# Практическая работа №3: Изучение дискриминантного анализа

## Цель работы

Ознакомиться с методами дискриминантного анализа на основе языка R.

## Основные теоретические положения

Дискриминантный анализ используется для принятия решения о том, какие переменные различают (дискриминируют) две или более возникающие совокупности (группы). Например, некий исследователь в области образования может захотеть исследовать, какие переменные относят выпускника средней школы к одной из трех категорий: (1) поступающий в колледж, (2) поступающий в профессиональную школу или (3) отказывающийся от дальнейшего образования или профессиональной подготовки. Для этой цели исследователь может собрать данные о различных переменных, связанных с учащимися школы. После выпуска большинство учащихся естественно должно попасть в одну из названных категорий. Затем можно использовать Дискриминантный анализ для определения того, какие переменные дают наилучшее предсказание выбора учащимися дальнейшего пути.

Медик может регистрировать различные переменные, относящиеся к состоянию больного, чтобы выяснить, какие переменные лучше предсказывают, что пациент, вероятно, выздоровел полностью (группа 1), частично (группа 2) или совсем не выздоровел (группа 3). Биолог может записать различные характеристики сходных типов (групп) цветов, чтобы затем провести анализ дискриминантной функции, наилучшим образом разделяющей типы или группы.

*Функции классификации.* Функции классификации предназначены для определения того, к какой группе наиболее вероятно может быть отнесен каждый объект. Имеется столько же функций классификации, сколько групп. Каждая функция позволяет вам для каждого образца и для каждой совокупности вычислить веса классификации по формуле: $$ S_i = c_i + \sum_{j = 1}^mw_{ij}x_j. $$ В этой формуле индекс $i$ обозначает соответствующую совокупность, а индекс $j$ обозначает переменную; $c_i$ являются константами для $i$-ой совокупности, $w_{ij}$ – веса для $j$-ой переменной при вычислении показателя классификации для $i$-ой совокупности; $x_j$ - наблюдаемое значение для соответствующего образца $j$-ой переменной. Величина $S_i$ является результатом показателя классификации. Переменные с наибольшими регрессионными коэффициентами вносят наибольший вклад в дискриминацию.

Расстояние Махаланобиса является мерой расстояния между двумя точками в пространстве, определяемым двумя или более коррелированными переменными. Например, если имеются всего две некоррелированных переменные, то можно нанести точки (образцы) на стандартную диаграмму рассеяния. Расстояние Махаланобиса между точками будет в этом случае равно расстоянию Евклида, т.е. расстоянию, измеренному, например, рулеткой. Если имеются три некоррелированные переменные, то для определения расстояния вы можно по-прежнему использовать рулетку (на 3М диаграмме). При наличии более трех переменных вы не можете более представить расстояние на диаграмме. Также и в случае, когда переменные

коррелированы, то оси на графике могут рассматриваться как неортогональные (они уже не направлены под прямыми углами друг к другу). В этом случае простое определение расстояния Евклида не подходит, в то время как расстояние Махаланобиса является адекватно определенным в случае наличия корреляций. Для расчёта ошибок классификации удобнее всего представить расстояние Махалонобиса, как симметричную матрицу с нулевой главной диагональю: $$ R = \begin{pmatrix} 0 & r_{12}^2 & \ldots & r_{1M}^2\\ r_{21}^2 & 0 & \ldots & r_{2M}^2\\ \vdots & \vdots & \ddots &\vdots\\ r_{M1}^2 & r_{M2}^2 & \ldots & 0 \end{pmatrix}, $$ $r_{ij} = r_{ji}$, $r_{ij} = 0$ при $i = j$. Элементы матрицы можно найти как: $$ r_{ij}^2 = (\mu_i - \mu_j)^T\Sigma^{-1}(\mu_i - \mu_j), $$ где $\mu_i$ и $\mu_j$ – вектора математических ожиданий для первого и второго класса соответственно, $\Sigma$ – ковариационная матрица. Вероятность ошибки можно определить следующим образом: $$ P(i \mid j) = \Phi\left(-\frac{r_{ij}}2\right) = 1 - \Phi\left(\frac{r_{ij}}2\right), $$ где $\Phi(\cdot)$ – функция ошибок.

*Пошаговый анализ с включением.* В пошаговом анализе дискриминантных функций модель дискриминации строится по шагам. Точнее, на каждом шаге просматриваются все переменные и находится та из них, которая вносит наибольший вклад в различие между совокупностями. Эта переменная должна быть включена в модель на данном шаге, и происходит переход к следующему шагу.

