Министерство образования и науки Российской Федерации

Калужский филиал
федерального государственного бюджетного образовательного
учреждения высшего образования
**«Московский государственный технический университет
имени Н.Э. Баумана
(национальный исследовательский университет)»**
(КФ МГТУ им. Н.Э. Баумана)

И.И. Кручинин
(к.т.н. доцент)

ЛАБОРАТОРНАЯ РАБОТА № 3
по курсу «Методы машинного обучения»

ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Калуга
2018

Дискриминантный анализ

Цель работы: изучение основных процедур дискриминантного анализа: дискриминации и классификации, построение и определение количества дискриминантных функций и их разделительной способности, нахождение классифицирующих функций с использованием функций Фишера и расстояния Махаланобиса.

ВВЕДЕНИЕ

Дискриминантный анализ является разделом многомерного статистического анализа, который позволяет изучать различия между двумя и более группами объектов по нескольким переменным одновременно.

Дискриминантный анализ – это общий термин, относящийся к нескольким тесно связанным статистическим процедурам. Эти процедуры можно разделить на методы *интерпретации межгрупповых различий – дискриминации* и методы *классификации наблюдений* по группам.

ЗАДАНИЕ

1. Составить программу на языке R и оценить следующие характеристики:
   - среднее значение переменных внутри классов, общее среднее;
   - матрицу перекрестных произведений и ковариационную матрицу общего рассеяния;
   - матрицу внутригрупповых квадратов и перекрестных произведений и корреляционную матрицу;
   - матрицу межгрупповых квадратов и перекрестных произведений и корреляционную матрицу;
   - коэффициенты канонической дискриминантной функции;
   - коэффициенты классифицирующей функции Фишера;
   - используя оценки априорных вероятностей принадлежности объектов к группам, определить расстояние Махаланобиса;
   - вычислить обобщенное расстояние Рао и его значимость.
2. Оформить отчет.

.

ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

1. ДИСКРИМИНАЦИЯ

Основной целью дискриминации является нахождение такой линейной комбинации переменных (в дальнейшем эти переменные будем называть *дискриминантными переменными*), которая бы оптимально разделила рассматриваемые группы. Линейная функция

$$d_{km}=\beta_0+\beta_1 x_{1km}+\ldots+\beta_p x_{pkm},\ m=1,\ldots,n,\ k=1,\ldots,g \qquad (1)$$

называется *канонической дискриминантной функцией* с неизвестными коэффициентами β_i. Здесь d_{km} — значение дискриминантной функции для *m*-го объекта в группе *k*; x_{ikm} — значение дискриминантной переменной X_i для *m*-го объекта в группе *k*. С геометрической точки зрения дискриминантные функции определяют гиперповерхности в *p*-мерном пространстве. В частном случае при *p* = 2 она является прямой, а при *p* = 3 – плоскостью.

Коэффициенты β_i первой канонической дискриминантной функции выбираются таким образом, чтобы центроиды различных групп как можно больше отличались друг от друга. Коэффициенты второй группы выбираются также, но при этом налагается дополнительное условие, чтобы значения второй функции были некоррелированы со значениями первой. Аналогично определяются и другие функции. Отсюда следует, что любая каноническая дискриминантная функция d имеет нулевую внутригрупповую корреляцию с $d_1,\ldots,d_{g-1}$. Если число групп равно *g*, то число канонических дискриминантных функций будет на единицу меньше числа групп. Однако по многим причинам практического характера полезно иметь одну, две или же три дискриминантных функций. Тогда графическое изображениее объектов будет представлено в одно-, двух- и трехмерных пространствах. Такое представление особенно полезно в случае, когда число дискриминантных переменных *p* велико по сравнению с числом групп *g*.

1.1. Коэффициенты канонической дискриминантной функции

Для получения коэффициентов β_i канонической дискриминантной функции нужен статистический критерий различения групп. Очевидно, что *классификация переменных будет осуществляться тем лучше, чем меньше рассеяние точек относительно центроида внутри группы и чем больше расстояние между центроидами групп*. Разумеется, что большая внутригрупповая вариация нежелательна, так как в этом случае любое за-

данное расстояние между двумя средними тем менее значимо в статистическом смысле, чем больше вариация распределений, соответствующих этим средним. Один из методов поиска наилучшей дискриминации данных заключается в нахождении такой канонической дискриминантной функции d, которая бы максимизировала отношение межгрупповой вариации к внутригрупповой [1, 2, 3, 4]

$$\lambda = B(d)/W(d), \tag{2}$$

где **B** – межгрупповая и **W** – внутригрупповая матрицы рассеяния наблюдаемых переменных от средних. В некоторых работах [3, 4] в (2) вместо **W** используют матрицу рассеяния **T** объединенных данных.

Рассмотрим максимизацию отношения (2) для произвольного

числа классов. Введем следующие обозначения:

g – число классов;
p – число дискриминантных переменных;
n_k – число наблюдений в k-й группе;
n – общее число наблюдений по всем группам;
x_{ikm} – величина переменной i для m-го наблюдения в k-й группе;
$\bar{x}_{ik}$ – средняя величина переменной i в k-й группе;
$\bar{x}_i$ – среднее значение переменной i по всем группам;
$T(u,v)$ – общая сумма перекрестных произведений для переменных u и v;
$W(u,v)$ – внутригрупповая сумма перекрестных произведений для переменных u и v;
$t_{ij} - T(x_i, x_j)$;
$w_{ij} - W(x_i, x_j)$.

В модели дискриминации должны соблюдаться следующие условия:

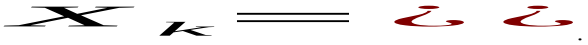
1) число групп: $g \geq 2$;
2) число объектов в каждой группе: $n_i \geq 2$;
3) число дискриминантных переменных: $0 < p < (n-2)$;
4) дискриминантные переменные измеряются в интервальной шкале;
5) дискриминантные переменные линейно независимы;
6) ковариационные матрицы групп примерно равны;
7) дискриминантные переменные в каждой группе подчиняются многомерному нормальному закону распределения.

Рассмотрим задачу максимизации отношения (2) когда имеются g групп. Оценим сначала информацию, характеризующую степень различия между объектами по всему пространству точек, определяемому переменными групп. Для этого вычислим матрицу рассеяния **T**, которая равна сумме квадратов отклонений и попарных произведений наблюдений от общих средних $\bar{x}_i,\ i=1,\ldots,p$ по каждой переменной. Элементы матрицы **T** определяются выражением [3, 4]

$$t_{ij}=\sum_{k=1}^{g}\sum_{m=1}^{n}\left(x_{ikm}-\bar{x}_i\right)\left(x_{jkm}-\bar{x}_j\right), \tag{3}$$

где $\bar{x}_i=(1/n)\sum_{k=1}^{g} n_i\bar{x}_{ik},\ i=1,\ldots,p$ $\bar{x}_{ik}=(1/n_i)\sum_{m=1}^{n_k} x_{ikm},\ i=1,\ldots,p;\ k=1,\ldots,g$;

Запишем это выражение в матричной форме. Обозначим *p*-мерную случайную векторную переменную *k*-й группы следующим образом



Тогда объединенная *p*-мерная случайная векторная переменная всех групп будет иметь вид

$$X=\left[X_1 X_2 \ldots X_g\right].$$

Общее среднее этой *p*-мерной случайной векторной переменной будет равен вектору средних отдельных признаков

$$\bar{x}=\left[\bar{x}_1 \bar{x}_2 \ldots \bar{x}_p\right].$$

Матрица рассеяния от среднего при этом запишется в виде

$$T=\sum_{k=1}^{g}\left(X_k-\bar{x}\right)\left(X_k-\bar{x}\right)'. \tag{4}$$

Если использовать векторную переменную объединенных переменных **X**, то матрица **T** определится по формуле $T=\left(X-\bar{x}\right)\left(X-\bar{x}\right)'$.

