Санкт-Петербургский государственный университет
Кафедра математической теории игр и статистических решений

Феофанов Василий Алексеевич

Выпускная квалификационная работа бакалавра

Дискриминантный анализ базы данных

Направление 010400
Прикладная математика, фундаментальная информатика и программирование

Научный руководитель,
доктор тех. наук,
профессор
Буре В.М.

Санкт-Петербург
2016

Содержание

Введение **4**

Постановка задачи **5**

Обзор литературы **6**

1 Обзор математических методов **7**

1.1 Дискриминантный анализ 7

1.1.1 Принцип дискриминации 7

1.1.2 Линейная дискриминация 7

1.1.3 Квадратичная дискриминация 10

1.2 Проверка выполнения условий 10

1.2.1 Критерий Шапиро-Уилка 10

1.2.2 Критерий Мардиа 11

1.2.3 Критерий Бартлетта 12

1.2.4 Box's M test 12

1.3 Отбор признаков 13

1.3.1 Необходимость отбора 13

1.3.2 Лямбда Уилкса и тест на добавочную информацию . 14

1.3.3 Пошаговый выбор: forward selection 15

1.4 Оценка величины ошибки 16

1.4.1 Ошибка обученной модели 16

1.4.2 Ошибка на обучении 17

1.4.3 Cross-validation leave-one-out 17

1.4.4 Bootstrap leave-one-out 18

1.4.5 Bootstrap 0.632 19

1.4.6 Bootstrap 0.632+ 20

2 Сведения из медицины **21**
2.1 Сочетанная травма груди 21
2.2 Травматический шок 21
2.3 Медицинские шкалы 22
2.3.1 Военно-полевая хирургия (ВПХ) 22
2.3.2 Шкала комы Глазго 23
2.3.3 AIS и ISS 23
2.4 Анализ сердечного ритма 23
2.5 Артериальное давление (АД) 25
2.6 Анализ газов крови 26
2.7 Анализ крови 26

3 Анализ Данных **28**
3.1 Описание задачи 28
3.2 Предварительная очистка данных 29
3.3 Пошаговый отбор признаков 30
3.3.1 Этап I 30
3.3.2 Этап II 33
3.3.3 Сравнение с результатами другого исследования . . . 34
3.4 Дополнительный анализ с целью улучшения результата . . . 35
3.5 Сравнение методов оценки величины ошибки 38

Выводы **41**

Заключение **43**

Список литературы **44**

Приложение **48**

Введение

Одним из самых актуальных разделов прикладной статистики на сегодняшний день является обучение классификационной модели с учителем. Методы классификации находят широкое применение в различных областях науки: в медицине [1], генетике [2], экономике [3], социологии [4].

Статистика всегда играла большую роль в медико-биологических системах. Благодаря этому, сам статистический анализ активно развивается. Достаточно вспомнить, что одной из ключевых фигур в статистике был и остается биолог Рональд Фишер. Но и сегодня в этой области возникают новые задачи, которые требуют нестандартного подхода к статистической обработке данных. Так, появляется класс задач анализа малого количества данных большой размерности, в биоинформатике получившие название Microarray Data [2].

Большая часть таких задач посвящена анализу различных смертельных болезней. Поэтому исследования формулируются в виде задачи классификации: предсказания летального или благоприятного исхода развития болезни для пациента. Основной проблемой в таких задачах при проведении статистического анализа является тот факт, что количество объектов в обучающей выборке в несколько раз меньше, чем признаков, описывающих каждого объекта. Поэтому возникает необходимость использования специальных методов, благодаря которым становится возможным использование классических методов классификации.

В данной работе будет анализироваться база данных по пациентам с травматической болезнью — сочетанной травмой груди. Будут рассматриваться актуальные методы, позволяющие расширить один из методов классификации, а именно дискриминантный анализ, на случай, когда число признаков превышает количество наблюдений.

Постановка задачи

Математически задача классификации формулируется следующим образом. Пусть имеется объект (наблюдение) x, характеризующийся набором из p признаков: $x = (x^{(1)}, x^{(2)}, \ldots, x^{(p)}) \in X$, где X — некоторое множество возможных значений. Пусть X разбито на не пересекающиеся между собой k классов (популяций) W_j:

$$Y = \{1, \ldots, k\},\ \bigcup_{j=1}^{k} W_j = X,\ W_j \bigcap W_i = \varnothing\ (i, j \in Y,\ i \neq j).$$

Требуется построить классификатор $m : X \to Y$, который будет определять принадлежность $x \in X$ к одному из классов. С практической точки зрения, имеется обучающая выборка из n наблюдений: $\{x_i\}_{i=1}^{n}$, для каждого из которых известно, к какой популяции он относится. Необходимо построить классификатор, предназначенный для последующего прогнозирования принадлежности к классам наблюдений, которые могут появиться впоследствии. Построение классификатора происходит за счет обучения на имеющихся данных [5].

В выпускной квалификационной работе рассматривается медицинская база данных пострадавших с сочетанной травмой груди, каждый из которых описывается большим количеством различных признаков. Для каждого пациента известно, исход полученной травмы был смертельный или благоприятный. Необходимо построить классификационное правило, позволяющее предсказать исход травмы для будущих пострадавших в ситуации, когда число наблюдений в обучающей выборке меньше числа признаков. С медицинской точки зрения, задача состоит в выявлении признаков, которые являются наиболее важными при оценке критического состояния пациента.

Обзор литературы

Основная часть теории из математической главы была взята преимущественно из книг:

1. Афифи А., Эйзен С. Статистический анализ: Подход с использованием ЭВМ [5].
2. Рао С. Р. — Линейные статистические методы и их применения [6].
3. Кобзарь А. И. Прикладная математическая статистика. Для инженеров и научных работников [9].
4. Rencher A. C. Methods of Multivariate Analysis [10].
5. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction [12].

Большая часть сведений из медицины была взята из книг:

1. Соколов В. А. Множественные и сочетанные травмы (практическое руководство для врачей травматологов) [22].
2. Мусалатов Х. А. Хирургия катастроф [26].
3. Военно-полевая хирургия [27].
4. Зудбинов Ю. И. Азбука ЭКГ [29].
5. Руководство по кардиологии [30].
6. Хеннеси А. А. М., Джапп А. Д. Анализ газов артериальной крови понятным языком [33].

Глава 1

Обзор математических методов

1.1 Дискриминантный анализ

1.1.1 Принцип дискриминации

Пусть наблюдения из рассматриваемых популяций W_j распределены нормально:

$$x \in W_j \Leftrightarrow x \sim N(\mu_j, \Sigma_j), j = \overline{1,k}$$

Классификация происходит из соображений минимизации величины $\sum_{j=1}^{k} p_j \Big(\sum_{\substack{i=1 \\ i \neq j}}^{k} P(i|j) \Big)$ — вероятности ошибочной класификации, где p_j — априорная вероятность принадлежности объекта к W_j, а $P(i|j)$ — вероятность ошибочно отнести элемент из j-го класса ко i-му. Из этого следует, что x будет отнесен к тому классу, который будет иметь наибольшую апостериорную вероятность [5, 6]:

$$P(j|x) = \frac{p_j f_j(x)}{\sum_{i=1}^{k} p_i f_i(x)}$$

1.1.2 Линейная дискриминация

Для начала, рассматривается частный случай, когда ковариационные матрицы популяций равны: $\Sigma_1 = \cdots = \Sigma_k = \Sigma$. Тогда плотности распределений имеют следующий вид:

$$f_j(x) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} e^{-\frac{1}{2}(x-\mu_j)^T \Sigma^{-1} (x-\mu_j)}$$

Все рассматриваемые апостериорные вероятности $P(j|x)$ имеют общий знаменатель $\sum_{i=1}^{k} p_i f_i(x)$. Поэтому вместо них можно рассмотреть выражения $p_j f_j(x)$, при этом принцип классификации не поменяется. Более того, будет удобным прологарифмировать эти выражения: $\ln(p_j f_j(x)) = \ln(p_j) +$

$\ln(f_j(x))$. Данная операция также не изменит принцип классификации, так как натуральный логарифм — монотонно возрастающая функция. Подставляя в эти выражения плотности $f_j(x)$ и сделав некоторые преобразования, получаем выражения, называемые дискриминантными функциями:

$$g_i(x) = \mu_i^T \Sigma^{-1} x - \frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \ln p_i$$

Таким образом, получаем следующее классификационное правило: объект x будет отнесен к той популяции, дискриминантная функция которой будет иметь наибольшее значение для x [6].

В случае двух классов достаточно рассмотреть разницу $g_2(x) - g_1(x)$ и ввести:

$$z(x) = (\mu_1 - \mu_2)^T \Sigma^{-1} x$$

$$c = (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 + \mu_2) + \ln \left(\frac{p_2}{p_1} \right)$$

Тогда классификационное правило будет выглядеть следующим образом:

$$W_1 : z(x) \geqslant c$$

$$W_2 : z(x) < c,$$

или:

$$W_1 : (\mu_1 - \mu_2)^T \Sigma^{-1} x \geqslant \frac{1}{2} (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 + \mu_2) + \ln \left(\frac{p_2}{p_1} \right)$$

$$W_2 : (\mu_1 - \mu_2)^T \Sigma^{-1} x < \frac{1}{2} (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 + \mu_2) + \ln \left(\frac{p_2}{p_1} \right)$$

Алгебраически, $z(x)$ представляет собой линейную комбинацию признаков объекта x, а геометрически, уравнение $z(x) = c$ является гиперплоскостью. Функцию $z(x)$ называют линейной дискриминантной функцией Рональда Фишера, который первым представил идеи дискриминации в 1936 году [7].

Для использования линейного дискриминантного анализа на практике необходимо проверить на заданном уровне значимости гипотезу о том, что обучающая выборка извлечена из нормальной генеральной совокупности, а также гипотезу о том, что ковариационные матрицы равны между собой. Однако, фактически, линейный дискриминантный анализ работает

и при отклонении от рассматриваемых гипотез. Лахенбрук в своей работе [8] говорит о том, что линейный дискриминантный анализ показывает хорошие результаты, имея дело с дискретными признаками. Также ничто не мешает использовать рассматриваемое классификационное правило и в других случаях, когда отклонена гипотеза о равенстве ковариационных матриц или гипотеза о нормальности распределения генеральной совокупности. Однако стоит иметь ввиду, что это может привести к большой ошибке при классификации [8].

Таким образом, для применения на практике описанного выше классификационного правила необходимо [5]:

1. оценить математические ожидания всех популяций средними по выборкам
   $\hat{w}_1 = \{x_{1i}\}_{i=1}^{n_1}, \ldots, \hat{w}_k = \{x_{ki}\}_{i=1}^{n_k}$ $(n_1 + \cdots + n_k = n)$— всех элементов принадлежащих первому, ..., k-му классу соответственно:
   $$\mu_1 \to \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}\,, \ldots,\ \mu_k \to \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ki},$$

2. оценить ковариационную матрицу, как:
   $$S = \frac{1}{n-k}((n_1-1)S_1 + (n_2-1)S_2 + \cdots + (n_k-1)S_k),$$
   где $S_1, S_2, \ldots, S_k$ — выборочные ковариационные матрицы $\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_k$.
   В случае двух популяций выражение будет выглядеть следующим образом:
   $$S = \frac{n_1-1}{n-2}S_1 + \frac{n_2-1}{n-2}S_2.$$

3. оценить априорные вероятности $p_1, \ldots, p_k$ выражениями $\frac{n_1}{n}, \ldots, \frac{n_k}{n}$ соответственно.

1.1.3 Квадратичная дискриминация

В общем случае, когда ковариационные матрицы не равны, плотности распределений представляются следующим образом:

$$f_i(x) = \frac{1}{\sqrt{(2\pi)^p |\Sigma_i|}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)}.$$

Тогда дискриминантные функции принимают вид [6]:

$$g_i(x) = -\frac{1}{2}\ln(|\Sigma_i|) - \frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) + \ln p_i$$

Аналогичным образом, как и в секции 1.1.2, строится классификационное правило: объект x принадлежит той совокупности, дискриминантная функция которой доставляет наибольшее значение в точке x среди всех рассматриваемых функций $g_i(x)$.

Как и в случае с линейной дискриминацией, здесь может быть введена одна дискриминантная функция для случая двух популяций, и правило примет следующий вид:

$$W_1 : z(x) \geqslant c$$

$$W_2 : z(x) < c,$$

$$z(x) = -\frac{1}{2}x^T(\Sigma_1^{-1} - \Sigma_2^{-1})x + (\mu_1^T\Sigma_1^{-1} - \mu_2^T\Sigma_2^{-1})x$$

$$c = \frac{1}{2}(\mu_1^T\Sigma_1^{-1}\mu_1 - \mu_2^T\Sigma_2^{-1}\mu_2) + \frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + \ln\left(\frac{p_2}{p_1}\right)$$

В отличие от линейного случая, здесь дискриминантная функция уже не будет представлять собой линейную комбинацию признаков, а уравнение $z(x) = c$ будет являться поверхностью второго порядка.

