвестно. Метод средних и иерархический анализ могут быть объединены для оптимального результата:

1. **Определение числа кластеров с помощью иерархического анализа**:
   - Построение дендрограммы.
   - Выбор оптимального уровня разбиения (на основе визуального анализа дендрограммы или специальных критериев, например, метода локтя или индекса Силуэтов).
2. **Использование k-means для уточнения кластеров**:
   - После определения количества кластеров (k) метод средних используется для детального распределения объектов, так как он быстрее и лучше работает на больших данных по сравнению с чисто иерархическими методами.

Преимущества совместного использования:

- **Иерархический анализ** обеспечивает наглядность и помогает выбрать оптимальное количество кластеров.
- **Метод средних** эффективен для обработки больших наборов данных и уточнения структуры кластеров.
- Вместе они уменьшают риск субъективного выбора числа кластеров и повышают точность анализа.

Недостатки и вызовы:

- Метод k-means чувствителен к выбору начальных центроидов, что может привести к локальным минимумам.
- Иерархическая кластеризация может быть вычислительно затратной на больших данных.
- Оба метода предполагают, что кластеры имеют определенную форму (чаще сферическую), что не всегда соответствует реальным данным.

Интернет про использование с иерархической кластеризацией

Сначала определяется **центр кластера**, а затем группируют все объекты в пределах заданного от центра порогового значения.

Недостатки:
- Чувствительность к выбросам
- Необходимо заранее задавать количество кластеров, а не как в иерархическом анализе, получать это в качестве результата

Проблему с выбором числа кластеров можно преодолеть проведением иерархического анализа со случайно отобранной выборкой наблюдений и, таким образом, определить оптимальное количество кластеров.

Достоинства:
- Простота использования
- В качестве метрики используется Евклидово расстояние
- Возможность наглядной интерпретации кластеров с использованием графика «Средних значений в кластерах»

7) Визуализация выборки в дискриминантном анализе, графики средних по классам в кластер-анализе.

Визуализация выборки в дискриминантном анализе.

В случае модели Фишера (16.13), (16.20) ($\Sigma_i = \Sigma$, $i \in \mathcal{S}$)

$$\boldsymbol{P}\{d^o(\omega) = i\} = \pi_i > 0,\ i \in \mathcal{S}; \quad \pi_1 + \ldots + \pi_L = 1, \tag{16.13}$$

$$p_i(x) = n_N(x|\mu_i, \Sigma),\ x \in \mathbb{R}^N,\ i \in \mathcal{S}, \tag{16.20}$$

~~ваться квадратичным РП (16.28), но~~ с точки зрения точности оценивания лучше проводить *линейный дискриминантный анализ*, основанный на БРП (16.22) ($x \in \mathbb{R}^N$):

$$\hat{d}_o(x) = \arg\min_{i \in \mathcal{S}} \left((x - \hat{\mu}_i)^T \hat{\Sigma}^{-1} (x - \hat{\mu}_i) - 2 \ln \hat{\pi}_i \right), \tag{16.29}$$

где в (16.29) вместо оценок ковариационных матриц $\{\hat{\Sigma}_i\}_{i \in \mathcal{S}}$ используется несмещенная оценка общей для всех классов ковариационной матрицы Σ:

$$\hat{\Sigma} = \frac{1}{n - L} \sum_{t=1}^{n} (x_t - \hat{\mu}_{d_t^o})(x_t - \hat{\mu}_{d_t^o})^T, \tag{16.30}$$

вычисляемая по всем n наблюдениям из выборки X (применяется, если $|\hat{\Sigma}| \neq 0$). Несме-

Замечание 16.4. Перед проведением дискриминантного анализа целесообразно проверить гипотезу однородности (см. пп. 16.4.3):

$$\mu_1 = \ldots = \mu_L, \quad \Sigma_1 = \ldots = \Sigma_L.$$

Если она принимается, то выборка считается однородной, и дискриминантный анализ не проводится. В противном случае проверяется гипотеза о совпадении ковариационных матриц (см. п. 16.4.2):

$$\Sigma_1 = \ldots = \Sigma_L.$$

Ее принятие или отклонение позволяет соответственно сделать выбор в пользу линейного (модель Фишера) или квадратичного дискриминантного анализа.