Матрица **T** содержит полную информацию о распределении точек по пространству переменных. Диагональные элементы представляют собой сумму квадратов отклонений от общего среднего и показывают как ведут себя наблюдения по отдельно взятой переменной. Внедиагональные элементы равны сумме произведений отклонений по одной переменной на отклонения по другой.

Если разделить матрицу **T** на $(n-1)$, то получим ковариационную матрицу. Для проверки условия линейной независимости переменных полезно рассмотреть вместо **T** нормированную корреляционную матрицу.

Для измерения степени разброса объектов внутри групп рассмотрим матрицу **W**, которая отличается от **T** только тем, что ее элементы определяются векторами средних для отдельных групп, а не вектором средних для общих данных. Элементы внутригруппового рассеяния определятся выражением

$$w_{ij}=\sum_{k=1}^{g}\sum_{m=1}^{n_k}(x_{ikm}-\bar{x}_{ik})(x_{jkm}-\bar{x}_{jk}) \quad (5)$$

Запишем это выражение в матричной форме. Данным g групп будут соответствовать векторы средних

$$\bar{X}_1=[\bar{x}_{11}\bar{x}_{21}\ldots\bar{x}_{p1}],$$
$$\ldots$$
$$\bar{X}_g=[\bar{x}_{1g}\bar{x}_{2g}\ldots\bar{x}_{pg}]. \quad (6)$$

Тогда матрица внутригрупповых вариаций запишется в виде

$$W=\sum_{k=1}^{g}(X_k-\bar{x}_k)(X_k-\bar{x}_k)' \quad (7)$$

Если разделить каждый элемент матрицы **W** на (n - g), то получим оценку ковариационной матрицы внутригрупповых данных.

Когда центроиды различных групп совпадают, то элементы матриц **T** и **W** будут равны. Если же центроиды групп различные, то разница

$$B=T-W \quad (8)$$

будет определять межгрупповую сумму квадратов отклонений и попарных произведений. Если расположение групп в пространстве различается (т.е. их центроиды не совпадают), то степень разброса наблюдений внутри групп будет меньше межгруппового разброса. Отметим, что элементы матрицы **B** можно вычислить и по данным средних

$$b_{ij}=\sum_{k=1}^{g}n_k(\bar{x}_{ik}-\bar{x}_i)(\bar{x}_{jk}-\bar{x}_j),\ i,\ j=1,\ldots,p \quad (9)$$

Матрицы **W** и **B** содержат всю основную информацию о зависимости внутри групп и между группами. Для лучшего разделения наблюдений на группы нужно подобрать коэффициенты дискриминантной функции из условия максимизации отношения межгрупповой матрицы рассеяния к внутригрупповой матрице рассеяния при условии ортогональности

дискриминантных плоскостей. Тогда нахождение коэффициентов дискриминантных функций сводится к решению задачи о собственных значениях и векторах [3]. Это утверждение можно сформулировать так: если спроектировать g групп *p*-мерных выборок на (g - 1) пространство, порожденное собственными векторами $(v_{1k},\ldots,v_{pk})$, $k=1,\ldots,g-1$, то отношение (2) будет максимальным, т. е. рассеивание между группами будет максимальным при заданном внутригрупповом рассеивании. Если бы мы захотели спроектировать g выборок на прямую при условии максимизации наибольшего рассеивания между группами, то следовало бы использовать собственный вектор $(v_{11},\ldots,v_{p1})$, соответствующий максимальному собственному числу λ_1. При этом дискриминантные функции можно получать: по *нестандартизованным* и *стандартизованным коэффициентам.*

Нестандартизованные коэффициенты. Пусть $\lambda_1 \geq \ldots \geq \lambda_p$ и $v_i,\ldots,v_p$ соответственно собственные значения и векторы. Тогда условие (2) в терминах собственных чисел и векторов запишется в виде

$$\lambda = \frac{\sum_k b_{jk} v_j v_k}{\sum_k w_{jk} v_j v_k},$$

что влечет равенство $\sum_k (b_{jk} - \lambda w_{jk}) v_k = 0$, или в матричной записи

$$(B - \lambda_i W) v_i = 0, \quad v_i' \mathbf{W} \mathbf{v}_j = \delta_{ij}, \qquad (10)$$

где δ_{ij} – символ Кронекера. Таким образом, решение уравнения $|B - \lambda W| = 0$ позволяет нам определить компоненты собственных векторов, соответствующих дискриминантным функциям. Если **B** и **W** невырожденные матрицы, то собственные корни уравнения $|B - \lambda W| = 0$ такие же, как и у $|W^{-1}B - \lambda I| = 0$. Решение системы уравнений (10) можно получить путем использования разложения Холецкого $L L'$ матрицы W^{-1} и решения задачи о собственных значениях

$$(L' \mathbf{B} \mathbf{L} - \lambda_i I) v_i = 0, \quad v_i' v_j = \delta_{ij}.$$

Каждое решение, которое имеет свое собственное значение λ_i и собственный вектор v_i, соответствует одной дискриминантной функции. Компоненты собственного вектора v_i можно использовать в качестве коэффициентов дискриминантной функции. Однако при таком подходе начало координат не будет совпадать с главным центроидом. Для того, что-

бы начало координат совпало с главным центроидом нужно нормировать компоненты собственного вектора [4]

$$\beta_i = v_i\sqrt{n-g}, \quad \beta_0 = -\sum_{i=1}^{p}\beta_i\bar{x}_i \,. \tag{11}$$

Нормированные коэффициенты (11) получены по нестандартизованным исходным данным, поэтому они называются *нестандартизованными.* Нормированные коэффициенты приводят к таким дискриминантным значениям, единицей измерения которых является стандартное квадратичное отклонение. При таком подходе каждая ось в преобразованном пространстве сжимается или растягивается таким образом, что соответствующее дискриминантное значение для данного объекта представляет собой число стандартных отклонений точки от главного центроида.

Стандартизованные коэффициенты можно получить двумя способами: 1) по формуле (11), если исходные данные были приведены к стандартной форме; 2) преобразованием нестандартизованных коэффициентов к стандартизованной форме:

$$c_i = \beta_i\sqrt{\frac{w_{ii}}{n-g}}, \tag{12}$$

где w_{ii}^{-} сумма внутригрупповых квадратов *i*-й переменной, определяемой по формуле (5). Стандартизованные коэффициенты полезно применять для уменьшения размерности исходного признакового пространства переменных. Если абсолютная величина коэффициента для данной переменной для всех дискриминантных функций мала, то эту переменную можно исключить, тем самым сократив число переменных.

Структурные коэффициенты определяются коэффициентами взаимной корреляции между отдельными переменными и дискриминантной функцией. Если относительно некоторой переменной абсолютная величина коэффициента велика, то вся информация о дискриминантной функции заключена в этой переменной.

Структурные коэффициенты полезны при классификации групп. Структурный коэффициент можно вычислить и для переменной в пределах отдельно взятой группы. Тогда получаем *внутригрупповой структурный коэффициент*, который вычисляется по формуле

$$s_{ij} = \sum_{k=1}^{p} r_{ik}c_{kj} = \sum_{k=1}^{p}\frac{w_{ik}c_{kj}}{\sqrt{w_{ii}w_{jj}}}, \tag{13}$$

где s_{ij} — внутригрупповой структурный коэффициент для *i*-й переменной и *j*-й функции; r_{ik} — внутригрупповые структурные коэффициенты корреля-ции между переменными *i* и *k*; c_{kj} — стандартизованные коэффициенты канонической функции для переменной *k* и функции *j*.