1.2 Проверка выполнения условий

1.2.1 Критерий Шапиро-Уилка

Пусть рассматривается одномерная выборка $\{x_i\}_{i=1}^n$. Критерий Шапиро-Уилка позволяет проверить, извлечены ли наблюдения из нормальной генеральной совокупности. В качестве нулевой гипотезы берется

нормальность распределения генеральной совокупности.Рассматривается статистика W:

$$W = \frac{\gamma^2}{s^2}; \;\; \gamma = \sum_{i=1}^{n} a_{n-i+1}(x_{n-i+1} - x_i)$$

$$s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2; \;\; \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Значение коэффициентов a_{n-i+1} берутся из таблицы, которую можно найти в [9]. Задаются значением α и вычисляется критическое значение $W(\alpha)$. Если $W < W(\alpha)$, тогда на уровне α гипотеза о нормальном распределении отклоняется [9].

1.2.2 Критерий Мардиа

Пусть теперь наблюдения представляют собой p-мерные вектора $x = (x^{(1)}, x^{(2)}, \ldots, x^{(p)})$ и требуется проверить многомерную нормальность выборки. Вводятся многомерные аналоги коэффициентов асимметрии и эксцесса:

$$b_{1,p} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left((x_i - \overline{x})^T S^{-1}(x_j - \overline{x})\right)^3,$$

$$b_{2,p} = \frac{1}{n}\sum_{i=1}^{n}\left((x_i - \overline{x})^T S^{-1}(x_i - \overline{x})\right)^2,$$

$$S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})^T.$$

Рассматриваются статистики:

$$B_1 = \frac{(p+1)(n+1)(n+3)}{6\left((n+1)(p+1) - 6\right)} b_{1,p}, \;\; B_2 = \frac{b_{2,p} - p(p+2)}{\sqrt{8p(p+2)/n}}$$

Статистика B_1 в пределе распределена по закону χ^2 со степенью свободы $\frac{1}{6}p(p+1)(p+2)$, тогда как B_2 — по стандартному нормальному закону. Поэтому на заданном уровне альфа гипотеза отклоняется, если $B_1 > \chi^2_{\alpha,\frac{1}{6}p(p+1)(p+2)}$ и $B_2 \in (-\infty; z_{\alpha/2}) \cup (z_{1-\alpha/2}; +\infty)$, где χ^2, z — соответственно квантили χ^2 и стандартного нормального распределений [10].

1.2.3 Критерий Бартлетта

Используется для проверки равенства дисперсий k выборок: $H_0 : \Sigma_1 = \cdots = \Sigma_k$. Пусть размеры выборок равны соответственно $n_1, n_2, \ldots, n_k$, при этом, $n = \sum_{i=1}^{k} n_i$, а $n_i > 3$. Рассматривается статистика B:

$$B = \frac{M}{C}$$

$$M = (n-k)\ln\frac{1}{n-k}\sum_{i=1}^{k}(n_i-1)s_i^2 - \sum_{i=1}^{k}(n_i-1)\ln s_i^2$$

$$C = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i-1} - \frac{1}{n-k}\right),$$

где S_i^2 — выборочная дисперсия выборки i. Статистика B подчинена закону χ^2 с $df_1 = k-1$ степенями свободы. Если $B > \chi^2_{\alpha,df_1}$, тогда гипотеза равенства дисперсий отклоняется на уровне α.

В случае, если гипотеза о нормальности не выполняется для какой-то из выборок, рассматривается модификация статистики B:

$$B^* = \frac{df_2 M}{df_1\left(\frac{df_2^2}{df_2(2-C)+C} - M\right)},$$

где $df_2 = \frac{k+1}{(C-1)^2}$. Статистика B^* распределена по закону Фишера со степенями свободы df_1 и df_2. Нулевая гипотеза отклоняется, если $B^* > F_{\alpha,df_1,df_2}$ [9].

1.2.4 Box's M test

Пусть нулевая гипотеза представляет собой предположение о равенстве ковариационных матриц: $H_0 : \Sigma_1 = \cdots = \Sigma_k$, а также выполняется предположение о нормальной распределенности наблюдений. Обозначим, как и в секции 1.2, за $S_i\ \left(i = \overline{1,k}\right)$ — выборочные ковариационные матрицы, а S — объединенную выборочную ковариационную матрицу. Рассматривается статистика U:

$$U = (C-1)\ln M$$

$$\ln M = \sum_{i=1}^{k}(n_i-1)\ln|S_i| - (n-k)\ln|S|$$

$$C = \frac{2p^2 + 3p - 1}{6(p+1)(k-1)} \left(\sum_{i=1}^{k} \frac{1}{n_i - 1} - \frac{1}{n-k} \right).$$

Статистика U в пределе распределена, как χ^2_{df}, где $df = \frac{1}{2}(k-1)p(p+1)$. Гипотеза отклоняется на уровне α, если $U > \chi^2_{\alpha,df}$ [10].

1.3 Отбор признаков

1.3.1 Необходимость отбора

В реальных задачах анализируемые данные довольно часто представлены в «сыром» виде. Например, наличествуют пропущенные значения и выбросы, рассматриваются большое количество признаков. На последнее особенно важно обратить внимание, если число признаков равно или даже превышает количество наблюдений. В таком случае, при оценке ковариационной матрицы оцениваемых параметров будет больше, чем наблюдений, по которым оцениваются эти параметры; а это означает, что ковариационная матрица будет плохо обусловленной. Более того, даже когда количество признаков меньше количества наблюдений, но, при этом, не значимо меньше, ковариационная матрица также может оказаться плохо обусловленной, что чревато неадекватному классификационному правилу, так как в формулах дискриминантных функций вычисляется обратная ковариационная матрица. Плохая обусловленность может быть вызвана не только количеством признаком, но и тем, что некоторые признаки близки к линейной зависимости [11].

Поэтому, прежде чем провести сам анализ, необходимо удалить некоторое число признаков, которые являются избыточными для исследования. Первоначально исследователи отбирают те характеристики, которые представляют для них интерес и которые они считают достаточно объективными для рассмотрения. Если оставшиеся характеристики равнозначны, нужно провести исследование каждого признака на значимость его вклада в классификационное правило, и отбросить наименее значимые. В действительности, такой отбор приводит к улучшению точности классификации [11]. Более того, при уменьшении размерности пространства признаков

увеличивается устойчивость линейного и квадратичного классификатора к отклонениям от предположения нормальности [10].

Существует множество различных методов отбора переменных. Дальше пойдет речь об одном из них — пошаговый дискриминантный анализ, еще называемый пошаговым MANOVA. В основе этого метода лежит тест на добавочную информацию, основанный на частной лямбде Уилкса. Существует три вида данного метода: forward, backward и stepwise selection [10]. В работе будет рассматриваться алгоритм forward selection.

1.3.2 Лямбда Уилкса и тест на добавочную информацию

Пусть на некотором этапе исследования данные описываются набором переменных, $t = (t^{(1)}, t^{(2)}, ..., t^{(r)})$. Сгруппируем наблюдения по классам:

$$(t_{ij}); i = 1, \ldots, n_j; j = 1, \ldots, k; \sum_{j=1}^{k} n_j = n,$$

где t_{ij} — i-е наблюдение из класса j. Лямбда Уилкса используется для проверки гипотезы о равенстве средних $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$. Выглядит она следующим образом:

$$\Lambda(t) = \frac{|W_{tt}|}{|W_{tt} + B_{tt}|},$$

где B и W меж- и внутри- групповые матрицы:

$$B = \sum_{j=1}^{k} n_j (\bar{t}_j - \bar{t})^T (\bar{t}_j - \bar{t})$$

$$W = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (t_{ij} - \bar{t}_j)^T (t_{ij} - \bar{t}_j)$$

$$\bar{t}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} t_{ij}, \bar{t} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} t_{ij}$$

Если гипотеза верна, то статистика распределяется по закону Уилкса: $\Lambda_{p,k-1,n-k}$, и значения, меньшие или равные квантиля заданного уровня значимости, позволяют отклонить гипотезу [10].

Дальше, на основе этого теста строится тест на добавочную информацию. Пусть имеется набор переменных,$t = (t^{(1)}, t^{(2)}, ..., t^{(r)})$, по которым описываются данные, и набор переменных $x = (x^{(1)}, x^{(2)}, ..., x^{(q)})$, которые затем были добавлены в модель. Суть теста заключается в определении значимости вклада новых переменных для проверки гипотезы $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$. Если вклад в отклонении гипотезы будет не значим, переменные можно будет удалить, так как нас интересует задача отделения друг от друга классов. Рассматриваются k групп, состоящие из наблюдений, принадлежащих соответствующим классам:

$$(t_{ij}, x_{ij}), i = 1, \ldots, n_j; j = 1, \ldots, k; \sum_{j=1}^{k} n_j = n$$

Внутри- и меж- групповые матрицы можно представить, как:

$$W = \begin{pmatrix} W_{tt} & W_{tx} \\ W_{xt} & W_{xx} \end{pmatrix}, B = \begin{pmatrix} B_{tt} & B_{tx} \\ B_{xt} & B_{xx} \end{pmatrix}$$

Нижние индексы обозначают переменные, на основе которых строится данный блок. Например, W_{tt} - внутригрупповая матрица, построенная только по переменным t, то есть, внутригрупповая матрица до добавления переменных x. Дальше вычисляются статистики Уилкса до и после добавления набора признаков x:

$$\Lambda(t, x) = \frac{|W|}{|W + B|}, \Lambda(t) = \frac{|W_{tt}|}{|W_{tt} + B_{tt}|}$$

И рассматривается статистика:

$$\Lambda(x|t) = \frac{\Lambda(t, x)}{\Lambda(t)},$$

которая распределяется по закону $\Lambda_{q,k-1,n-k-r}$. Значения, меньшие или равные квантиля $\Lambda_{q,k-1,n-k-r,\alpha}$, позволяют отклонить гипотезу H_0. Таким образом, чем меньше $\Lambda(t, x)$ по сравнению с $\Lambda(t)$, тем меньше $\Lambda(x|t)$, а значит, тем больше x несет в себе дополнительной информации [10].

1.3.3 Пошаговый выбор: forward selection

Пусть снова произвольное наблюдение x описывается p признаками: $x = (x^{(1)}, x^{(2)}, \ldots, x^{(p)})$. Алгоритм forward selection выглядит следующим

образом:

1. Вначале из модели удаляются все рассматриваемые признаки, число которых p.

2. Задается значение Λ-включения.

3. Для каждого $x^{(j)}$ высчитывается $\Lambda(x^{(j)})$ $\left(j = \overline{1,p}\right)$ и затем включается в модель та переменная, значение соответствующей статистики которой наименьшее среди рассматриваемых. Включенный в модель признак обозначается за $t^{(1)}$.

4. Среди остальных $p-1$ переменных ищется признак с наименьшей частной лямбдой Вилкса:

$$\Lambda(x_j|t_1) = \frac{\Lambda(t^{(1)}, x^{(j)})}{t^{(1)}}, j = \overline{1, p-1}$$

   Вместе с тем, статистика должна удовлетворять условию:
   $\Lambda \leqslant \Lambda$-включения. Если условие не выполняется ни для одной статистики из рассматриваемых, то пошаговый процесс останавливается.

5. Процесс продолжается аналогичным образом до тех пор, пока ни у одной из переменных соответствующая статистика не будет удовлетворять условию включения или все переменные не войдут в модель.

Как итог, получается набор переменных $t_1, \ldots, t_r$, который может дать лучший результат, чем изначальный набор признаков [10].

1.4 Оценка величины ошибки

1.4.1 Ошибка обученной модели

С практической точки зрения, очень важным этапом исследования является оценка точности классификатора, который был построен на обучающей выборке. Для этого вводится понятие величины ошибки обученной модели, которое определяется, как математическое ожидание ошибочной классификации произвольного наблюдения x:

$$Err = E(I(y \neq m\,(x))),$$

где y — метка принадлежности к одному из классов, $m(x)$ — обученная модель, а I — функция-индикатор, которая равна 1 в случае, если аргумент, являющийся логической переменной, равен истине, и равна 0 в противном случае. На основе оценки данной величины можно судить о качестве классификационной модели, а также имеется возможность сравнить несколько таких моделей. В случае, когда исследуемая выборка содержит большое число наблюдений, для оценки величины ошибки выборку принято делить на три части: обучающую, контрольную и тестовую. На основе первой подвыборки происходит обучение рассматриваемых методов классификации, на второй выбирается модель, имеющая наименьшую ошибку, тогда как на третьей подвыборке происходит окончательная оценка величины ошибки выбранного классификатора [12]. Однако, при отсутствии большого количества наблюдений такая процедура является нецелесообразной, поэтому на малых выборках используются специальные методы оценки величины ошибки, которые будут рассмотрены ниже.