Рассмотрим типичные для задач статистической классификаци известные данные Фишера по ирисам [4] (Fisher Iris Data, 1936), представляющие собой наблюдения $x = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \tilde{x}_4)^T$ над четырьмя признаками ($N = 4$): $\tilde{x}_1$, $\tilde{x}_2$ – длина и ширина чашелистика, $\tilde{x}_3$, $\tilde{x}_4$ – длина и ширина лепестка цветка ириса. Всего 150 наблюдений ($n = 150$), принадлежащих к $L = 3$ классам ($n_1 = n_2 = n_3 = 50$): $\mathbf{\Omega}_1$ – ирис цветной (*Iris versicolor*), $\mathbf{\Omega}_2$ – ирис махровый (*Iris setosa*), $\mathbf{\Omega}_3$ – ирис чистый (*Iris virginica*).

Согласно замечанию 16.4 сначала решаем вопрос о целесообразности проведения классификации, проверяя гипотезу однородности (она отклоняется с уровнем значимости $\alpha = 0{,}05$). Затем устанавливаем, что ковариационные матрицы по классам совпадают (с уровнем значимости $\alpha = 0{,}05$), что приводит нас к модели Фишера и линейному дискриминантному анализу, проведя который, вычисляем долю ошибочных решений: $\gamma_n = 3/150 = 0,02$. На рис. 16.1 данные Фишера отображены в пространстве первых

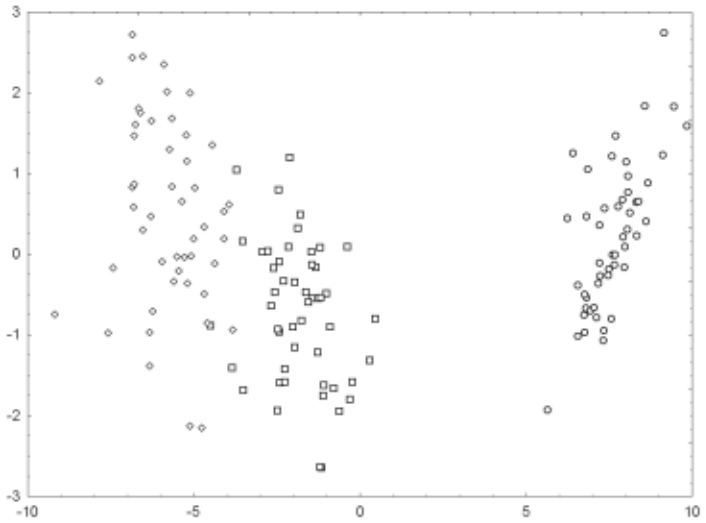

Рис. 16.1. Диаграмма рассеяния данных Фишера в пространстве двух главных компонент: □ – *Iris versicolor*; ○ – *Iris setosa*; ◇ – *Iris virginica*

двух главных компонент, вычисленных на основе построенной при дискриминантном анализе оценки ковариационной матрицы, а в табл. 16.1 приведен фрагмент полученной классификации с вычисленными оценками апостериорных вероятностей классов. Видно, что на ошибочных решениях (помечены «*») доминирующее значение апостериорной вероятности далеко от единицы.

Графики средних по классам в кластер-анализе.

Проведя кластер-анализ методом L-средних в метрике Евклида, имеем долю ошибочных решений: $\gamma_n = 16/150 = 0,11$, которая больше, чем в дискриминантном анализе, но по-прежнему приемлема. Полученные в кластер-анализе оценки «центров» классов показаны на рис. 16.2. Видно, что по каждому признаку они «достаточно» различаются. Отметим также, что для этих данных в предположении неизвестного числа классов псевдо F-статистика Фишера, подсчитанная по результатам метода L-средних при различных предполагаемых значениях числа классов, имеет глобальный максимум на истинном числе классов $L = 3$.