Структурные коэффициенты по своей информативности несколько отличаются от стандартизованных коэффициентов. Стандартизованные коэффициенты показывают вклад переменных в значение дискриминантной функции. Если две переменные сильно коррелированы, то их стандартизованные коэффициенты могут быть меньше по сравнению с теми случаями, когда используется только одна из этих переменных. Такое распределение величины стандартизованного коэффициента объясняется тем, что при их вычислении учитывается влияние всех переменных. Структурные же коэффициенты являются парными корреляциями и на них не влияют взаимные зависимости прочих переменных.

1.2. Число дискриминантных функций

Общее число дискриминантных функций не превышает числа дискриминантных переменных и, по крайней мере, на единицу меньше числа групп. Степень разделения выборочных групп зависит от величины собственных чисел: чем больше собственное число, тем сильнее разделение. Наибольшей разделительной способностью обладает первая дискриминантная функция, соответствующая наибольшему собственному числу λ_1, вторая обеспечивает максимальное различение после первой и т. д. Различительную способность *i*-й функции оценивают по относительной величине в процентах собственного числа λ_i от суммы всех λ.

Коэффициент канонической корреляции. Другой характеристикой, позволяющей оценить полезность дискриминантной функции является коэффициент канонической корреляции r_i. Каноническая корреляция является мерой связи между двумя множествами переменных. Максимальная величина этого коэффициента равна 1. Будем считать, что группы составляют одно множество, а другое множество образуют дискриминантные переменные. Коэффициент канонической корреляции для *i*-й дискриминантной функции определяется формулой:

$$r_i = \sqrt{\frac{\lambda_i}{1+\lambda_i}}. \qquad (14)$$

Чем больше величина r_i, тем лучше разделительная способность дискриминантной функции.

Остаточная дискриминация. Так как дискриминантные функции находятся по выборочным данным, они нуждаются в проверке статистической значимости. Дискриминантные функции представляются аналогично главным компонентам. Поэтому для проверки этой значимости можно воспользоваться критерием, аналогичным дисперсионному критерию в методе главных компонент. Этот критерий оценивает остаточную дискриминантную способность, под которой понимается способность различать группы, если при этом исключить информацию, полученную с помощью ранее вычисленных функций. Если остаточная дискриминация мала, то не имеет смысла дальнейшее вычисление очередной дискриминантной функции. Полученная статистика носит название «$\Lambda-$*статистики Уилкса*» и вычисляется по формуле:

$$\Lambda=\prod_{i=k+1}^{g}(1/(1+\lambda_i)), \tag{15}$$

где k – число вычисленных функций. Чем меньше эта статистика, тем значимее соответствующая дискриминантная функция. Величина

$$\chi^2=-[n-((p+g)/2)-1]\ln\Lambda_k,\ k=0,1,\ldots,g-1 \tag{16}$$

имеет хи-квадрат распределение с $(p-k)(g-k-1)$ степенями свободы.

Вычисления проводим в следующем порядке.

1. Находим значение критерия χ^2 при $k = 0$. Значимость критерия подтверждает существование различий между группами. Кроме того, это доказывает, что первая дискриминантная функция значима и имеет смысл ее вычислять.

2. Определяем первую дискриминантную функцию, и проверяем значимость критерия при $k = 1$. Если критерий значим, то вычисляем вторую дискриминантную функцию и продолжаем процесс до тех пор, пока не будет исчерпана вся значимая информация.

2. КЛАССИФИЦИРУЮЩИЕ ФУНКЦИИ

До сих пор мы рассматривали получение канонических дискриминантных функций при известной принадлежности объектов к тому или иному классу. Основное внимание уделялось определению числа и значимости этих функций, и использованию их для объяснения различий между классами. Все сказанное относилось к интерпретации результатов ДА. Однако наибольший интерес представляет задача предсказания класса, которому принадлежит некоторый случайно выбранный объект. Эту задачу

можно решить, используя информацию, содержащуюся в дискриминантных переменных. Существуют различные способы классификации.

В процедурах классификации могут использоваться как сами дискриминантные переменные так и канонические дискриминантные функции. В первом случае применяется метод максимизации различий между классами для получения функции классификации, различие же классов на значимость не проверяется и, следовательно, дискриминантный анализ не проводится. Во втором случае для классификации используются непосредственно дискриминантные функции и проводится более глубокий анализ.

2.1. Применение элементарных классифицирующих функций

Рассмотрим случай отнесения случайно выбранного объекта $x=(x_1 \ldots x_p)'$ к одной из групп G_k, $k=1,\ldots,g$, $k\geq 2$. Пусть $f_k(x)$ плотность распределения **x** в G_k и q_k- априорная вероятность того, что вектор **x** принадлежит к группе G_k. Предполагается, что сумма априорных вероятностей $\sum_{k=1}^{g} q_k$ равна 1.

Определим условную вероятность $\Pr(x|G_k)$ получения некоторого вектора **x**, если известно, что объект принадлежит к группе G_k, $k=1,\ldots, g$. Обозначим через $\Pr(G_k|x)$ условную вероятность принадлежности объекта к группе G_k при заданном **x**. Величины $\Pr(x|G_k)$ и $\Pr(G_k|x)$ называются апостериорными вероятностями. Различие между априорными и апостериорными вероятностями заключается в следующем. Априорная вероятность q_k равна вероятности принадлежности объекта к данной группе G_k до получения вектора наблюдений **x**. Апостериорная вероятность $\Pr(G_k|x)$ определяет вероятность принадлежности объекта к группе G_k только после анализа вектора наблюдений **x** этого объекта.

Из *теоремы Байеса* получаем

$$\Pr(G_k|x)=\frac{q_k \Pr(x|G_k)}{\sum_{j=1}^{g} q_j \Pr(x|G_j)}. \qquad (17)$$

Выражение (17) справедливо для любого распределения вектора **x**. Байесовская процедура минимизирует ожидаемую вероятность ошибочной классификации $¿\ j\neq k\ ¿¿¿¿\ g$.

Так, например, для двух групп получим $q_1 \Pr(2|1)+q_2 \Pr(1|2)$.

Эта величина является вероятностью того, что объект, принадлежащий к группе G_1, ошибочно классифицируется, как принадлежащий G_2, или наоборот, объект из G_2 ошибочно относится к G_1.

Если **x** имеет *p*-мерный нормальный закон распределения $N(\mu_k^{p\times 1}, \Sigma^{p\times p})$, то вероятности $\Pr(x|G_k), k=1,\ldots, g$ можно заменить соответственно на плотности распределений $f_k(x),\ k=1,\ldots, g$. В результате получим

$$\Pr(G_k|x)=\frac{q_k f_k(x)}{\sum_{j=1}^{g} q_j f_j(x)},\ k=1,\ldots, g \quad . \tag{18}$$

Байесовская процедура классификации состоит в том, что вектор наблюдений **x** относится к группе G_k, если $\Pr(G_k|x)$ имеет наибольшее значение.