1.4.2 Ошибка на обучении

Естественна мысль величину ошибки оценить значением ошибки на обучении:

$$\overline{err} = \frac{1}{n} \sum_{i=1}^{n} I\left(y_i \neq m(x_i)\right)$$

То есть, классификатор, обученный на выборке, применили к тем же самым обучающим данным и посчитали среднее-арифметическое сделанных ошибок. Но такой подход, как правило, дает оптимистически заниженную оценку в связи с таким явлением, как переобучение — модель «адаптируется» к обучающей выборке и на других данных предсказывает значимо хуже [13]. Таким образом, ошибка на обучении является смещенной оценкой и, как правило, дает недостоверные результаты [12].

1.4.3 Cross-validation leave-one-out

Одним из самых известных подходов к оценке величины ошибки — кросс-валидация, известный в русской литературе также, как метод скользящего контроля [13]. Рассматриваемую выборку делят на k примерно рав-

ных частей. Последовательно, k-раз выбирается одна контрольная группа, на которой будет тестироваться алгоритм, а остальная часть выборки берется в качестве обучающих данных, по которым будет строится классификационное правило. В итоге вычисляется величина:

$$\widehat{Err}^{(CV)} = \frac{1}{n}\sum_{i=1}^{n} I(y^{(i)} \neq m^{-\ell(i)}(x^{(i)})),$$

где $\ell : \{1, ..., n\} \mapsto \{1, .., k\}$ — функция, аргументом которой является номер наблюдения, а значением — номер группы, к которой принадлежит это наблюдение, и тогда $m^{(1)}, m^{(2)}, \ldots, m^{(k)}$ — классификаторы, построенные по всем группам, кроме первой,. . . , k-й группы соответственно, на каждой из которых соответствующий классификатор будет тестироваться [12]. Таким образом, $CV(m)$ и есть оценка величины ошибки классификационной модели m. Самыми распространенными подходами кросс-валидации являются ситуации, когда $k = 5, k = 10, k = n$ [12]. Последний подход получил название метода leave-one-out, который был описан еще в 1968 году [14]. В этом случае $\ell(i) = i$, что означает, что каждый элемент будет поочередно использоваться для контроля, тогда как на остальных модель будет обучаться. С одной стороны, leave-one-out дает несмещенную оценку, но с другой — имеет, как правило, высокую дисперсию оценки [12]. Однако, на выборках малых размеров данный подход довольно хорошо себя зарекомендовал. Некоторый сравнительный анализ с другими методами оценки величины ошибки можно найти в работах [15, 16].

1.4.4 Bootstrap leave-one-out

Идея бутстрапинга, развитая Брэдли Эфроном [17], заключается в следующем: из обучающей выборки $(x^{(i)}, y_i)_{i=1}^{n}$ случайным образом n раз выбирается одно из наблюдений, причем в дальнейшем оно может быть выбрано снова. Таким образом, получается новая выборка того же объема, содержащая, однако, в общем случае не все объекты из обучающих данных и какие-то объекты содержащая по несколько раз. Такая процедура проделывается B раз. Теперь, полученные B выборок можно рассмотреть, как B обучающих наборов данных, а изначальные данные, как контрольную

выборку. Итак, величину ошибки можно оценить выражением:

$$\widehat{Err}^{(Boot)} = \frac{1}{n}\frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n} I(y^{(i)} \neq m^{(b)}(x^{(i)})),$$

где $m^{(b)}$— классификатор, построенный по выборке b. Но, нетрудно видеть, что по факту и обучающие, и контрольные данные используют одни и те же наблюдения, что влечет за собой, как правило, эффект переобучения [12]. Использовав здесь идеи кросс-валидации, эту проблему можно решить. Пусть для каждого наблюдения ошибка будет подсчитываться только на тех классификаторах, которые построены по бутстрап-выборкам, не содержащих данного наблюдения. Такой подход напоминает метод leave-one-out, потому он и получил название bootstrap leave-one-out. Таким образом, оценка величины ошибочной классификации будет выглядеть следующим образом:

$$\widehat{Err}^{(LOOB)} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} I(y^{(i)} \neq m^{(b)}(x^{(i)})),$$

где C^{-i} — набор индексов, идентифицирующие те бустрап-выборки, которые не содержат объект i, а $|C^{-i}|$ — количество таких выборок. Из формулы следует, что B должно быть достаточно большим для того, чтобы $|C^{-i}|$ не равнялись нулю [12].

1.4.5 Bootstrap 0.632

Данный метод является развитием бутстрап leave-one-out и основывается на следующем факте. Вероятность того, что конкретное наблюдение из исследуемой выборки при однократном выборе будет взято, равна $\frac{1}{n}$. Отсюда с вероятностью $1 - \frac{1}{n}$ наблюдение выбрано не будет. Значит событие, что наблюдение не принадлежит бутстрап-выборки, имеет вероятность $(1 - \frac{1}{n})^n \approx e^{-1} = 0.368$. Из этого следует, что, в среднем, в бутстрап выборке будет $0.632 \cdot n$ различных наблюдений. Оценка величины ошибочной классификации бутстрап 0.632 выглядит следующим образом [12, 18]:

$$\widehat{Err}^{(0.632)} = 0.368 \cdot \overline{err} + 0.632 \cdot \widehat{Err}^{(LOOB)}.$$

1.4.6 Bootstrap 0.632+

Рассматриваемый подход также развивает бутстрап leave-one-out, учитывая при этом такую величину, как относительная частота переобучения:

$$\widehat{R} = \frac{\widehat{Err}^{(LOOB)} - \overline{err}}{\widehat{\gamma} - \overline{err}},$$

где $\widehat{\gamma} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} I(y^{(i)} \neq m(x^{(j)}))$ — оценка величины ошибки в условиях отсутствия информации, то есть оценка получена посредством классификатора, построенного на всех возможных комбинациях $y^{(i)}$ и наблюдений $x^{(j)}$. Как видно, относительная частота обучения лежит в интервале [0,1], где левое значение говорит о том, что эффекта переобучения нет. Таким образом, оценка бутстрап 0.632+ определяется, как:

$$\widehat{Err}^{(0.632+)} = (1-\alpha)\cdot\overline{err} + \alpha\cdot\widehat{Err}^{(LOOB)},$$

где $\alpha = \frac{0.632}{1-0.368\widehat{R}}$, которая равна 0.632, при $\widehat{R} = 0$, и равна 1, при $\widehat{R} = 1$. Таким образом, bootstrap 0,632+ варьируется от bootstrap 0.632 до bootstrap leave-one-out, являясь компромиссом между ними, в зависимости от относительной частоты переобучения [12, 19]. Как правило, все три рассматриваемых в настоящей работе бутстрапа имеют низкую дисперсию, как оценки величины ошибки. Но, вместе с этим, у всех методов наличествует смещение, в особенности у bootstrap leave-one-out, который во многих случаях дает пессимистически завышенную оценку величины ошибки [12]. Это подтверждается экспериментами, проделанными в работах [15, 16], рассматривающие выборки малых размеров. Несмотря на это, бутстрап методы не теряют свою актуальность, и широко используются для оценки ошибочной классификации для маленьких выборок [18, 19, 20].

Глава 2

Сведения из медицины

2.1 Сочетанная травма груди

Повреждение головы, нижних конечностей и груди являются самыми распространенными травмами в дорожно-транспортных происшествиях [21]. Как правило, водители легковых автомобилей получают повреждение грудной клетки при ударе о рулевое колесо. Нередки ситуации получения множественной или сочетанной травмы. Сочетанной травмой называют одновременное повреждение травмирующим агентом более двух анатомических областей тела, которых выделяют всего семь: голова, шея, грудь, живот, таз, позвоночник, конечности [22]. При этом, считается, что одно из повреждений является опасным для жизни. Травма груди практически не обходится без перелома ребер или грудины и часто сопровождается повреждением спины [23], живота [24], легких и органов средостения, наличествуют гемоторакс и пневмоторакс [25]. Пневмоторакс возникает при повреждении легкого или бронха — воздух или газы поступают в плевральную полость, что может привести к нехватке кислорода и снижению артериального давления, что в свою очередь может вызвать остановку сердца. Гемоторакс же представляет собой ситуацию, когда в плевральной полости скапливается кровь из поврежденных сосудов, легких и средостения, что влечет нарушения функции внешнего дыхания, а также острую кровопотерю при истечении более 1 литра крови [26].

2.2 Травматический шок

При получении тяжелых травм первые 6–12 часов особенно важны с точки зрения оказания медицинской помощи, так как этот период характеризуется острым нарушением жизненно важных функций. Более того,

наиболее эффективным является оказание срочной медицинской помощи в течение первого часа. Согласно [27], повреждения жизненно важных функций организма в 63% случаев проявляется в виде травматического шока, в 18% — травматической комой, в 13% — острой дыхательной недостаточностью, а в 6% — острой сердечной недостаточностью. Травматический шок является тяжелым и жизнеугрожающим состоянием, проявляющимся в виде острого нарушения кровообращения, которое происходит вследствие острой кровопотери, повреждения жизненно важных органов, расстройства газообмена, интоксикации организма. Главным симптомом в определении травматического шока является понижение артериального давления [27, 26]. Шок может стать необратимым в случае, если полноценная помощь задерживается более, чем на 2 часа. Важными симптомами развития необратимого шока являются уменьшение объема циркулирующей плазмы, повышение гематокритного числа, а также наличием в плазме крови свободного гемоглобина [26].

2.3 Медицинские шкалы

2.3.1 Военно-полевая хирургия (ВПХ)

Объективная оценка тяжести состояния является одной из главных задач при диагностики пострадавшего. Для решения этой задачи используются различные шкалы, помогающие определить лечебную тактику, тем самым улучшая качество лечения [27]. В начале 1990-х годов на кафедре военно-полевой хирургии Военно-медицинской академии им. С.М.Кирова сформировалась собственная методология оценки тяжести травм. Одним из преимуществ этих методов является возможность объективно оценить тяжесть сочетанных, множественных и комбинированных травм. Шкала ВПХ-П(МТ) (П-повреждение, МТ-механическая травма) применяется для количественной оценки тяжести повреждений, позволяя предсказать окончательный исход травмы. Шкалирование осуществляется путем присвоения конкретному повреждению соответствующего балла тяжести. При оценке тяжести сочетанных, множественных и комбинированных травм определяется степень тяжести каждого повреждения с последующим сум-

мированием баллов. Шкала ВПХ-СП (СП - состоянии при поступлении) используется для объективной оценки тяжести состояния. ВПХ-СП состоит из 12 наиболее важных признаков, которые могут быть определены медиками при поступлении пострадавшего [27].

2.3.2 Шкала комы Глазго

В 1974 году Б. Дженнетом и Дж. Тисдейлом была предложена шкала для оценки сознания пациентов старше 4 лет, которая состоит из трех компонент: проверка двигательной (motor), вербальной (verbal) реакции и реакции открывания глаз (eye) на внешнее воздействие. Начисляются баллы: от 1 до 6, от 1 до 5, от 1 до 4 соответственно. Затем баллы суммируются и определяется состояние больного. 15 баллов означает, что пациент находится в ясном сознании. Ниже 8 баллов говорит о том, что больной находится в коме, 3 балла - в состоянии запредельной комы [28].

2.3.3 AIS и ISS

При применении шкалы Abbreviated Index Severity(AIS) все повреждения ранжируются в баллах от 1 до 6. Легкие повреждения имеют 1 балл, тогда как балл 6 соответствует безусловно смертельным травмам. Ранжирование проводится по таким областям человеческого тела, как голова, шея, лицо, грудь, живот, конечности, наружные покровы. AIS применима скорее для оценки тяжести отдельных повреждений, поэтому для сочетанных травм используется шкала Injury Severity Scale — сумма квадратов баллов трех наиболее поврежденных областей тела по шкале AIS [22].

2.4 Анализ сердечного ритма

Одним из важнейших этапов диагностики пострадавшего является анализ работы сердца. Для этого используются запись электрокардиограммы и измерение вариабельности сердечного ритма с помощью спироартериокардиоритмографа.