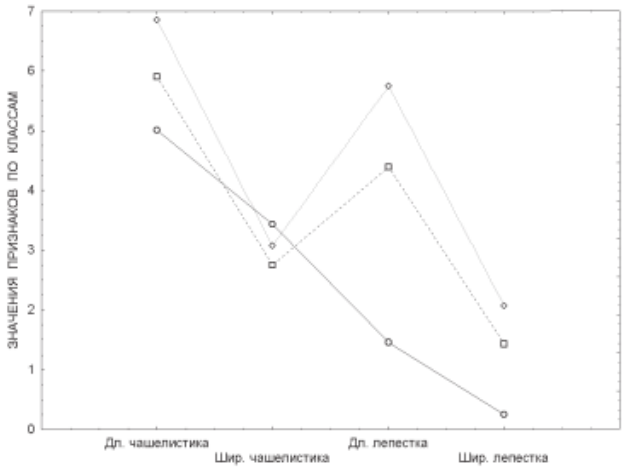

Рис. 16.2. Графики средних для данных Фишера по результатам кластер-анализа: □ – *Iris versicolor*; ○ – *Iris setosa*; ◇ – *Iris virginica*

ЧатГПТ говорит:

Определение: Графики средних по классам показывают средние значения признаков для каждого класса (или кластера), что позволяет сравнивать, как разные классы отличаются по своим характеристикам.

Цель: Используется для анализа и понимания различий между классами по каждому признаку. Это может быть представлено в виде столбчатых диаграмм, линейных графиков или тепловых карт.

8) Проведение кластер-анализа неоднородных выборок и интерпретация его результатов.

Рассмотрим типичные для задач статистической классификаци известные данные Фишера по ирисам [4] (Fisher Iris Data, 1936), представляющие собой наблюдения $x = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \tilde{x}_4)^T$ над четырьмя признаками ($N = 4$): $\tilde{x}_1$, $\tilde{x}_2$ – длина и ширина чашелистика, $\tilde{x}_3$, $\tilde{x}_4$ – длина и ширина лепестка цветка ириса. Всего 150 наблюдений ($n = 150$), принадлежащих к $L = 3$ классам ($n_1 = n_2 = n_3 = 50$): $\mathbf{\Omega}_1$ – ирис цветной (*Iris versicolor*), $\mathbf{\Omega}_2$ – ирис махровый (*Iris setosa*), $\mathbf{\Omega}_3$ – ирис чистый (*Iris virginica*).

Проведя кластер-анализ методом L-средних в метрике Евклида, имеем долю ошибочных решений: $\gamma_n = 16/150 = 0,11$, которая больше, чем в дискриминантном анализе, но по-прежнему приемлема. Полученные в кластер-анализе оценки «центров» классов показаны на рис. 16.2. Видно, что по каждому признаку они «достаточно» различаются. Отметим также, что для этих данных в предположении неизвестного числа классов псевдо F-статистика Фишера, подсчитанная по результатам метода L-средних при различных предполагаемых значениях числа классов, имеет глобальный максимум на истинном числе классов $L = 3$.

Алгоритм метода L-средних:

Шаг 0. Из выборки $X = \{x_1, \ldots, x_n\}$ выбираем какие-либо $L \geqslant 2$ наблюдений в качестве начальных приближений $\left\{\hat{\mu}_i^{(0)}\right\}_{i \in \mathcal{S}}$ для «центров» классов $\{\mu_i\}_{i \in \mathcal{S}}$. При использовании метрики Махаланобиса в качестве начального приближения для ковариационной матрицы Σ выбираем единичную матрицу: $\hat{\Sigma}^{(0)} = I_N$.