Можно показать, что байесовская процедура эквивалентна отнесению вектора **x** к группе G_k, если оценочная функция

$$\delta_k(x)=q_k f_k(x) \tag{19}$$

является максимальной. Подставим в оценочную функцию (19) формулу нормального закона распределения

$$\delta_k(x)=q_k(2\pi)^{-2p}|\Sigma|^{-\frac{1}{2}}\exp[-\frac{1}{2}(x-\mu_k)'\Sigma^{-1}(x-\mu_k)] \quad .$$

Удаляя общую константу $(2\pi)^{-2p}|\Sigma|^{-\frac{1}{2}}$ и логарифмируя, получим

$$d_k(x)=-\frac{1}{2}(x-\mu_k)'\Sigma^{-1}(x-\mu_k)+\ln q_k \quad . \tag{20}$$

Преобразуем выражение (20)

$$d_k=-\frac{1}{2}x'\Sigma^{-1}x+\frac{1}{2}x'\Sigma^{-1}\mu_k+\frac{1}{2}\mu_k'\Sigma^{-1}x-\frac{1}{2}\mu_k'\Sigma^{-1}\mu_k+\ln q_k$$

и, удалив постоянную $-\frac{1}{2}x'\Sigma^{-1}x$, получим

$$d_k=\mu_k'\Sigma^{-1}x-\frac{1}{2}\mu_k'\Sigma^{-1}\mu_k+\ln q_k,\ \ k=1,\ldots, g \quad .$$

Заменим векторы средних и ковариационную матрицу их оценками $\bar{x}_k=(\bar{x}_{k1}\dots\bar{x}_{kp}),\ k=1,\dots,g$ и **Σ**. Тогда получим *классифицирующую функцию* вида

$$d_k=\bar{x}_k^{'}\Sigma^{-1}x-\frac{1}{2}\bar{x}_k^{'}\Sigma^{-1}\bar{x}_k+\ln q_k \quad . \tag{21}$$

Введем обозначения $b_k=\bar{x}_k^{'}\Sigma^{-1}$ и $b_{k0}=-\frac{1}{2}\bar{x}_k^{'}\Sigma^{-1}\bar{x}_k$, $k=1,\dots,g$, где $b_k=(b_{k1},\dots,b_{kp})$ и b_{k0} – коэффициенты *k*-й классифицирующей функции *i*-го объекта (простой дискриминантной функции Фишера)

$$d_{ik}=b_{k0}+b_{k1}x_{i_1}+\dots+b_{kp}x_{i_p}+\ln q_k,\ k=1,\dots,\ g \quad . \tag{22}$$

Объект $x_i=(x_{i1}\dots x_{ip})$ относится к классу, у которого значение *d* оказывается наибольшим. Коэффициенты классифицирующих функций удобнее вычислять по скалярным выражениям

$$b_{ki}=(n-g)\sum_{j=1}^{p}(w^{-1})_{ij}\bar{x}_{jk},\ k=1,\dots,\ g \quad , \tag{23}$$

где $b_{ki}-$ коэффициент для переменной *i* в выражении, соответствующему классу *k*, $(w^{-1})_{ij}-$ обратный элемент внутригрупповой матрицы сумм попарных произведений **W**. Постоянный член находится по формуле

$$b_{k0}=-0{,}5\cdot\sum_{j=1}^{p}b_{kj}\bar{x}_{jk},\ k=1,\dots,\ g \quad . \tag{24}$$

Функции, определяемые соотношением (22), называются «*простыми классифицирующими функциями*» потому, что они предполагают лишь равенство групповых ковариационных матриц и не требуют других дополнительных свойств.

2.2. Классификация объектов с помощью функции расстояния

Выбор функций расстояния между объектами для классификации является наиболее очевидным способом введения меры сходства для векторов объектов, которые интерпретируются как точки в евклидовом пространстве. В качестве меры сходства можно использовать евклидово расстояние между объектами. Чем меньше расстояние между объектами, тем больше сходство. Однако в тех случаях, когда переменные коррелированы, измерены в разных единицах и имеют различные стандартные отклонения, трудно четко определить понятие «расстояния». В этом случае по-

лезнее применить не евклидовое расстояние, а *выборочное расстояние Махаланобиса*

$$D^2(x/G_k)=(n-g)\cdot\sum_{v=1}^{p}\sum_{j=1}^{p}(w^{-1})_{vj}(x_{iv}-\bar{x}_{vk})(x_{ij}-\bar{x}_{jk}),\ k=1,\ldots,g \quad (25)$$

или в матричной записи

$$D^2(x/G_k)=(n-g)\cdot(x-\bar{x}_k)'W^{-1}(x-\bar{x}_k),\ k=1,\ldots,g, \quad (25')$$

где **x** представляет объект с *p* переменными, $\bar{x}_k$-вектор средних для переменных *k*-й группы объектов. Если вместо W использовать оценку внутригрупповой ковариационной матрицы $\Sigma=W/(n-g)$, то получим стандартную запись выборочного расстояния Маханалобиса

$$D^2(x/G_k)=(x-\bar{x}_k)'\Sigma^{-1}(x-\bar{x}_k),\ k=1,\ldots,g. \quad (26)$$

При использовании функции расстояния, объект относят к той группе, для которой расстояние D^2 наименьшее.

Относя объект к ближайшему классу в соответствии с D^2, мы неявно приписываем его к тому классу, для которого он имеет наибольшую вероятность принадлежности $\Pr(x|G_k)$. Если предположить, что любой объект должен принадлежать одной из групп, то можно вычислить вероятность его принадлежности для любой из групп

$$\Pr(G_k|x)=\frac{\Pr(x|G_k)}{\sum_{i=1}^{g}\Pr(x|G_i)}. \quad (27)$$

Объект принадлежит к той группе, для которой апостериорная вероятность $\Pr(x|G_k)$ максимальна, что эквивалентно использованию наименьшего расстояния.

До сих пор при классификации по D^2 предполагалось, что априорные вероятности появления групп одинаковы. Для учета априорных вероятностей нужно модифицировать расстояние D^2, вычитая из выражений (25)–(26) удвоенную величину натурального логарифма от априорной вероятности q_k. Тогда, вместо выборочного расстояния Махаланобиса (26), получим

$$D_q^2(x/G_k)=(x-\bar{x}_k)'\Sigma^{-1}(x-\bar{x}_k)-2\ln(q_k). \quad (28)$$

Это изменение расстояния математически идентично умножению величин $f_k(x)$ или $\Pr(x|G_k)$ на априорную вероятность группы q_k. Форму-

лу (28) можно получить, умножив правые и левые части выражения (20) на два. Тогда после замены векторов средних и ковариационной матрицы их оценками имеем $D_q^2(x/G_k)=(x-\bar{x}_k)'\Sigma^{-1}(x-\bar{x}_k)-2\ln(q_k)$.

Отметим, тот факт, что априорные вероятности оказывают наибольшее влияние при перекрытии групп и, следовательно, многие объекты с большой вероятностью могут принадлежать ко многим группам. Если группы сильно различаются, то учет априорных вероятностей практически не влияет на результат классификации, поскольку между классами будет находиться очень мало объектов.

V-статистика Рао. В некоторых работах для классификации используется *обобщенное расстояние Махаланобиса V* – обобщение величины D^2. Эта мера, известная как V-статистика Рао, измеряет расстояния от каждого центроида группы до главного центроида с весами, пропорциональными объему выборки соответствующей группы. Она применима при любом количестве классов и может быть использована для проверки гипотезы $H_0: \mu_1=\cdots=\mu_g$. Если гипотеза H_0 верна, а объемы выборок n_i стремятся к ∞, то распределение величины V стремится к χ^2 с $p(g-1)$ степенями свободы. Если наблюдаемая величина $\chi^2>\chi^2_{1-\alpha}(p(g-1))$, то гипотеза H_0 отвергается. *V*-статистика вычисляется по формуле

$$V=(n-g)\sum_{i=1}^{p}\sum_{j=1}^{p}(w^{-1})_{ij}\sum_{k=1}^{g}n_k(\bar{x}_{ik}-\bar{x}_i)(\bar{x}_{jk}-\bar{x}_j). \quad (29)$$

Матричное выражение оценки V имеет вид

$$V=\sum_{k=1}^{g}n_k(\bar{x}_k-\bar{x})'\Sigma^{-1}(\bar{x}_k-\bar{x}). \quad (30)$$

Отметим, что при включении или исключении переменных V-статистика имеет распределение хи-квадрат с числом степеней свободы, равным (g - 1), умноженное на число переменных, включенных (исключенных) на этом шаге. Если изменение статистики не значимо, то переменную можно не включать. Если после включения новой переменной V-статистика оказывается отрицательной, то это означает, что включенная переменная ухудшает разделение центроидов.