Электрокардиограмма (ЭКГ) — метод графической регистрации прохождения электрических импульсов, исходящих из синусового узла сердца,

по проводящей системе сердца. Синусовый узел, который посылает 60–90 импульсов в минуту, располагается в правом предсердии, в месте слияния полых вен. Прохождение импульсов по проводящей системе сердца изображается в виде кривой, которая характеризуется своими пиками подъемов и спадов. Эти пики называются зубцами электрокардиограммы, которых обычно выделяют 5 и обозначают латинскими буквами P, Q, R, S и T [29].

Зубец P отражает в себе совокупное отображение прохождения синусового импульса по проводящей системе предсердий и последовательное возбуждение правого и левого предсердий. Амплитуда зубцов P, как правило, не превышает 1,5–2,5 миллиметра, а продолжительность — 0,1 секунды. Интервал PQ измеряется от начала зубца P до начала желудочкового комплекса QRS. Параллельно с возбуждением предсердий импульс идет к предсердножелудочковому соединению, где возникает физиологическая задержка импульса. PQ интервал отражает время распространения импульса в предсердножелудочковом соединении. Длительность интервала варьируется от 0,12 до 0,20 секунд. Желудочковый комплекс QRST представляет собой процесс распространения, который характеризуется комплексом QRS, и угасания, характеризующийся сегментом RS–T и зубцом T, возбуждения по миокарду желудочков. Интервал QT, который принято называть электрической систолой желудочков, измеряется от начала комплекса QRS до конца зубца T. В течение этого интервала происходит возбуждения всех отделов желудочков сердца [29, 30].

Интервал R–R измеряется от зубца R одного сердечного цикла до зубца R следующего зарегистрированного цикла. На его основе определяется частота сердечных сокращений (ЧСС). При правильном ритме ЧСС определяется, как $\frac{k}{R-R}$, где k — показатель, зависящий от скорости ленты ЭКГ, а R–R — длительность интервала в секундах. В состоянии покоя ЧСС составляет от 60–90 ударов в минуту. Превышение нормы называют тахикардией, а снижение — брадикардией [30].

Вариабельность сердечного ритма — явление, заключающееся в постоянном изменении интервала времени от начала цикла одного сердечного сокращения до начала другого. Показатели ВСР отражают состояние вегетативного баланса, являющегося жизненно важным показателем контроля

физиологическими функциями организма. При анализе ВСР оценивается функциональное состояние организма и его динамика. Снижение показателей ВСР говорит о нарушении вегетативного контроля сердечной деятельности, а также может свидетельствовать о неблагоприятном прогнозе — смерти пациента с высокой долей вероятности. Как правило, высокие показатели ВСР имеют здоровые лица молодого возраста и спортсмены, низкие — люди с различными органическими заболеваниями сердца [31].

2.5 Артериальное давление (АД)

Артериальное давление — это давление крови в артериях, варьирующееся при каждом ударе сердца между систолическим и диастолическим давлением. Систолическое артериальное давление (САД) — давление крови в момент систолы сердца, то есть когда желудочки сердца сжимаются и выталкивают кровь в кровеносные сосуды. Величина систолического давления зависит преимущественно от состояния миокарда, силы и скорости сокращения сердца [32]. Является одной из важнейших характеристик, способствующих определению тяжести шока [26]. Диастолическое артериальное давление (ДАД) — это давление, поддерживающееся в сосудах в момент диастолы, расслабления сердца, и формирующееся за счет сокращения периферических артерий, по которым кровь поступает к органам и тканям. На величину диастолического давления влияет эластичность сосудов, общий объем крови и частота сердечных сокращений [32].

Снижение артериального давления уменьшает возможности оптимизации тканевого кровотока. Повышение уровня АД создает более благоприятные условия для обеспечения тканевого кровоснабжения, но в то же время сопряжен с резким возрастанием нагрузки на сердце и его работу, что может стать причиной развития заболеваний сердечно-сосудистой системы. В связи с этим, уровень АД должен варьироваться в узких пределах относительно оптимального уровня — САД порядка 120 мм рт. ст, ДАД порядка 80 мм рт. ст. [30]. Артериальное давление отражает состояние органного и тканевого кровотока только в таких частях тела, как мозг и сердце. Например, при систолическом артериальном давлении ниже 60

мм рт. ст. нарушается регуляция мозговых сосудов, вследствие чего резко уменьшается перфузия мозга [26].

2.6 Анализ газов крови

Анализ газов крови предназначен для измерения концентрации свободных ионов водорода и парциальных давлений кислорода и углекислого газа в артериальной крови. На основе этих величин оценивается кислотно-основное состояние в крови и эффективность газообмена в легких [33]. Согласно [26], в 56,3–61,3% случаев тяжелые травмы сопровождаются расстройством газообмена, поэтому при критическом состоянии пациента необходимо исследование газового состава.

Парциальное давление описывает вклад отдельного газа воздуха в общее давление, посредством чего можно определить количество растворенного газа в крови. Концентрация водорода в крови измеряется в специальной шкале pH. В норме этот показатель варьируется от 7,35 до 7,45 наномоль на литр. Его повышение приводит к алкалемии, а понижение — к ацедимии. Значительное отклонение концентрации водорода (больше 7,8 или меньше 6,8) неминуемо влечет за собой смерть человека.

Анализ газов крови является одним из ключевых факторов в постановки таких диагнозов, как дыхательная недостаточность, гипервентиляция (падение парциального давления углекислого газа (pCO_2) и рост pH), нарушение оксигенации (снижение парциального давления кислорода (pO_2)), гипоксемия — недостатка кислорода в крови (понижение содержания гемоглобина или результат нарушения оксигенации), ишемия — недостаточный приток крови. Также, благодаря показателю доли кислорода во вдыхаемом воздухе (FiO_2), осуществляется контроль над лечением пациентов, подключенных к аппарату искусственной вентиляции легких [33].

2.7 Анализ крови

Биохимический анализ крови — лабораторное исследование, с помощью которого можно получить информацию об обмене липидов, белков

и углеводов, выяснить потребность в микроэлементах. Анализ позволяет оценить работу печени, почек, поджелудочной железы, желчного пузыря и других внутренних органов. На основе креатина, калия, мочевины и мочевой кислоты определяются почечные заболевания, на основе общего билирубина, аспартатаминотрансфераза (АСТ), аланинаминотрансфераза (АЛТ) — патологии печени. Повышение показателя общего количества белков в крови происходит при заболеваниях крови и инфекционно-воспалительных процессах. Важным показателем является уровень натрия, отвечающий за работу нервной и мышечной ткани, пищеварительных ферментов, водный обмен и кровяное давление [34].

Общий анализ крови — лабораторная диагностика, включающая в себя подсчет всех видов клеток крови, определение их параметров, лейкоцитарную формулу, измерение уровня гемоглобина, определение соотношения клеточной массы к плазме (гематокрит). Лейкоциты отвечают за обезвреживание от вирусов и бактерий, создание клеточного иммунитета. Поэтому увеличение числа лейкоцитов говорит о наличии воспалительного процесса в организме. Тромбоциты участвуют в процессе свёртывания крови. Снижение их числа — признак плохой свертываемости крови. Снижение концентрации гемоглобина и числа эритроцитов говорит о синдроме малокровии (анемии) [34]. Также низкая концентрация гемоглобина в крови может говорить о гипоксемии, которая может привести к кислородному голоданию (гипоксии) [33]. По уровню гемоглобина судят о травматичности операции, определяющейся объемом операционной кровопотери. Есть вероятность, что у обескровленного пациента даже самое минимальное вмешательство закончиться фатально. Поэтому учитывается концентрация гемоглобина в крови, а также масса тела пострадавшего для определения объема операционной кровопотери, которая может быть им перенесена [22].

Глава 3

Анализ Данных

3.1 Описание задачи

В этой главе будет проводится анализ базы данных, предоставленной сотрудниками Военно-медицинской академии имени С.М.Кирова. Рассматриваются 51 пациента с сочетанной травмой грудной клетки, госпитализированные в экстренном порядке. База содержит общую информацию о пострадавших, результаты лабораторных и инструментальных исследований, проведенных в течение первых 12 часов с момента поступления в больницу пострадавшего. Также, предоставлена информация о том, оказался ли исход летальным или нет для каждого пациента после получения увечий. С математической точки зрения это означает, что базу данных можно рассматривать как обучающую выборку, на основе которой имеется возможность построить функцию, дающую прогноз о степени серьезности повреждений для потенциального пострадавшего на основе различных показателей. С медицинской точки зрения, задача заключается в выявлении признаков, способствующие оценке критического состояния больных.

Таким образом, имеется 51 наблюдения и 261 признак. Как уже отмечалось выше, такое соотношение числа наблюдений и признаков влечет за собой большую ошибку при непосредственной классификации. Поэтому, необходимо перейти к пространству признаков меньшей размерности. Так как число переменных превышает число объектов в несколько раз, будет целесообразным провести процедуру отбора признаков с помощью пошагового дискриминантного анализа, разделив базу на несколько групп и в каждой найти «хорошо классифицирующие» переменные. После этого, отобранные признаки будут объединены. Если число признаков все равно будет избыточным, тогда для полученной совокупности можно также провести отбор признаков.

Обучение на выборке будет производится методами линейного и квадратичного дискриминантного анализа. Теоретически, если не отклоняется гипотеза о равенстве матриц, следует использовать формулы линейной дискриминации, в ином случае — квадратичной. Однако, в работе будет применяться как линейный, так и квадратичный анализ независимо от результатов проверки гипотезы. Следует заметить, что на практике проверка гипотез о нормальности и равенстве ковариационных матриц или дисперсий, в одномерном случае, скорее формальность. Такого объема выборки все-таки не достаточно, чтобы говорить о нормальном распределении рассматриваемых популяций. Следовательно, критерий равенства ковариационных матриц, который основан на предположении о нормальности выборок, также не может давать по-настоящему достоверных результатов. Поэтому в работе все данные будут классифицироваться как линейным, так и квадратичным дискриминантным анализом. Но несмотря на это, проверка на выполнение условий дискриминантного анализа будет осуществляться. При этом, критерий равенства ковариационных матриц будет применяться даже в случае отклонений от нормальности популяций. Все вышесказанное позволит наглядно сравнить работу двух подходов дискриминантного анализа в различных ситуациях.

В работе рассматриваются современные методы, такие как cross-validation leave-one-out, bootstrap leave-one-out, bootstrap 0.632 и 0.632+, которые позволяют более достоверно оценить величину ошибочной классификации, по сравнению с классическим подходом — вычисление ошибки на обучении. Таким образом, можно дать качественную оценку работы обучающего алгоритма даже в случае невыполнение условий дискриминантного анализа. Для проведения всех вычислений была написана программа на языке R, код которой доступен в приложении работы.

3.2 Предварительная очистка данных

Перед проведением самого анализа, проводится очистка данных от тех признаков, включение в модель которых не является целесообразным. Сначала были удалены из рассмотрения те признаки, для которых харак-

терно наличие довольно большего числа пропущенных данных. Всего было выделено 38 таких признаков.

В базе данных содержался раздел с анамнестическими данными, который пришлось удалить полностью, в связи с тем, что имелся существенный недостаток количества информации по невыжившим больным. Из результатов электрокардиограммы было решено убрать из рассмотрения показания длительности R–R интервала, так как данная величина связана обратной зависимостью с частотой сердечных сокращений. Фактически, имеются два зависимых признака, что может привести к необратимости ковариационной матрицы. Поэтому в исследовании R–R интервал не рассматривается.

Также были удалены данные, не содержащие в себе никакой важной информации. Например, ни у одного из пострадавших после инцидента не возникло осложнений, связанных с функционированием желудочно-кишечного тракта. Помимо этого, в данных изначально содержалась информация, полученная уже впоследствии. Например, количество дней, проведенных в реанимации. Такие данные не представляют интереса и будут удалены, так как в работе была поставлена задача провести анализ только за счет информации, полученной в первые 12 часов от поступления.

3.3 Пошаговый отбор признаков

3.3.1 Этап I

После проведения вышесказанных действий, имеется 158 признаков. Было решено разбить базу на одиннадцать групп. Для удобства, группы были выбраны тематически:

Группа 1. Общие данные.
Группа 2. Объективный статус при поступлении и шкалы.
Группа 3. Шкалы.
Группа 4. Структура повреждений внутренних органов.
Группа 5. Параметры искусственной вентиляции легких (ИВЛ) и мониторов витальных функций.

Группа 6. Вариабельность сердечного ритма (ВСР), вариабельность систолического артериального давления (ВСАД) и вариабельность диастолического артериального давления (ВДАД).
Группа 7. Вариабельность дыхания и общие данные спироартериокардиоритмографа (САКР).
Группа 8. Электрокардиография (ЭКГ).
Группа 9. Общий анализ крови (ОАК).
Группа 10. Биохимический анализ крови.
Группа 11. Маркеры повреждения сердца и анализ газов крови.