Шаг k $(k = 1, 2, \ldots)$. Классифицируем наблюдения из выборки по «близости» к «центрам» классов:

$$\hat{d}_t^{(k)} = \arg\min_{i \in \mathcal{S}} \rho\left(x_t, \hat{\mu}_i^{(k-1)}\right), \ t = 1, \ldots, n,$$

где

$$\rho\left(x_t, \hat{\mu}_i^{(k-1)}\right) = \left|x_t - \hat{\mu}_i^{(k-1)}\right| = \sqrt{\left(x_t - \hat{\mu}_i^{(k-1)}\right)^T \left(x_t - \hat{\mu}_i^{(k-1)}\right)} \quad -$$

метрика Евклида.

Получаем $\hat{D}^{(k)} = \left(\hat{d}_1^{(k)}, \ldots, \hat{d}_n^{(k)}\right)^T \in \mathcal{S}^n$ – оценка вектора истинной классификации $D^o \in \mathcal{S}^n$ на k-м шаге.

Уточняем оценки «центров» классов:

$$\hat{\mu}_i^{(k)} = \frac{1}{n_i^{(k)}} \sum_{t=1}^{n} \delta_{\hat{d}_t^{(k)}, i} \, x_t, \quad n_i^{(k)} = \sum_{t=1}^{n} \delta_{\hat{d}_t^{(k)}, i}, \ i \in \mathcal{S},$$

и ковариационной матрицы, если используется метрика Махаланобиса:

$$\hat{\Sigma}^{(k)} = \frac{1}{n - L} \sum_{t=1}^{n} \left(x_t - \hat{\mu}_{\hat{d}_t^{(k)}}^{(k)}\right) \left(x_t - \hat{\mu}_{\hat{d}_t^{(k)}}^{(k)}\right)^T.$$

Шаг-остановка. При $\hat{D}^{(k)} = \hat{D}^{(k-1)}$ $(k \geqslant 2)$ итерационный процесс останавливаем и полагаем: $\hat{\mu}_i := \hat{\mu}_i^{(k)}$, $i \in \mathcal{S}$, – оценки «центров» классов $\{\mu_i\}_{i \in \mathcal{S}}$; $\hat{\Sigma} := \hat{\Sigma}^{(k)}$ – оценка ковариационной матрицы Σ при использовании метрики Махаланобиса; $\hat{D} := \hat{D}^{(k)}$ – оценка вектора истинной классификации $D^o \in \mathcal{S}^n$.

Замечание 16.5. Эффективность всех методов кластер-анализа существенно зависит от межклассовых расстояний $\rho(\mu_i, \mu_j)$, $i \neq j \in \mathcal{S}$: чем они больше, тем меньше доля ошибочных решений (16.18) и ниже «чувствительность» метода к выбору начальных приближений для «центров» классов и метрики (в методе L-средних).

С методом L-средних связана еще одна характеристика – *псевдо-F-статистика Фишера*:

$$\mathrm{PFS}(L) = \frac{\dfrac{1}{L-1}\displaystyle\sum_{i\in\mathcal{S}} n_i |\hat{\mu}_i - \overline{x}|^2}{\dfrac{1}{n-L}\displaystyle\sum_{t=1}^{n} |x_t - \hat{\mu}_{\hat{d}_t}|^2};$$

$$\hat{\mu}_i = \frac{1}{n_i}\sum_{t=1}^{n} \delta_{\hat{d}_t,i}\, x_t,\ \ n_i = \sum_{t=1}^{n} \delta_{\hat{d}_t,i},\ \ i \in \mathcal{S}; \quad \overline{x} = \sum_{i\in\mathcal{S}} \frac{n_i}{n}\hat{\mu}_i = \frac{1}{n}\sum_{t=1}^{n} x_t,$$

которая при истинном числе классов L и $\hat{D} := D^o$ имеет F-распределение Фишера с $L-1$ и $n-L$ степенями свободы. Но более важно другое ее свойство: она имеет глобальный максимум на истинном числе классов. Это позволяет использовать ее для оценивания неизвестного числа классов L, задавая верхнюю границу L_+ и проводя кластер-анализ при каждом значении $L = 2, \ldots, L_+$:

$$\hat{L} = \arg \max_{2\leqslant L\leqslant L_+} \mathrm{PFS}(L).$$