2.3. Классификационная матрица

В дискриминантном анализе процедура классификации используется для определения принадлежности к той или иной группе случайно выбранных объектов, которые не были включены при выводе дискриминантной и классифицирующих функций. Для проверки точности классификации применим классифицирующие функции к тем объектам, по которым они были

получены. По доле правильно классифицированных объектов можно оценить точность процедуры классификации. Результаты такой классификации представляют в виде *классификационной матрицы*. Рассмотрим пример классификационной матрицы, приведенной в табл. 1.

Таблица 1

КЛАССИФИКАЦИОННАЯ МАТРИЦА

Группы	Предсказанные группы (число, процент)								
	1		2		3		4		Всего
1	9	90.0	0	0.0	0	0.0	1	10.0	10
2	0	0.0	4	80.0	1	20.0	0	0.0	5
3	8	14.8	4	7.4	37	68.5	5	9.3	54
4	1	7.7	0	0.0	1	7.7	11	84.6	13

В первой группе точно предсказаны из 10 объектов 9, что составляет 90 %, один объект отнесен к 4-й группе. Во второй группе правильно предсказаны 80 % объектов, один объект (20 %) отнесен к третьей группе. В третьей группе процент правильного предсказания самый низкий и составляет 68,5 %, причем из 54 объектов 8 отнесены к первой группе, 4 – ко второй и 5 – к четвертой группе. В четвертой группе правильно предсказаны 84,6%, по одному объекту отнесено к первой и третьей группам.

Процент правильной классификации объектов является дополнительной мерой различий между группами и ее можно считать наиболее подходящей мерой дискриминации. Следует отметить, что величина процентного содержания пригодна для суждения о правильном предсказании только тогда, когда распределение объектов по группам производилось случайно. Например, для двух групп при случайной классификации можно правильно предсказать 50 %, а для четырех групп эта величина составляет 25 %. Поэтому если для двух групп имеем 60 % правильного предсказания, то нужно считать эту величину слишком малой, тогда как для четырех групп эта величина говорит о хорошей разделительной способности.

Пример. Больные увеличением щитовидной железы общим числом 23 человека были разделены на три группы.

Группа 1. Лечение оказалось успешным; проведенное через большой промежуток времени клиническое обследование показало, что пациент здоров.

Группа 2. Лечение безуспешно, т. е. состояние больного осталось без изменения.

Группа 3. Исход лечения успешен, но в дальнейшем возможен рецидив.

По результатам обследования 23 пациентов имеются следующие измерения:

у6 – йод, регистрируемый через 3 часа после принятия испытательной дозы;

у9 – йод, регистрируемый через 48 часов после принятия испытательной дозы;

у10 – содержание в крови белковосвязанного йода ($PB^{131}J$) через 48 часов;

kl – номер группы.

Конкретные результаты приведены в табл.2.

Таблица 2

ДАННЫЕ О 23 БОЛЬНЫХ ГИПЕРТИРЕОЗОМ, РАЗДЕЛЕННЫЗ НА ТРИ ГРУППЫ

№	Kl	у6	у9	у10	№	kl	у6	у9	Y10
1	1	14.4	25.1	0.20	13	1	54.0	57.0	0.19
2	1	20.1	40.1	0.11	14	1	16.1	20.6	0.22
3	1	24.1	32.1	0.17	15	1	57.5	74.5	0.49
4	1	11.1	16.9	0.12	16	1	37.8	63.0	0.32
5	1	16.3	32.1	0.36	17	2	55.8	48.0	2.74
6	1	40.5	64.4	0.21	18	2	75.0	60.0	1.37
7	1	52.7	50.0	0.53	19	2	72.0	65.0	0.70
8	1	20.8	22.3	0.13	20	2	70.6	45.0	1.40
9	1	14.0	3.1	0.18	21	3	24.1	45.0	0.22
10	1	27.0	41.7	0.19	22	3	33.2	55.0	0.01
11	1	44.3	63.8	0.22	23	3	30.4	44.6	0.09
12	1	47.5	50.1	0.29					

По матрице исходных данных находятся средние и стандартные отклонения дискриминантных переменных (табл. 3, 4), общая **T** и внутригрупповые **W** матрицы сумм квадратов и перекрестных произведений (табл. 5, 6).

Таблица 3

СРЕДНИЕ ДИСКРИМИНАНТНЫХ ПЕРЕМЕННЫХ $\overline{X}_{jk}$

Группы GR	Y6	Y9	Y10	Кол-во
1 ($\overline{X}_{i1}$)	31,1375	41,0500	0,2456	16
2 ($\overline{X}_{i2}$)	68,3500	54,5000	1,5525	4
3 ($\overline{X}_{i3}$)	29,2333	48,2000	0,1067	3
Все группы $\overline{X}_{i}$)	37,3609	44,3217	0,4548	23

Таблица 4

СТАНДАРТНЫЕ ОТКЛОНЕНИЯ S_{jk}

Группы GR	Y6	Y9	Y10	Кол-во
1. (S_{j1})	16,2739	20,4760	0,1237	16
2. (S_{j2})	8,5656	9,5394	0,8551	4
3. (S_{j3})	4,6608	5,8924	0,1060	3
Все группы (S_{j})				23

Таблица 5

МАТРИЦА ОБЩЕЙ СУММЫ ПЕРЕКРЕСТНЫХ ПРОИЗВЕДЕНИЙ T

Переменная	Y6	Y9	Y10
Y6	8895,3148	6025,1896	163,2293
Y9	6025,1896	7262,2391	53,5466
Y10	163,2293	53,5466	8,3290

Таблица 6

МАТРИЦА ВНУТРИГРУППОВОЙ СУММЫ ПЕРЕКРЕСТНЫХ ПРОИЗВЕДЕНИЙ W

Переменная	Y6	Y9	Y10
Y6	4236,1542	4532,3100	–2,1545
Y9	4532,3100	6631,4600	1,9565
Y10	–2,1545	1,9565	2,4455

Если разделить каждый элемент **T** на (n - 1)), а каждый элемент **W** – на (n – g), то получим ковариационные матрицы. Для оценки меры связи между дискриминантными переменными матрицы **T** и **W** преобразованы в корреляционные матрицы, которые приведены в табл. 7 и 8. Элементы этих матриц найдены по формулам $r_{ij}^{(t)} = \frac{T_{ij}}{(n-1)S_i S_j}$ и $r_{ij}^{(w)} = \frac{W_{ij}}{(n-q)S_i S_j}$.

Из общей корреляционной матрицы видно, что переменные некоррелированы на уровне 0.01. Отсюда следует, что ни одна переменная не может быть предсказана по значению, соответствующему другой переменной.

Таблица 7

ОБЩАЯ КОРРЕЛЯЦИОННАЯ МАТРИЦА

Переменная	Y6	Y9	Y10
Y6	1,0000	-0,1759	0,0664
Y9	-0,1759	1,0000	0,3480
Y10	0,0664	0,3480	1,0000

Для измерения меры разброса наблюдений внутри классов используется внутригрупповая корреляционная матрица, которая приведена в табл. 8. Эта матрица не совпадает с общей корреляционной матрицей. Из таблицы видно, что многие коэффициенты отличаются от значений, приведенных в табл.7.

Таблица 8

ВНУТРИГРУППОВАЯ КОРРЕЛЯЦИОННАЯ МАТРИЦА

Переменная	Y6	Y9	Y10
Y6	1,0000	0,8551	–0,0212
Y9	0,8551	1,0000	0,0154
Y10	–0,0212	0,0154	1,00

Из табл. 5 и 6 видно, что большая часть элементов матрицы **W** меньше соответствующих элементов матрицы **T**. Разница этих матриц

$B=T-W$ определяет межгрупповую сумму квадратов отклонений и попарных произведений. Эта матрица приведена в табл. 9.