Перед проведением отбора признаков из данных удаляются те наблюдения, которые содержат пропущенные данные, поэтому в каждой группе будет разное количество объектов. Было решено не проводить сам дискриминантный анализ перед отбором признаков. Это было бы полезным в целях сравнения результатов до и после отбора. Однако, в большинстве групп содержится порядка двадцати признаков, тогда как наблюдений — около 40, а где-то и меньше. Такое соотношение значительно затрудняет проведение дискриминантного анализа. На рис. 3.1 приведена таблица результатов применения пошагового дискриминантного анализа.

Группа	**Признак 1**	**Признак 2**	**Признак 3**	**Признак 4**	**Признак 5**	Λ_{total}
Группа 1	Давность травмы					0,85
Группа 2	ВПХ - П (МТ)	Речевой контакт	Величина кровопотери	САД при поступлении	Частота пульса При поступлении	0,4
Группа 3	ВПХ - голова	ВПХ - грудь	ВПХ - таз	AIS - грудь	AIS - таз	0,41
Группа 4	Повреждение ЦНС САК					0,85
Группа 5	ИВЛ ДАД					0,92
Группа 6	ВСР LF (n. u.)	Вар. САД ТР				0,65
Группа 7	САКР PQ	САКР ЧССср	САКР САДмакс	САКР ДАДср	САКР Вар. Дых. HF n.	0,36
Группа 8	ЭКГ ЧСС					0,92
Группа 9	ОАК Гемоглобин					0,77
Группа 10	Биохимия Натрий					0,63
Группа 11	Газы крови FiO2	Газы крови RI				0,67

Рис. 3.1: Результат отбора признаков с итоговыми статистиками лямбда Уилкса

В последнем столбце для каждой группы представлено значение ста-

тистики лямбда Уилкса после окончания процедуры пошагового отбора, то есть значение лямбды от совокупности отобранных переменных в рассматриваемой группе. Чем ближе значение статистики к нулю, тем больше вклад подгруппы признаков в отклонение нулевой гипотезы дисперсионного анализа. Однако, это совсем не означает, что одна подгруппа признаков будет лучше классифицировать данные, чем другая подгруппа, если значении ее статистики меньше. Убедиться в этом можно по результатам оценок величины ошибочной классификации на каждой полученной подгруппе, которые представлены на рис. 3.2.

Группа	Метод	Ошибка на обучении	Cross-validation Leave-one-out	Bootstrap Leave-one-out	Bootstrap 0.632	Bootstrap 0.632+	Нормальность популяции 1	Нормальность популяции 2	Равенство ковариационных матриц
Группа 1	ЛДА	0,31	0,333	0,354	0,338	0,339	Нет	Да	Нет
	КДА	0,31	0,333	0,346	0,333	0,334			
Группа 2	ЛДА	0,118	**0,157**	**0,177**	**0,155**	**0,157**	Нет	Нет	Да
	КДА	-	-	-	-	-			
Группа 3	ЛДА	0,178	0,222	0,242	0,219	0,221	Нет	Нет	Нет
	КДА	-	-	-	-	-			
Группа 4	ЛДА	0,280	0,280	0,305	0,296	0,296	Нет	Нет	Да
	КДА	0,280	0,280	0,289	0,286	0,286			
Группа 5	ЛДА	0,294	0,324	0,344	0,326	0,327	Да	Да	Нет
	КДА	0,327	0,347	0,376	0,358	0,36			
Группа 6	ЛДА	0,227	0,25	0,255	0,245	0,245	Да	Нет	Нет
	КДА	0,182	0,182	0,221	0,207	0,208			
Группа 7	ЛДА	**0,074**	0,185	0,202	**0,155**	0,163	Нет	Нет	Нет
	КДА	-	-	-	-	-			
Группа 8	ЛДА	0,333	0,357	0,383	0,365	0,366	Да	Да	Нет
	КДА	0,262	0,31	0,346	0,315	0,319			
Группа 9	ЛДА	0,362	0,426	0,385	0,376	0,377	Нет	Да	Да
	КДА	0,362	0,426	0,387	0,378	0,378			
Группа 10	ЛДА	0,267	0,267	0,291	0,282	0,282	Да	Да	Нет
	КДА	0,233	0,233	0,281	0,264	0,265			
Группа 11	ЛДА	0,250	0,296	0,309	0,287	0,289	Да	Да	Нет
	КДА	0,273	0,318	0,325	0,306	0,308			

Рис. 3.2: Оценки величины ошибочной классификации линейным и квадратичным дискриминантным анализом для каждой группы.

Прочерки в таблице означают, что программе не удалось построить квадратичное дискриминационное правило. Такое наблюдается в группах 2, 3 и 7. Во всех названных группах пошаговая процедура оставила по 5 признаков. Вероятно, именно по этой причине построить модель классификации не удалось, так как для построения классификационного правила в этом случае требуется оценить 41 параметр, что очень много для выборок

в 40–50 наблюдений. В группах 1, 4, 9 квадратичная и линейная дискриминации показали практически идентичные результаты. Линейный дискриминантный анализ в группах 5 и 11 показал результаты лучше, чем квадратичный анализ. Обратная ситуация в группах 6, 8, 10 — квадратичная дискриминация здесь оказалась более точной. Наименьшую ошибку показал линейный дискриминантный анализ на отобранных признаках группы 2. Также стоит отметить ЛДА на признаках групп 3 и 7, а также КДА на переменных из группы 6.

3.3.2 Этап II

После первого этапа отбора имеется 25 признаков. Такое количество все равно избыточно, поэтому имеющиеся признаки разделим на три группы. Все отобранные признаки из групп 1–3 образуют группу A, из групп 4–6 — группу B, из групп 7–11 — группу C. В каждой группе проводится пошаговый отбор. Полученные признаки объединяются в финальную группу F, в которой также происходит отбор признаков. Результаты отбора и точности классификаторов представлены на рис. 3.3 и рис. 3.4 соответственно.

Группа	Признак 1	Признак 2	Признак 3	Признак 4	Признак 5	Признак 6	Λ_{total}
Группа A	ВПХ - П (МТ)	Речевой контакт					0,38
Группа B	ВСР LF (n. u.)	Вар. САД ТР					0,67
Группа C	САКР PQ	ОАК Гемоглобин	САКР Вар. Дых. HF n.u.	САКР ДАДср	САКР САДмакс	Газы крови FiO2	0,14
Группа F	ВСР LF (n. u.)	Вар. САД ТР	САКР Вар. Дых. HF n.u.	САКР САДмакс	ОАК Гемоглобин	САКР ДАДср	0,23

Рис. 3.3: Финальный отбор признаков.

Как видно из рис. 3.4, не квадратичный анализ не удалось провести в группах C и F, которые содержат по 6 признаков. Анализируя группы 1–11 и A, B, C и F, в итоге получается, что в среднем лучшего результаты добились отобранные признаки из группы F, а именно:

— мощность спектра низких частот вариабельности сердечного ритма,

— общая мощность спектра вариабельности систолического артериального давления,

Группа	Метод	Ошибка на обучении	Cross-validation Leave-one-out	Bootstrap Leave-one-out	Bootstrap 0.632	Bootstrap 0.632+	Нормальность популяции 1	Нормальность популяции 2	Равенство ковариационных матриц
Группа A	ЛДА	0,157	0,196	0,194	0,18	0,181	Нет	Нет	Да
	КДА	0,176	0,176	0,193	0,187	0,187			
Группа B	ЛДА	0,227	0,25	0,255	0,245	0,245	Да	Нет	Нет
	КДА	0,182	0,182	0,221	0,207	0,208			
Группа C	ЛДА	0,12	0,16	0,224	0,186	0,192	Да	Нет	Да
	КДА	-	-	-	-	-			
Группа F	ЛДА	0,037	0,185	0,184	0,13	0,14	Да	Нет	Да
	КДА	-	-	-	-	-			

Рис. 3.4: Оценка величины ошибки в группах A, B, C и F.

— мощность спектра высоких частот вариабельности дыхания,

— максимальный уровень систолического артериального давления, измеренного на САКР,

— среднее диастолическое артериальное давление, измеренное на САКР,

— концентрация гемоглобина в крови.

На основе сведений из медицины, описанных в главе 2, можно проанализировать адекватность результатов применения пошагового дискриминантного анализа. Как было отмечено, артериальное давление является важным показателем в определении степени тяжести травматического шока, уровня кровотока тканей; вариабельность сердечного ритма и частота дыхания представляют собой одни из важнейших показателей, характеризующих состояние организма; по уровню гемоглобина можно судить о нарушении газообмена. Таким образом, использовав пошаговую процедуру отбора признаков, был построен классификатор с хорошей точностью (81,5–87%), который является правдоподобным с точки зрения медицины.

3.3.3 Сравнение с результатами другого исследования

В прошлом году рассматриваемая база данных уже исследовалась [35]. Тогда, на основе рекомендаций врачей для классификации были выбраны следующие признаки:

— мощность спектра низких частот вариабельности сердечного ритма,

– мощность спектра низких частот вариабельности систолического артериального давления,

– мощность спектра очень низких частот вариабельности систолического артериального давления,

– показатель вегетативного баланса вариабельности систолического артериального давления,

– концентрация мочевины в крови,

– концентрация креатина в крови.

По сравнению с прошлым годом, база данных претерпела небольшие изменения: были добавлены несколько новых пациентов. Также, в той работе величина ошибки оценивалась только посредством ошибки на обучении, а классифицировались данные только линейным дискриминантным анализом. В связи с этим, была заново построена линейная дискриминационная модель, а точность была оценена с помощью всех подходов, рассматриваемых в настоящей работе. Были получены следующие результаты:

Метод	Ошибка на обучении	Cross-validation Leave-one-out	Bootstrap Leave-one-out	Bootstrap 0.632	Bootstrap 0.632+
ЛДА	0,25	0,325	0,332	0,302	0,306

Рис. 3.5: Оценка точности прошлогоднего исследования.

Как видно из результатов, пошаговой процедуре удалось найти набор признаков, улучшающих точность классификации примерно на 16%, по сравнению с набором признаков, рекомендованным врачами.

3.4 Дополнительный анализ с целью улучшения результата

В этой секции предпринимается попытка увеличить точность классификации, за счет нахождения более оптимального набора признаков, чем набор, который был получен на основе пошагового отбора. Идея о том, что

полученный после отбора набор может быть не оптимальным, имеет медицинские предпосылки. В финальную модель не вошла ни одна из медицинских шкал, которые создаются с целью оценки состояния пострадавшего и являются сами по себе предсказательными. Также, для пострадавшего очень важно, через какое время после получения травмы ему будет оказана срочная медицинская помощь. В связи с этим рассматриваются признаки: давность травмы в минутах, шкала Военно-Полевой Хирургии П(МТ) и речевой контакт в баллах, которые хорошо себя зарекомендовали в течение отбора признаков. После различных экспериментов, было решено также рассмотреть возраст пострадавшего. Таким образом, на рис. 3.6 представлены результаты восьми экспериментальных наборов признаков.

Признаки	Наличие в модели выброса	Ошибка на обучении	Cross-validation Leave-one-out	Bootstrap Leave-one-out	Bootstrap 0.632	Bootstrap 0.632+
Давность травмы, речевой контакт, ВПХ МТ(П)	Да	0,095	0,143	0,13	0,117	0,118
	Нет	0,073	0,098	0,125	0,106	0,107
Возраст, давность травмы, речевой контакт, ВПХ МТ(П)	Да	0,048	0,071	0,117	0,092	0,094
	Нет	0,024	0,073	0,09	**0,066**	**0,067**
Возраст, давность травмы, речевой контакт, ВПХ МТ(П), Группа F	Да	**0**	0,217	0,244	0,154	0,181
	Нет	**0**	0,182	0,207	0,131	0,149
Речевой контакт, ВПХ МТ(П), ВСР LF (n. u.)	–	0,159	0,205	0,243	0,212	0,216
Возраст, давность травмы, речевой контакт, ВПХ МТ(П), гемоглобин	Да	0,05	0,1	0,107	0,086	0,087
	Нет	0,051	**0,051**	**0,089**	0,075	0,076
Возраст, давность травмы, речевой контакт, ВПХ МТ(П), ВСР LF (n. u.)	Да	0,054	0,135	0,141	0,109	0,112
	Нет	0,028	**0,056**	**0,087**	**0,065**	**0,066**
Возраст, давность травмы, речевой контакт, ВПХ МТ(П), САКР ДАДср, гемоглобин	Да	0,057	0,086	0,138	0,108	0,111
	Нет	0,029	0,059	0,095	0,071	0,072
Возраст, давность травмы, речевой контакт, ВПХ МТ(П), САКР ДАДср	Да	0,054	0,081	0,121	0,096	0,098
	Нет	0,028	0,083	0,098	0,072	0,074

Рис. 3.6: Оценки величины ошибки.