*Пошаговый анализ с исключением.* Можно также двигаться в обратном направлении, в этом случае все переменные будут сначала включены в модель, а затем на каждом шаге будут устраняться переменные, вносящие малый вклад в предсказания. Тогда в качестве результата успешного анализа можно сохранить только “важные” переменные в модели, то есть те переменные, чей вклад в дискриминацию больше остальных. Эта пошаговая процедура “руководствуется” соответствующим значением $F$ для включения и соответствующим значением $F$ для исключения. Значение $F$ статистики для переменной указывает на ее статистическую значимость при дискриминации между совокупностями, то есть, она является мерой вклада переменной в предсказание членства в совокупности.

# Постановка задачи

# Порядок выполнения работы

#### Задание на разработку статистических данных

Модель представляет собой набор многомерных векторов $\vec x = (x_1, \dots, x_m)$ , $m = 2, 3$, имеющих заданные вектора математических ожиданий $\mu_i$, $i = 1..M$ и заданные ковариационные матрицы (одинаковые по классам), которые имеют вид $\Sigma = \mathop{\mathrm{diag}}\nolimits\{\sigma_1, \dots, \sigma_m\}$. Компоненты векторов имеют нормальное распределение. Количество классов равно $M = 2, 3$.

| № варианта | Размерность $m$ | Объём выборки $N$ по классу | Вектора $\mu_i$ | Значения $\sigma_i$ | Количество классов $M$ |
|---|---|---|---|---|---|

| № варианта | Размерность $m$ | Объём выборки $N$ по классу | Вектора $\mu_i$ | Значения $\sigma_i$ | Количество классов $M$ |
|---|---|---|---|---|---|
| 1 | 2 | 100 | $\mu_1 = (1,\,2)^T \\ \mu_2 = (1,\,-2)^T$ | $\sigma_1 = 1 \\ \sigma_2 = 1$ | 2 |
| 2 | 2 | 300 | $\mu_1 = (1.5,\,3)^T \\ \mu_2 = (3,\,4)^T \\ \mu_3 = (-1.5,\,-1)^T$ | $\sigma_1 = 1 \\ \sigma_2 = 1$ | 3 |
| 3 | 3 | 150 | $\mu_1 = (1,\,1,\,1)^T \\ \mu_2 = (2,\,2,\,2)^T$ | $\sigma_1 = 0.5 \\ \sigma_2 = 1 \\ \sigma_3 = 0.5$ | 2 |
| 4 | 3 | 150 | $\mu_1 = (1,\,1,\,1)^T \\ \mu_2 = (2.5,\,2.5,\,2.5)^T \\ \mu_3 = (4,\,4,\,4)^T$ | $\sigma_1 = 0.5 \\ \sigma_2 = 1 \\ \sigma_3 = 2$ | 3 |
| 5 | 2 | 200 | $\mu_1 = (-0.5,\,2)^T \\ \mu_2 = (-1,\,4)^T$ | $\sigma_1 = 1.5 \\ \sigma_2 = 1$ | 2 |
| 6 | 2 | 250 | $\mu_1 = (1,\,1)^T \\ \mu_2 = (4,\,2.5)^T \\ \mu_3 = (-1,\,3)^T$ | $\sigma_1 = 1.3 \\ \sigma_2 = 0.8$ | 3 |
| 7 | 3 | 100 | $\mu_1 = (0,\,0,\,0)^T \\ \mu_2 = (3,\,3,\,3)^T$ | $\sigma_1 = 1.5 \\ \sigma_2 = 1 \\ \sigma_3 = 2$ | 2 |
| 8 | 3 | 150 | $\mu_1 = (4,\,4.5,\,3.7)^T \\ \mu_2 = (5.2,\,4.9,\,4.1)^T \\ \mu_3 = (2.2,\,3.9,\,3.8)^T$ | $\sigma_1 = 0.3 \\ \sigma_2 = 0.5 \\ \sigma_3 = 0.6$ | 3 |
| 9 | 2 | 250 | $\mu_1 = (1.5,\,1)^T \\ \mu_2 = (3,\,2.7)^T$ | $\sigma_1 = 1 \\ \sigma_2 = 1.2$ | 2 |
| 10 | 2 | 180 | $\mu_1 = (-0.9,\,2.7)^T \\ \mu_2 = (0,\,-5)^T \\ \mu_3 = (-1,\,-1)^T$ | $\sigma_1 = 0.25 \\ \sigma_2 = 2$ | 3 |
| 11 | 3 | 200 | $\mu_1 = (-1,\,2,\,-3)^T \\ \mu_2 = (2,\,-1,\,0)^T$ | $\sigma_1 = 0.75 \\ \sigma_2 = 0.3 \\ \sigma_3 = 1.6$ | 2 |
| 12 | 3 | 200 | $\mu_1 = (-1,\,-1,\,-1)^T \\ \mu_2 = (0,\,0,\,0)^T \\ \mu_3 = (2,\,2,\,2)^T$ | $\sigma_1 = 1 \\ \sigma_2 = 0.5 \\ \sigma_3 = 1$ | 3 |

# Содержание отчёта