Таблица 9

МАТРИЦА МЕЖГРУППОВОЙ СУММЫ ПЕРЕКРЕСТНЫХ ПРОИЗВЕДЕНИЙ B

Переменная	Y6	Y9	Y10
Y6	4659,1606	1492,8796	165,3838
Y9	1492,8796	630,7791	51,5901
Y10	165,3838	51,5901	5,8834

Для нахождения коэффициентов канонической дискриминантной функции решаем задачу (2) в терминах собственных чисел и векторов, которая в матричной записи имеет вид (10). Систему уравнений (10) решаем с помощью разложения Холецкого матрицы $W^{-1}=L\,L'$,
$(L'\mathbf{B}L-\lambda_i I)v_i=0,\ \ v_i'v_j=\delta_{ij}$.

$$L=\begin{bmatrix} 0,0297 & 0 & 0 \\ -0,0203 & 0,0123 & 0 \\ 0,0424 & -0,0098 & 0,6395 \end{bmatrix}$$

Наибольшее собственное значение для системы равно $\lambda_1=5,3514$ и $\lambda_3=0,0452$, которым соответствуют собственные векторы $v_1=\begin{bmatrix}0,7360 & 0,0990 & 0,6697\end{bmatrix}'$ и $v_3=\begin{bmatrix}-0,4368 & 0,8252 & 0,3581\end{bmatrix}'$. Положив $b=\mathbf{Lv}$, получаем коэффициенты канонической дискриминантной функции $b_1=\begin{bmatrix}0,0219 & -0,0137 & 0,4585\end{bmatrix}$ и $b_3=\begin{bmatrix}-0,0130 & 0,0190 & 0,2024\end{bmatrix}$.

При использовании коэффициентов **b** начало координат не будет совпадать с главным центроидом. Для того чтобы начало координат совпало с главным центроидом нужно нормировать компоненты вектора **b**, используя формулы (11). Для оценки относительного вклада каждой переменной в значение дискриминантной функции вычислим стандартизованные дискриминантные коэффициенты по формуле (12). Результаты вычислений приведены в табл.10. Из табл.10 видно, что две наиболее значимо коррелированные переменные Y6 и Y9 имеют примерно одинаковые стандартизованные коэффициенты. Значения нестандартизованной канонической функции для каждого пациента сведены в табл.15. Координаты центроидов первой, второй и третьей групп соответственно равны:
$\begin{bmatrix}-0,8363 & 4,6553 & -1,7466 \\ -0,1063 & 0,0604 & 0,4862\end{bmatrix}$.

Таблица 10

КОЭФФИЦИЕНТЫ ДИСКРИМИНАНТНОЙ ФУНКЦИИ

Нестандартизованные коэффициенты		Стандартизованные Коэффициенты	
Переменная	Коэффициенты	Переменная	Коэффициенты

Y6	0,0978	-0,0580	Y6	1,4228	-0,8445
Y9	-0,0614	0,0850	Y9	-1,1184	1,5479
Y10	2,0504	0,9050	Y10	0,7170	0,33165
Константа	-1,8628	-0,20112	Собств. нач.	5,3514	0,0452

Для определения взаимной зависимости отдельной переменной и дискриминантной функции рассмотрим внутригрупповые структурные коэффициенты, значения которых находим по формуле (13). Результаты вычислений представлены в табл. 11.

Таблица 11

ВНУТРИГРУППОВЫЕ СТРУКТУРНЫЕ КОЭФФИЦИЕНТЫ

Переменная	Коэффициент	
Y6	1,4580	-0,8653
Y9	-1,1460	1,5861
Y10	0,7347	0,3243

Переменные Y6 и Y9 имеют небольшие структурные коэффициенты, но у них относительно большие стандартизованные коэффицинты. Это объясняется значимой корреляцией переменной Y6 с другими переменными и может оказаться, что вклад переменных Y6 и Y9 в дискриминантые значения невелик. Для оценки реальной полезности канонической дискриминантной функции вычисляем по формулам (14)–(16) коэффициент канонической корреляции, Λ-статистику Уилкса , статистику хи-квадрат, уровень значимости. Результаты вычислений приведены в табл. 12.

Таблица 12

ОСНОВНЫЕ СТАТИСТИКИ

Дискриминантная функция	Собственное значение	Каноническая корреляция R	Λ-статистика Уилкса	Статистика хи-квадрат	Степень свободы	Уровень значимости.
1	5,3514	0,9179	0,1506	35,9655	6	$4{,}076\ 10^{-6}$
2	0,0452	0,2080	0,9567	0,8405	2	0,6569

Данные таблицы указывают на хорошую дискриминацию групп: большая величина канонической корреляции соответствует тесной связи дискриминантной функции с группами; малая величина Λ-статистики Уилкса означает, что четыре используемых переменных эффективно участвуют в различении групп и, наконец, статистика хи-квадрат значима с уровнем $1{,}6\ 10^{-8}$.

Процедура классификации. Процедуры классификации могут использовать канонические дискриминантные функции или сами дискриминантные переменные. Для классификации с помощью дискриминантных переменных коэффициенты классифицирующей функции вычисляем по формуле (22). Результаты вычислений приведены в табл. 13. Значения классифицирующей функции для каждого больного вычислены по формуле (21), результаты классификации в виде классификационной матри-

цы представлены в табл. 14. Так как процент правильной классификации составляет 100 %, то таблицу классифицирующих функций для отдельных пациентов можно не представлять.

Таблица 13

Коэффициенты классифицирующих функций

Переменная	Группа 1	Группа 2	Группа 3
Y6	0,0603	0,5875	–0,0631
Y9	0,0820	–2,4110	0,1883
Y10	1,9962	13,4071	0,6661
Константа	–2,8760	–23,9141	–3,6512

Таблица 14

Классификационная матрица

Группы	Предсказанные группы (число, процент)						
	1		2		3		Всего
1	10	62,50	0	0,0	6	37,50	16
2	0	0.00	4	100,00	0	00.0	4
3	0	0,00	0	0,00	3	100,00	3

Результаты классификации с помощью расстояния Махаланобиса (формулы (25), (26)) и апостериорной вероятности принадлежности к группе в предположении нормальности распределения (формула 19) приведены в табл. 15.

Таблица 15

Сводка результатов классификации

№ больного	Нестандартизованные канонические функции d_i			Квадрат расстояния Махаланобиса $D^2(\boldsymbol{x}/G_k)$		
	Группа	Значение		Группа 1	Группа 2	Группа 3
1	1	–1,6258	–0,5453	1,3941	39,9613	1,7126
2	1	–2,1879	0,3389	2,1281	46,4330	0,4254
3	1	–1,1576	–0,5402	0,3037	33,8515	1,4480
4	1	–1,6083	–1,1376	2,1155	40,6888	3,1499
5	1	–1,5398	0,0998	1,6444	39,0807	1,3698
6	1	–1,4635	1,3352	2,4410	38,6575	0,8729
7	1	–1,3373	–0,3477	5,3223	12,0657	10,6765
8	1	–1,2347	–0,9555	1,2544	32,8613	3,5611
9	1	–2,4564	–0,3223	5,7100	30,9378	10,5528
10	1	0,1421	–1,4293	0,4101	36,6478	0,2827
11	1	1,0663	–1,0241	1,6739	33,2676	1,1976
12	1	–0,2524	0,3058	0,1102	19,8784	5,5216
13	1	–0,1306	0,3126	3,2852	20,941	6,5678
14	1	–1,0198	–1,1302	1,2853	34,6955	3,0330
15	1	1,4639	0,1921	4,0840	22,5124	5,3097

16	1	1,4759	–1,4148	2,6895	38,3378	1,0454
17	2	1,3432	6,4170	60,6784	12,4824	73,1019
18	2	–0,0236	4,7068	29,9684	0,4904	40,9341
19	2	–0,0311	2,6839	14,5114	6,785	21,8918
20	2	–1,0408	5,2731	36,9560	1,7390	50,1042
21	3	0,6296	–1,8645	1,7390	42,4824	0,2744
22	3	0,7651	–2,0234	2,1344	44,5377	0,2310
23	3	0,0998	–1,4813	0,4413	37,2501	0,2704

Практическая компьютерная реализация метода дискриминантного анализа с помощью языка статистического моделирования R.