В базе данных есть пациент, который попал в руки врачей только через 36 часов с момента получения травмы. Этот показатель резко отличается от других пострадавших, до которых врачебная помощь добиралась, как правило, в течение первых двух часов. Стоит ли оставлять наблюдение с таким выбросом, или нет, вопрос спорный. Поэтому в приведенных вычислениях для каждой группы рассматриваются классификационные мо-

дели, построенные с и без данного наблюдения. Как видно из результатов, во всех случаях из таблицы удаление этого наблюдение давало улучшение точности, поэтому следует такой объект убрать из рассмотрения.

Результаты, представленные на рис. 3.6 вызывают большой интерес. Сочетание признаков: давность травмы, речевой контакт и ВПХ-П(МТ) дало уже результат лучше, чем полученный после отбора набор признаков (группа F). Добавив же возраст в группу признаков, ошибка становится еще меньше. Затем были объединены 4 рассматриваемых признака с группой F. В этом случае ошибка на обучении оказалась равна нулю, тогда как на других методах оценка варьировалась от 13,1% до 20,7%. Это свидетельствует о том, что ошибка на обучении может иногда вводить в заблуждение. После этого на данной группе признаков был применен пошаговый дискриминантный анализ, который оставил такие признаки, как: речевой контакт, ВПХ-П(МТ) и ВСР LF(n.u.). Как это ни странно, но ошибка только увеличилась в сравнении с ошибкой на наборе, который был до проведения процедуры отбора. В дальнейшим были проделаны различные эксперименты сочетаний некоторых признаков из группы F и 4 признаков, введенных в рассмотрения в этой секции. В таблице приведены 4 таких сочетания. Оказалось, что это принесло успех: при присоединении показателя уровня гемоглобина и мощности низкочастотного спектра вариабельности сердечного ритма по отдельности, но не вместе, дало наилучшие в работе результаты. При добавлении гемоглобина точность классификатора варьируется от 91,1% до 94,9%, при добавлении ВСР LF(n.u.) — от 91,3% до 94,4%.

Интересно, что при классификации оказалась велика роль возраста пациента. Сказать точно, с чем это связано, не представляется возможным. Не исключено, что при значительном увеличении размера выборки этот признак выпадет из рассмотрения. Хотя, если посмотреть на рис. 3.7, то можно видеть, что какой-то прямой зависимости между возрастом и исходом нету. Красным цветом здесь обозначены те пациенты, которые не выжили, а синим — те, которые были выписаны. Возможно, именно в комбинации с другими признаками возраст дает ценную информацию, и возможно именно поэтому пошаговая процедура не отбирала этот признак, ведь она ориентируется на личный информационный вклад переменной в

отклонение нулевой гипотезы.

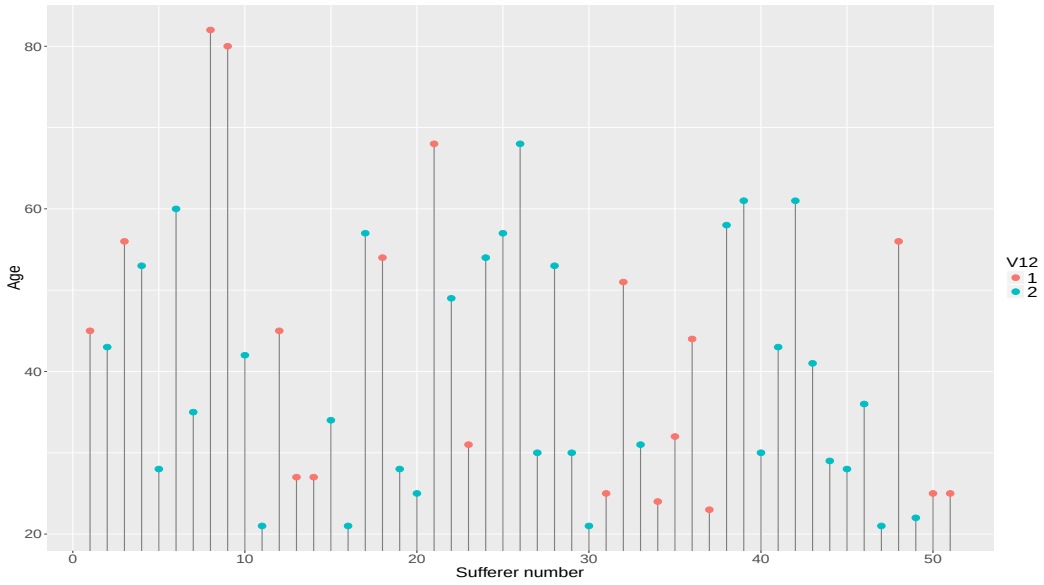

Рис. 3.7: Возраст пациентов.

3.5 Сравнение методов оценки величины ошибки

После проделанных всех вычислений имеется возможность провести сравнительный анализ всех рассматриваемых в этой работе подходов к оценке величины ошибки. Для этого на рис. 3.8 были изображены графически результаты оценок ошибочной классификации линейным дискриминантным анализом на группах 1–11, рассматриваемых в секции 3.3.1. Точки на графике были соединены прямыми для наглядности результатов.

Как и ожидалось, из графика видно, что ошибка на обучении имеет тенденцию оптимистически занижать величину ошибки, в то время как bootstrap leave-one-out — пессимистически завышать. Cross-validation leave-one-out, bootstrap 0.632 и bootstrap 0.632+ показали приблизительно одинаковые результаты, разве лишь за исключением группы 9, где кросс-валидация пессимистичнее всех оценила ошибку. Следовательно, можно предположить, что истинное значение величины ошибки лежит в «коридоре» между bootstrap leave-one-out и ошибкой на обучении, ближе к оценкам cross-validation leave-one-out, bootstrap 0.632 и 0.632+. Последние два

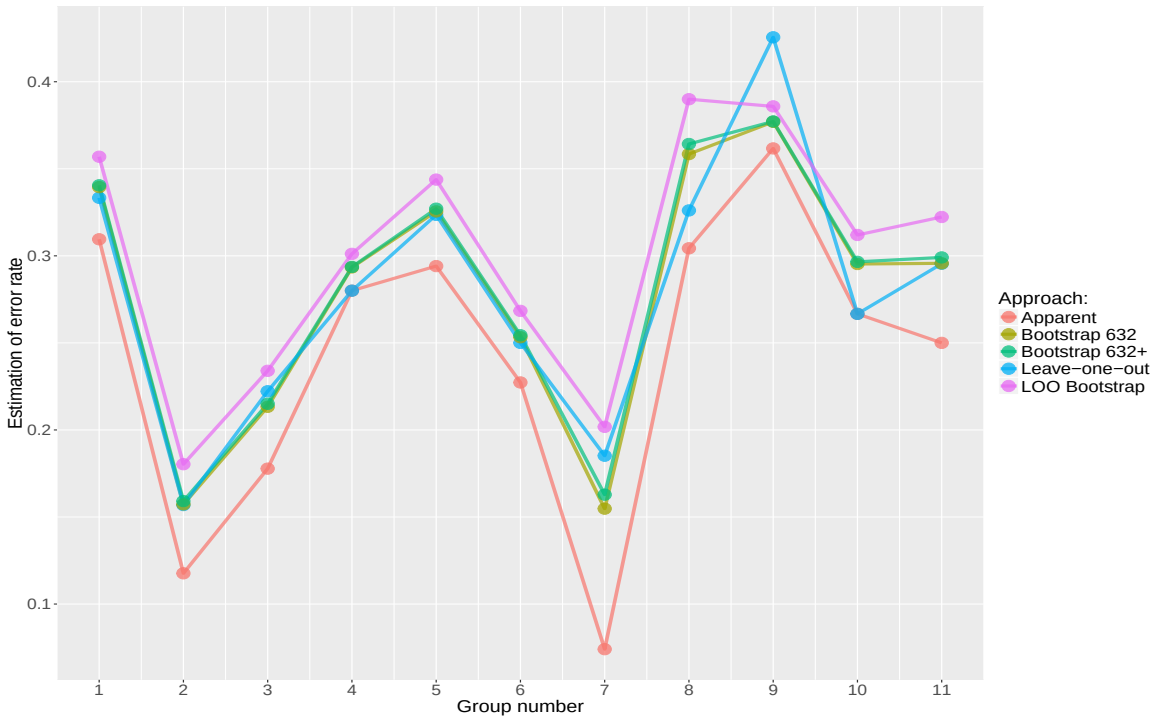

Рис. 3.8: Графическое изображение оценок величины ошибки на группах 1–11.

названных метода продемонстрировали практически идентичные результаты. Это говорит о том, что относительная частота обучения, описанная в секции 1.4.6, была близка к нулю, и в модели не наблюдается большого влияния эффекта переобучения.

Дальше был построен изображенный на рис. 3.9 аналогичный график, но уже на экспериментальных группах признаков, описанных в секции 3.4. В сравнении с рис. 3.8, здесь более наглядно изображена разница между методами. С одной стороны это, связано с большем масштабом, чем на рис. 3.8. Но имеется и другая причина. Цифрой 3 на графике обозначена третья на рис. 3.6 группа, которая состоит из 10 признаков. Можно видеть, что «коридор» на этой группе очень широкий, тогда как на группах под остальными цифрами, где содержится меньшее число признаков, этот «коридор» сужается. Конечно, это может быть связано с количеством наблюдений, ведь для группы из 10 признаков пришлось удалить большее число объектов, имевших пропущенные значения. Однако, можно предположить, что на методы оценки величины ошибки обученной модели влияет также соотношение числа наблюдений и числа признаков, которое, как известно, оказывает значимое влияние на методы классификации. При этом оказалось, что ошибка на обучении и bootstrap leave-one-out очень чувстви-

тельны к этому соотношению. Подтверждает сделанное предположение и рис. 3.8, где некоторые группы имели только по одному признаку, поэтому «коридор» в этих местах был довольно узок.

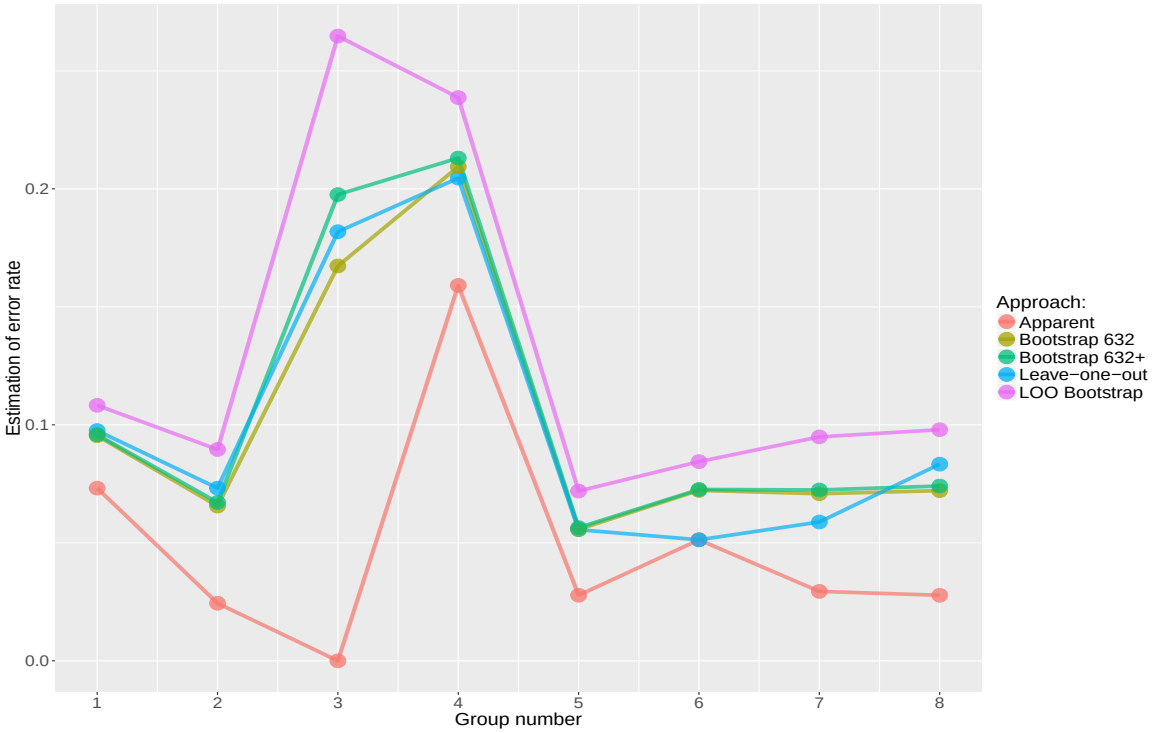

Рис. 3.9: Графическое изображение оценок величины ошибки на группах признаков из раздела 3.4.