В языке R линейный дискриминантный анализ кроме предположения о нормальности распределения данных в каждом классе, выдвигает предположение о статистическом равенстве внутригрупповых матриц дисперсий и корреляций. Если между ними нет серьезных отличий, их объединяют в расчетную ковариационную матрицу.

Для проверки гипотезы о многомерном нормальном распределении данных используется многомерная версия критерия согласия Шапиро-Уилка, которая реализована в функции mshapiro.test() из пакета mvnormtest. На вход этой функции подается матрица, строки которой соответствуют переменным, а столбцы – наблюдениям:

```
DGlass <- read.table(file = "Glass.txt", sep = ",",
                     header = TRUE, row.names = 1)
DGlass$FAC <- as.factor(ifelse(DGlass$Class == 2, 2, 1))
library(mvnormtest)
mshapiro.test(t(DGlass[DGlass$FAC == 1, 1:9]))
```

Для проверки гипотезы о гомогенности матриц ковариаций используется так называемый М-критерий Бокса, который реализован в функции boxM() из пакета biotools:

```
library(biotools)
## ---
## biotools version 3.0
```

```
boxM(as.matrix(DGlass[, 1:9]), DGlass$FAC)
```

Дискриминантный анализ реализован в нескольких пакетах для R, рассмотрим применение функции lda() из базового пакета MASS. Поскольку важной характеристикой прогнозирующей эффективности модели является ее ошибка при перекрестной проверке, то в функции lda() пакета MASS заложена реализация скользящего контроля (leave-one-out CV).

Составим предварительно функцию, которая по построенной модели выводит нам важные показатели для оценки ее качества: матрицы неточностей на обучающей выборке и при перекрестной проверке, ошибку распознавания и расстояние Махалонобиса между центроидами двух классов.

```
# Функция вывода результатов классификации
Out_CTab <- function(model, group, type = "lda") {
   # Таблица неточностей "Факт/Прогноз" по обучающей выборке
   classified <- predict(model)$class
   t1 <- table(group, classified)
   # Точность классификации и расстояние Махалонобиса
   Err_S <- mean(group != classified)
   mahDist <- NA
   if (type == "lda")
   { mahDist <- dist(model$means %*% model$scaling) }
   # Таблица "Факт/Прогноз" и ошибка при скользящем контроле
   t2 <-  table(group, update(model, CV = T)$class -> LDA.cv)
   Err_CV <- mean(group != LDA.cv)
   Err_S.MahD <- c(Err_S, mahDist)
   Err_CV.N <- c(Err_CV, length(group))
   cbind(t1, Err_S.MahD, t2, Err_CV.N)
}
# --- Выполнение расчетов
library(MASS)
lda.all <- lda(FAC ~ ., data = DGlass[, -10])
Out_CTab(lda.all, DGlass$FAC)
```

Отметим существенный рост ошибки распознавания до 31% при выполнении скользящего контроля. Естественно задаться вопросом, какие из имеющихся 9 признаков являются информативными при разделении, а какие - сопутствующим балластом. Шаговая процедура выбора переменных при классификации, реализованная функцией stepclass() из пакета klaR, основана на вычислении сразу четырех параметров качества моделей-претендентов: а) индекса ошибок (correctness

rate), б) точности (accuracy), основанной на евклидовых расстояниях между векторами “факта” и “прогноза”, в) способности к разделимости (ability to seperate), также основанной на расстояниях, и г) доверительных интервалах центроидов классов. Все эти параметры оцениваются в режиме многократной перекрестной проверки.

```
library(klaR)
stepclass(FAC ~ ., data = DGlass[, -10], method = "lda")
## correctness rate: 0.70515;  in: "Al";  variables (1): Al
## correctness rate: 0.77978;  in: "Ca";  variables (2): Al, Ca
##
##  hr.elapsed min.elapsed sec.elapsed
##        0.00        0.00        0.72
## method      : lda
## final model : FAC ~ Al + Ca
## <environment: 0x000000001761a930>
##
## correctness rate = 0.7798
lda.step <- lda(FAC ~ Mg + Al, data = DGlass[, -10])
```

В результате получили компактную дискриминантную функцию

$$z(x)=2.69Al−0.83Mg,z(x)=2.69Al−0.83Mg,$$

зависящую только от двух переменных. Найдем ошибку предсказания как на обучающей выборке, так и при скользящем контроле:

```
partimat(FAC ~ Mg + Al, data = DGlass[, -10], main = '', method = "lda")
```

Вместо функции `stepclass()` из пакета `klaR` для выбора оптимального набора предикторов можно воспользоваться функцией `rfe()` из пакета `caret` (процедура рекурсивного исключения - см. раздел 4.1):

```
ldaProfile <- rfe(DGlass[, 1:9], DGlass$FAC,
sizes = 2:9,
                  rfeControl = rfeControl(func-
tions = ldaFuncs,
```

```
                                                    method
= "repeatedcv", repeats = 6))
```

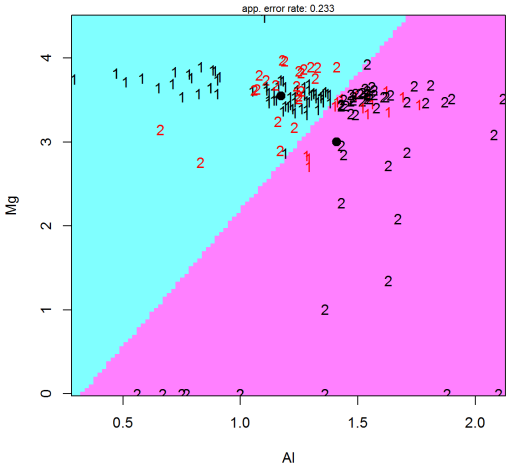
Для того чтобы уточнить, какую из трех построенных моделей следует предпочесть, выполним их тестирование на основе 10-кратной перекрестной проверки с 5 повторами:

```
DGlass$FAC <- as.factor(ifelse(DGlass$Class == 2,
"C2", "C1"))
# Модель на основе всех 9 предикторов
lda.full.pro <- train(DGlass[, 1:9], DGlass$FAC,
                      data = DGlass, method =
"lda",
                      trControl =
trainControl(method = "repeatedcv", repeats = 5,

classProbs = TRUE), metric = "Accuracy")
# Модель на основе 2 предикторов stepclass
lda.step.pro <- train(FAC ~ Mg + Al, data =
DGlass, method = "lda",
                      trControl =
trainControl(method = "repeatedcv", repeats = 5,

classProbs = TRUE), metric = "Accuracy")
# Модель на основе 3 предикторов rfe
lda.rfe.pro <- train(FAC ~ Al + K + Fe,
                     data = DGlass, method =
"lda",
                     trControl =
trainControl(method = "repeatedcv", repeats = 5,

classProbs = TRUE), metric = "Accuracy")
plot(varImp(lda.full.pro))
```



Функция `rfe()` провела отбор переменных в полном соответствии с рейтингом их важности, но это решение оказалось неоптимальным. Модель `lda.step`, полученная с использованием функции `stepclass()`, оказалась существенно эффективней.

Также в состав языка R входит возможность применить наивный байесовский классификатор.

library("quantmod") # *Для скачивания котировок.*

library("lubridate") # *Для работы с датами.*

library("e1071") # *Алгоритмы машинного обучения.*

startDate = as.Date("2009-01-01") # *Начальная дата.*

endDate = as.Date("2013-12-31") # *Конечная дата.*

 # Скачиваем котировки индекса S&P500 с Yahoo Finance:

```
stock = getSymbols("^GSPC", src = "yahoo", from = startDate, to = endDate,
auto.assign=F)
colnames(stock) = c("Open", "High", "Low", "Close", "Volume", "Adjusted")
# Имена столбцов.

dayOfWeek = wday(stock, label=T, abbr=T) # День недели для каждой
даты.
dr = dailyReturn(Ad(stock), type='arithmetic')  # Относительные приращения цены за каждый день.
# Удалим первый элемент в обоих наборах данных (для него нет предыдущего дня):
dr = dr[-1];  dayOfWeek = dayOfWeek[-1]

# Максимальное, минимальное и среднее значения относительного приращения цен:
max(dr); min(dr); mean(dr)
s = 0.0035 # Максимальное относительное приращение цены, которое
отличает флет от тренда.
# Если относительное приращение цены больше s, то имеем тренд, иначе флет:
classes = ifelse(dr > s, "1", ifelse(dr < -s, "-1", "0"))

# Набор данных с двумя предикторами и столбцом значений классов:
dataset2 = data.frame(classes[-1], dayOfWeek[-1], classes[-length(classes)])
colnames(dataset2) = c("Класс", "ДеньНедели", "НаправлениеВчера")
n = nrow(dataset2) # Длина всего набора данных (кол-во дней)
```

ntrain = round(3 * n / 4, digits=0) # *Первые три четверти - обучающие.*
train2 = dataset2[1:ntrain-1,] # *Получили обучающую выборку из исходного набора данных.*
test2 = dataset2[ntrain:n,] # *Получили проверочную выборку из исходного набора данных.*

Теперь построим наивный байесовский классификатор и передадим ему данные для обучения:

 # Обучающий набор предикторов - столбцы 2 и 3, значения классов - столбец 1:
model2 = naiveBayes(train2[,2:3], train2[,1])
model2 # *Вывели параметры модели на экран.*

Теперь проверим, каким оказалось качество предсказаний. Сначала на обучающей выборке, а потом на проверочной:

pr2train = predict(model2, train2[,2:3]) # *Предсказанные состояния рынка для обучающей выборки.*
pr2test = predict(model2, test2[,2:3]) # *Предсказанные состояния рынка для проверочной выборки.*

 # Таблица правильности предсказаний для обучающей выборки:
t2train = table(pr2train, train2[,1], dnn=list('Предсказано', 'На самом деле'))
t2train # *Вывели таблицу на экран*

 # Таблица правильности предсказаний для проверочной выборки:

t2test = table(pr2test, test2[,1], dnn=list('Предсказано', 'На самом деле'))

t2test *# Вывели таблицу на экран*

ВАРИАНТЫ заданий

№ варианта	Вид заболевания	Количество групп	Количество измерений	Количество пациентов
1	Вирусный Гепатит	3 степени тяжести легкая, средняя, тяжелая	4 - Билирубин, гемоглобин, Альбумин, Гематокрит	33
2	Воспаление легких (пневмония)	5 видов: бактериальная, вирусная, микроплазменная, грибковая, смешанная	5 - Эритроциты, лейкоциты, лимфоциты, тромбоциты, нейтрофилы	28
3	Ларингит	4 вида: катаральный, аутрофический, гипертрофический, туберкулезный	3- Базофилы, плазмоциты, Эозинофилы	37
4	Бронхит	3 – острый, хронический, обструктивный бронхиты	5 – уровень сиаловых кислот, серомукоида, меноциты, лим-	14

			фоциты, Ретикулоциты	
5	Конъюктивит	5 - бактериальный, аллергический, ангулярный, вирусный, хламидийный	4 - эозинофилы, катионный белок, лимфициты, нейтрофилы	41
6	Кератит – воспаление роговицы	3 – экзогенный, эндогенный, герпетический	4- эритроциты, лейкоциты, гемоглобин, базофилы	27
7	Инвазия печени	2 – токсокара, шистосома	4 - Альбумин, ферритин, Билирубин, креатинин	39
8	Хламидиоз	2 - респираторный, офтальмохламидиоз	3- Глобулин, креатинин, ферритин	52
9	Грипп	3 – афебрильный, акатаральный, токсический,	3 - лейкоциты, моноциты, лимфоциты	43
10	Фарингит	2- атрофический, гипертрофический	4 - Трансферрин, плазмоциты, Билирубин, Гематокрит	13
11	Эмфизема	3- викарная, межуточная, идеопатическая	4 – уровень сиаловых кислот, серомукоида, меноциты, гемоглобин, Ретикулоциты	29
12	Плеврит	4- острый, инфекционный, хилезный, фибринозный	5- Альбумин, ферритин, плазмоциты, креа-	38

			тинин, моноциты	
13	Саркаидоз	3 – активный, регрессия, стабилизационный	3-ферритин, содержание белка и глюкозы	53
14	Пневмоторакс	5 – открытый, закрытый, напряженный, первичный, вторичный	3 - эритроциты, тромбоциты, глюкоза	49
15	Катаракта	3 – ядерная, кортикальная, веретенообразная	4 - содержание белка, креатинин, базофилы, нейтрофилы	52
16	Глаукома	4 – инфантильная, Ювенильная, врожденная, гипертензивная	5 - Ретикулоциты, глобулин, катионный белок, глюкоза	30

Для выполнения лабораторной работы необходимо:

1. Выбрать вариант задания из таблицы
2. Найти средние и стандартные отклонения дискриминантных переменных
3. Рассчитать матрицу рассеивания T (формула 4)
4. Рассчитать внутригрупповую матрицу W (формула 7)
5. Определить межгрупповые суммы квадратов отклонений и перекрестных (попарных) произведений (формулы 8 и 9)
6. Рассчитать общую ковариационную матрицу
7. Рассчитать внутригрупповую корреляционную матрицу
8. Рассчитать матрицу межгрупповых сумм перекрестных произведений
9. Найти коэффициенты дискриминантной функции
10. Найти внутригрупповые структурные коэффициенты
11. Найти коэффициенты канонической корреляции
12. Найти коэффициенты классифицирующих функций
13. Провести классификацию выбранного множества пациентов

14. С помощью средств языка R исследовать выбранное множество пациентов с помощью критерия Шапиро-Уилка, М-критерия Бокса
15. Классифицировать выбранное множество пациентов с помощью функций языка R: LDA(), STEPCLASS(), RFE() (использовать расстояние Махаланобиса)
16. Проверить качество предсказания классификации пациентов функцией языка R – NaiveBayes()

ЛИТЕРАТУРА

1. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: Классификация и снижение размерности. – М.: Финансы и статистика, 1989. – 607 с.
2. Афифи А., Эйзенс С. Статистический анализ. Подход с использованием ЭВМ. Пер. с англ. – М.: Мир, 1982. – 488 с.
3. Болч Б., Хуань К.Дж. Многомерные статистические методы для экономики: Пер.с англ. – М.: Статистика, 1979. – 317 с.
4. Каримов Р.Н. Обработка экспериментальной информации. Учеб. пособие. Ч. 3. Многомерный анализ. – Саратов: СГТУ, 2000. – 108 с.
5. Кендалл М.Дж., Стьюарт А. Многомерный статистический анализ временные ряды: Пер. с англ . – М.: Наука. Гл. ред. физ.-мат. лит., 1976. – 736 с.
6. Статистические методы для ЭВМ. Пер.с англ. – М.: Наука, Гл. ред. физ. мат. лит., 1986. – 464 с.
7. Факторный, дискриминантный и кластерный анализ: Пер. с англ., /Дж.-Он Ким, Ч. У. Мьюллер и др. – М.: Финансы и статистика, 1989. – 215 с.