Выводы

Таким образом, после всех проделанных вычислений, можно выделить признаки, которые по результатам данной работы способствуют оценке критического состояния пострадавшего с сочетанной травмой груди:

- мощность спектра низких частот вариабельности сердечного ритма,
- общая мощность спектра вариабельности систолического артериального давления,
- мощность спектра высоких частот вариабельности дыхания,
- максимальный уровень систолического артериального давления, измеренного на САКР,
- среднее диастолическое артериальное давление, измеренное на САКР,
- концентрация гемоглобина в крови,
- количество минут, прошедшее с момента получения травмы во время прибытия скорой помощи,
- шкала военной-полевой хирургии П(МТ) (П–повреждение, МТ–механическая травма),
- речевой контакт,
- возраст пострадавшего.

Говоря о математических выводах, можно отметить роль пошагового дискриминантного анализа в получении результатов, даже несмотря на тот, факт, что процедура не получила оптимального набора признаков. Те результаты, которые мы получили уже после отбора признаков, в ходе различных экспериментов с выбором признаков, были получены благодаря тому же отбору. Возможно, для большей эффективности в исследованиях

такого масштаба следует применять даже несколько процедур отбора, которые при этом основаны на разных критериях выбора признаков.

Сравнивая работу линейного и квадратичного дискриминантного анализа, можно прийти к выводу, что линейная дискриминация оказывается более предпочтительной, так как на выборках малого размера, квадратичная дискриминация сильно ограничена в возможностях. Исключение составляет ситуация, когда количество признаков в модели минимальное (1–2). В этом случае квадратичный анализ может оказаться более эффективным, чем линейный, как это наблюдалось в группах 6, 8, 10 из секции 3.3.1.

Анализ методов оценки величины ошибки на выборках малых размеров является непростой задачей. Однако, по результатам данной исследования, неплохо себя зарекомендовали методы cross-validation leave-one-out, bootstrap 0.632 и 0.632+, тогда как ошибка на обучении не отличалась стабильностью при увеличении количества признаков в модели. Bootstrap leave-one-out, с одной стороны, также заметно отклонился от других методов в случае 10 признаков. С другой стороны, в остальных случаях данный метод можно рассматривать в качестве верхней оценки величины ошибки.

Заключение

В выпускной квалификационной работе проводилось исследование базы данных пострадавших с сочетанной травмой груди. Для непосредственной классификации использовались линейный и квадратичный дискриминантный анализ, для отбора признаков — пошаговый дискриминантный анализ, для оценки величины ошибки — классический подход вычисления ошибки на обучении и современные методы cross-validation leave-one-out, bootstrap leave-one-out, bootstrap 0.632 и bootstrap 0.632+. Благодаря такому подходу, удалось добиться довольно высокой точности (81,5–87%) в ситуации, когда число признаков превышает количество наблюдений. Данный результат оказался лучше, по сравнению с прошлогодним исследованием тех же данных. После проведения пошагового анализа были проведены различные эксперименты с целью увеличения точности классификации. В результате, удалось найти более оптимальный набор признаков, на котором достигается еще более высокая точность (91,1–94,9%).

Список литературы

[1] Буре В. М., Щербакова А. А. Применение дискриминантного анализа и метода деревьев принятия решений для диагностики офтальмологических заболеваний // Вестник Санкт-Петербургского университета. Серия 10: Прикладная математика. Информатика. Процессы управления. 2013. № 1. С. 70–76.

[2] Dudoit S., Fridlyand J., Speed T. P. Comparison of discrimination methods for the classification of tumors using gene expression data // Journal of the American Statistical Association. 2002. Vol. 97 (457). P. 77–87.

[3] Hand D. J., Henley W. E. Statistical Classification Methods in Consumer Credit Scoring: A Review // Journal of the Royal Statistical Society. Series A (Statistics in Society). 1997. Vol. 160 (3). P. 523–541.

[4] Мальцева А. В., Шилкина Н. Е., Махныткина О. В. Data minig в социологии: опыт и перспективы проведения исследования // Социологические исследования. 2016. № 3. С. 35–44.

[5] Афифи А., Эйзен С. Статистический анализ: Подход с использованием ЭВМ / пер. с англ. Енюкова И. С. и Новикова И. Д. / под ред. Башарина Г. П. М.: Мир, 1982. 488 с.

[6] Рао С. Р. — Линейные статистические методы и их применения / науч. ред. Линник Ю. В. / пер. с англ. Калинина В. М. и др. М.: Наука, 1968. 548 с.

[7] Fisher R. A. The use of multiple measurements in taxonomic problems // Annals of Eugenics. 1936. № 7. P. 179–188.

[8] Lachenbruch P. A. Some unsolved practical problems in discriminant analysis. Chapel Hill: University of North Carolina, 1975. 10 p.

[9] Кобзарь А. И. Прикладная математическая статистика. Для инженеров и научных работников. М.: ФИЗМАТЛИТ, 2006. 816 с.

[10] Rencher A. C. Methods of Multivariate Analysis. 2nd Ed. New York: John Wiley & Sons, Inc., 2002. 738 p.

[11] Воронцов К. В. Лекции по статистическим (байесовским) алгоритмам классификации.
http://www.machinelearning.ru/wiki/images/e/ed/Voron-ML-Bayes.pdf

[12] Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Ed. New York: Springer-Verlag, 2009. 745 p.

[13] Воронцов К. В. Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики. 2004. № 13. С. 5–36.

[14] Lachenbruch P. A., Mickey M. R. Estimation of error rates in discriminant analysis // Technometrics. 1968. № 10 (1) P. 1–11.

[15] Molinaro A. M., Simon R., Pfeiffer R. M. Prediction error estimation: a comparison of resampling methods // Bioinformatics. 2005. Vol. 21 (15). P. 3301–3307.

[16] Braga-Neto U. M., Dougherty E. R. Is cross-validation valid for small-sample microarray classification? // Bioinformatics. 2004. Vol. 20 (3). P. 374–380.

[17] Эфрон Б. Нетрадиционные методы многомерного статистического анализа: Сб. статей / пер. с англ. / предисловие Адлера Ю. П., Кошевника Ю. А. М.: Финансы и статистика, 1988. 263 с.

[18] Efron B. Estimating the error rate of a prediction rule: improvement on Cross-Validation // Journal of the American Statistical Association. 1983. Vol. 78 (382). P. 316–331.

[19] Efron B., Tibshirani R. Improvements on Cross-Validation: The .632+ Bootstrap Method // Journal of the American Statistical Association. 1997. Vol. 92 (438). P. 548–560.

[20] Vu T., Sima C., Braga-Neto U. M., Dougherty E. R. Unbiased bootstrap error estimation for linear discriminant analysis // EURASIP Journal on Bioinformatics and Systems Biology. 2014. Vol. 2014 (15). P. 1–14.

[21] Нестеров А. В. Состояние вопроса травмы внутри салона автомобиля при ДТП // Избранные вопросы судебно-медицинской экспертизы, 2007. №82. С. 10–22.

[22] Соколов В. А. Множественные и сочетанные травмы (практическое руководство для врачей травматологов). М.: ГЭОТАРМедиа, 2006. С. 512

[23] Altieri R., Citarelli C., Cofano F., Zenga F., Ducati A., Garbossa D. — Concomitant Thoracic and Spinal Injuries in Politraumatized Patient, a Frequent but Few Discussed Entity. A Case Report. // Journal of Universal Surgery, 2015. Vol. 3 (5).

[24] Гринцов А. Г., Куницкий Ю. Л., Христуленко А. А. Особенности клиники и диагностики при сочетанной травме груди и живота. // Травма, 2012. Т. 13(4). С. 154–156.

[25] Вагнер Е. А. Хирургия повреждений груди. М.: Медицина, 1981. 288 с.

[26] Мусалатов Х. А. Хирургия катастроф. М.: Медицина, 1998. 592 с.

[27] Военно-полевая хирургия / под ред. проф. Гуманенко Е. К. 2-е издание. СПб: Изд-во Фолиант, 2008. 464 с.

[28] Teasdale G. & Jennett B. Assessment of coma and impaired consciousness. A practical scale. // Lancet, 1974. Vol. 2(7872) P. 81-84

[29] Зудбинов Ю. И. Азбука ЭКГ. Изд. 3-е. Ростов-на-Дону: изд–во «Феникс», 2003. 160 с.

[30] Руководство по кардиологии / под ред. Коваленко В. Н. Киев: МОРИОН, 2008. 1424 с.

[31] Бокерия Л. А., Бокерия О. Л., Волковская И. В. Вариабельность сердечного ритма: методы измерения, интерпретация, клиническое использование. // Анналы Аритмологии, 2009. № 4. С. 21–32.

[32] Medweb. http://www.medweb.ru (дата обращения: 25.03.16).

[33] Хеннеси А. А. М., Джапп А. Д. Анализ газов артериальной крови понятным языком / пер. с англ. под ред. Кассиля В. Л. М.: Практическая медицина, 2009. 140 с.

[34] Medportal. http://medportal.ru (дата обращения: 27.03.16).

[35] Семенчиков Д. Н. Классификация больных с тяжёлой сочетанной травмой грудной клетки // Процессы управления и устойчивость. 2015. Т. 2. № 1. С. 317–321.

Приложение

Код программы, написанной на языке R:

```
library(MASS)
library(klaR)
library(ipred)
library(stats)
library(biotools)
library(methods)
library(nortest)
library(moments)
library(robustbase)
library(psych)
library(plyr)
library(robCompositions)
library(mvoutlier)
library(MVN)
library(andrews)
library(Daim)
library(ggplot2)

mypredict <- function(object, newdata)
  predict(object, newdata = newdata)$class

mylda <- function(formula, train, test){
  model <- lda(formula, train)
  predict(model, test)$posterior[,"pos"]
}

myqda <- function(formula, train, test){
  model <- qda(formula, train)
```

```
  predict(model, test)$posterior[,"pos"]
}

estimate_err <- function(formula, y, df, q=F, B=50){
  a<-cv(y, formula, data = df, model=lda,
k=length(y),predict=mypredict)$error
  d<-Daim(formula, data = df, model=mylda,
control=Daim.control(method="boot", number=B))
  if(q==T){
    b<-cv(y, formula, data = df, model=qda,
k=length(y),predict=mypredict)$error
    e<-Daim(formula, data = df, model=myqda,
control=Daim.control(method="boot", number=B))
    res<-c(d$errapp,a,d$errloob,d$err632,d$err632p,
e$errapp,b,e$errloob,e$err632,e$err632p)
    names(res)<-c("LDA_apparent","LDA_L-O-O",
"LDA_L-O-O-B","LDA_B632","LDA_B632+",
"QDA_apparent","QDA_L-O-O","QDA_L-O-O-B",
"QDA_B632","QDA_B632+")
    return(res)
  } else{
    res<-c(d$errapp,a,d$errloob,d$err632,d$err632p)
    names(res)<-c("LDA_apparent","LDA_L-O-O",
"LDA_L-O-O-B","LDA_B632","LDA_B632+")
    return(res)
  }
}

df <- read.csv("Обновленная база.csv",
header = FALSE, sep = ";")
df$V12 <- as.factor(df$V12)

firstgroup <- df[,c((1:4),10,11,12)]
```

```
firstgroup <- na.omit(firstgroup)
firstgroup <- firstgroup[-18,]
greedy.wilks(firstgroup[,-7],
firstgroup$V12,niveau=0.1)
firstgroup <- df[,c(3,12)]
firstgroup <- na.omit(firstgroup)
shapiro.test(subset
(firstgroup[,-2],firstgroup$V12==1))
shapiro.test(subset
(firstgroup[,-2],firstgroup$V12==2))
bartlett.test(firstgroup[,-2],firstgroup[,2])
err1<-estimate_err(V12~., df = firstgroup, y=firstgroup$V12)
err12<-estimate_err(V12~.,
df = firstgroup, y=firstgroup$V12,q=T)

secondgroup <- df[,c((22:38),12)]
secondgroup <- na.omit(secondgroup)
greedy.wilks(secondgroup[,-18],secondgroup$V12,niveau=0.1)
secondgroup <- df[,c(25,30,31,32,37,12)]
secondgroup <- na.omit(secondgroup)
mardiaTest(subset
(secondgroup[,-6], secondgroup$V12==1))
mardiaTest(subset
(secondgroup[,-6], secondgroup$V12==2))
boxM(secondgroup[,-6],secondgroup$V12)
err2<-estimate_err(V12~.,
df = secondgroup, y=secondgroup$V12)
err22<-estimate_err(V12~.,
df = secondgroup, y=secondgroup$V12,q=T)

thirdgroup <- df[,c(39,(43:54),12)]
thirdgroup <- na.omit(thirdgroup)
greedy.wilks(thirdgroup[,-14],thirdgroup$V12,niveau=0.1)
```

```
thirdgroup <- df[,c(43,44,45,49,50,12)]
thirdgroup <- na.omit(thirdgroup)
thirdgroup$V49<-as.factor(thirdgroup$V49)
mardiaTest(subset
(thirdgroup[,-6], thirdgroup$V12==1))
mardiaTest(subset
(thirdgroup[,-6], thirdgroup$V12==2))
boxM(thirdgroup[,-6],thirdgroup$V12)
err3<-estimate_err(V12~.,
df = thirdgroup, y=thirdgroup$V12)
err32<-estimate_err(V12~.,
df = thirdgroup, y=thirdgroup$V12,q=T,B=10)

fourthgroup <- df[,c(79,(88:90),92,93,12)]
fourthgroup <- na.omit(fourthgroup)
greedy.wilks(fourthgroup[,-7],fourthgroup$V12,niveau=0.1)
fourthgroup <- df[,c(93,12)]
fourthgroup <- na.omit(fourthgroup)
shapiro.test(subset
(fourthgroup[,-2],fourthgroup$V12==1))
shapiro.test(subset
(fourthgroup[,-2],fourthgroup$V12==2))
bartlett.test(fourthgroup[,1],fourthgroup$V12)
err4<-estimate_err(V12~V93,
df = fourthgroup, y=fourthgroup$V12)
err42<-estimate_err(V12~V93,
df = fourthgroup, y=fourthgroup$V12,q=T)

fifthgroup <- df[,c((103:108),113,12)]
fifthgroup <- na.omit(fifthgroup)
greedy.wilks(fifthgroup[,-8],fifthgroup$V12,niveau=0.2)
fifthgroup <- df[,c(105,12)]
fifthgroup <- na.omit(fifthgroup)
```

```
shapiro.test(subset(fifthgroup[,-2],fifthgroup$V12==1))
shapiro.test(subset(fifthgroup[,-2],fifthgroup$V12==2))
bartlett.test(fifthgroup[,1],fifthgroup$V12)
err5<-estimate_err(V12~V105,
df = fifthgroup, y=fifthgroup$V12)
err52<-estimate_err(V12~V105,
df = fifthgroup, y=fifthgroup$V12,q=T)

sixthgroup <-
df[,c(120,(123:128),130,(132:135),137,139,140,12)]
sixthgroup <- na.omit(sixthgroup)
greedy.wilks(sixthgroup[,-16],sixthgroup$V12,niveau=0.1)
sixthgroup <- df[,c(124,127,12)]
sixthgroup <- na.omit(sixthgroup)
mardiaTest(subset
(sixthgroup[,-3], sixthgroup$V12==1))
mardiaTest(subset
(sixthgroup[,-3], sixthgroup$V12==2))
boxM(sixthgroup[,-3],sixthgroup$V12)
err61<-estimate_err(V12~V124+V127,
df = sixthgroup, y=sixthgroup$V12)
err62<-estimate_err(V12~V124+V127,
df = sixthgroup, y=sixthgroup$V12, q=T)

seventhgroup <- df[,c(141,142,(145:163),(165:168),12)]
seventhgroup <- na.omit(seventhgroup)
greedy.wilks(seventhgroup[,-26],seventhgroup$V12,niveau=0.1)
seventhgroup <- df[,c(146,148,156,157,161,12)]
seventhgroup <- na.omit(seventhgroup)
mardiaTest(subset(seventhgroup[,-6], seventhgroup$V12==1))
mardiaTest(subset(seventhgroup[,-6], seventhgroup$V12==2))
boxM(seventhgroup[,-6],seventhgroup$V12)
err7<-estimate_err(V12~.,
```

```
df = seventhgroup, y=seventhgroup$V12)
err72<-estimate_err(V12~.,
df = seventhgroup, y=seventhgroup$V12,q=T)

eighthgroup <- df[,c((170:173),(175:184),(186:190),194,12)]
eighthgroup <- na.omit(eighthgroup)
greedy.wilks(eighthgroup[,-21],eighthgroup$V12,niveau=0.11)
eighthgroup <- df[,c(175,12)]
eighthgroup <- na.omit(eighthgroup)
shapiro.test(subset
(eighthgroup[,-2],eighthgroup$V12==1))
shapiro.test(subset
(eighthgroup[,-2],eighthgroup$V12==2))
bartlett.test(eighthgroup[,1],eighthgroup$V12)
err8<-estimate_err(V12~.,
df = eighthgroup, y=eighthgroup$V12)
err82<-estimate_err(V12~.,
df = eighthgroup, y=eighthgroup$V12,q=T)

ninthgroup <- df[,c((195:203),(205:213),12)]
ninthgroup <- na.omit(ninthgroup)
greedy.wilks(ninthgroup[,-19],ninthgroup$V12,niveau=0.1)
ninthgroup <- df[,c(197,12)]
ninthgroup <- na.omit(ninthgroup)
shapiro.test(subset
(ninthgroup[,-2],ninthgroup$V12==1))
shapiro.test(subset
(ninthgroup[,-2],ninthgroup$V12==2))
bartlett.test(ninthgroup[,1],ninthgroup$V12)
err9<-estimate_err(V12~V197,
df = ninthgroup, y=ninthgroup$V12)
err92<-estimate_err(V12~V197,
df = ninthgroup, y=ninthgroup$V12,q=T)
```

```
tenthgroup <- df[,c(221,(223:229),232,233,12)]
tenthgroup <- na.omit(tenthgroup)
greedy.wilks(tenthgroup[,-11],tenthgroup$V12,niveau=0.2)
tenthgroup <- df[,c(233,12)]
tenthgroup <- na.omit(tenthgroup)
shapiro.test(subset(tenthgroup[,-2],tenthgroup$V12==1))
shapiro.test(subset(tenthgroup[,-2],tenthgroup$V12==2))
bartlett.test(tenthgroup[,1],tenthgroup$V12)
err10<-estimate_err(V12~V233,
df = tenthgroup, y=tenthgroup$V12)
err102<-estimate_err(V12~V233,
df = tenthgroup, y=tenthgroup$V12,q=T)

eleventhgroup <- df[,c(244,245,(248:255),261,262,12)]
eleventhgroup <- na.omit(eleventhgroup)
greedy.wilks(eleventhgroup[,-13],
eleventhgroup$V12,niveau=0.1)
eleventhgroup <- df[,c(261,262,12)]
eleventhgroup <- na.omit(eleventhgroup)
mardiaTest(subset
(eleventhgroup[,-3], eleventhgroup$V12==1))
mardiaTest(subset
(eleventhgroup[,-3], eleventhgroup$V12==2))
boxM(eleventhgroup[,-3],eleventhgroup$V12)
err112<-estimate_err(V12~V261+V262,
df = eleventhgroup, y=eleventhgroup$V12, q=T)
err11<-estimate_err(V12~V261+V262,
df = eleventhgroup, y=eleventhgroup$V12)

onethreegroups <- df[,c(3,25,30,31,32,37,43,44,45,49,50,12)]
onethreegroups <- na.omit(onethreegroups)
onethreegroups <- onethreegroups[-13,]
```

```
greedy.wilks(onethreegroups[,-12],
onethreegroups$V12,niveau=0.1)

onethreegroups_2 <- df[,c(25,37,12)]
onethreegroups_2 <- na.omit(onethreegroups_2)
onethreegroups_2 <- onethreegroups[-18,]
mardiaTest(subset
(onethreegroups_2[,-3], onethreegroups_2$V12==1))
mardiaTest
(subset(onethreegroups_2[,-3], onethreegroups_2$V12==2))
boxM(onethreegroups_2[,-3],onethreegroups_2$V12)
estimate_err(V12~.,
df = onethreegroups_2, y=onethreegroups_2$V12,q=T)

foursixgroups <- df[,c(93,105,124,127,12)]
foursixgroups <- na.omit(foursixgroups)
greedy.wilks(foursixgroups[,-5],
foursixgroups$V12,niveau=0.1)

restgroups <-
df[,c(146,148,156,157,161,175,197,233,261,262,12)]
restgroups <- na.omit(restgroups)
greedy.wilks(restgroups[,-11],restgroups$V12,niveau=0.2)
restgroups <- df[,c(146,156,157,161,197,262,12)]
restgroups <- na.omit(restgroups)
mardiaTest(subset(restgroups[,-7], restgroups$V12==1))
mardiaTest(subset(restgroups[,-7], restgroups$V12==2))
boxM(restgroups[,-7],restgroups$V12)
estimate_err(V12~., df = restgroups, y=restgroups$V12,q=T)

unite <- df[,c(25,37,124,127,146,156,157,161,197,262,12)]
unite <- na.omit(unite)
greedy.wilks(unite[,-11],unite$V12,niveau=0.2)
```

```
unite_2 <- df[,c(124,127,146,156,157,197,12)]
unite_2 <- na.omit(unite_2)
mardiaTest(subset(unite_2[,-7], unite_2$V12==1))
mardiaTest(subset(unite_2[,-7], unite_2$V12==2))
boxM(unite_2[,-7], unite_2$V12)
estimate_err(V12~., df = unite_2, y=unite_2$V12,q=T)

#Совет врачей
adv <- df[,c(124,128,130,133,221,225,12)]
adv <- na.omit(adv)
estimate_err(V12~., df = adv, y=adv$V12)

experiment_01<-df[,c(3,25,37,12)]
experiment_01 <- na.omit(experiment_01)
experiment_01<- experiment_01[-18,]
e01<-estimate_err(
V12~., df = experiment_01, y=experiment_01$V12)
experiment_02<-df[,c(1,3,25,37,12)]
experiment_02 <- na.omit(experiment_02)
experiment_02<- experiment_02[-18,]
e02<-estimate_err(
V12~., df = experiment_02, y=experiment_02$V12)
experiment_1<-df[,c(1,3,25,37,124,127,146,156,157,197,12)]
experiment_1 <- na.omit(experiment_1)
experiment_1<- experiment_1[-16,]
e1<-estimate_err(
V12~., df = experiment_1, y=experiment_1$V12, B=25)
greedy.wilks(experiment_1[-11],
experiment_1$V12,niveau = 0.2)
experiment_2<-df[,c(25,37,124,12)]
experiment_2 <- na.omit(experiment_2)
e2<-
estimate_err(V12~., df = experiment_2, y=experiment_2$V12)
```

```
experiment_3<-df[,c(1,3,25,37,124,12)]
experiment_3<- na.omit(experiment_3)
experiment_3<- experiment_3[-16,]
e3<-estimate_err(
V12~., df = experiment_3, y=experiment_3$V12)
experiment_4<-df[,c(1,3,25,37,197,12)]
experiment_4<- na.omit(experiment_4)
experiment_4<- experiment_4[-18,]
e4<-estimate_err(
V12~., df = experiment_4, y=experiment_4$V12)
experiment_5<-df[,c(1,3,25,37,157,197,12)]
experiment_5<- na.omit(experiment_5)
experiment_5<- experiment_5[-16,]
e5<-estimate_err(
V12~., df = experiment_5, y=experiment_5$V12)
experiment_6<-df[,c(1,3,25,37,157,12)]
experiment_6<- na.omit(experiment_6)
experiment_6<- experiment_6[-16,]
e6<-estimate_err(V12~.,
df = experiment_6, y=experiment_6$V12)
```

Построение графиков

```
age<-df[,c(1,12)]
age<-na.omit(age)
age$num<-1:51
ggplot(age, aes(y=V1, x=num)) +
geom_segment(aes(xend=num), yend=0, colour="grey50") +
geom_point(size=3, aes(colour=V12))+
xlab("Sufferer number")+
ylab("Age")
ggsave("Age.pdf", width=14, height=10)
```

```
results <- data.frame(err=
c(err1,err2,err3,err4,err5,err6,err7,err8,err9,err10,err11))
results$num <- 1:55
results$type[results$num%%5==1]<-"Apparent"
results$type[results$num%%5==2]<-"Leave-one-out"
results$type[results$num%%5==3]<-"LOO Bootstrap"
results$type[results$num%%5==4]<-"Bootstrap 632"
results$type[results$num%%5==0]<-"Bootstrap 632+"
results$number <- results$num%/%5+1
results$number[results$num%%5==0] <- 1:11
results<-results[,-2]
ggplot(results,aes(x=number,y=err,col=type))+
geom_line(size=1.5,alpha=0.7)+
geom_point(size=5,alpha=0.75)+
xlab("Group number")+
ylab("Estimation of error rate")+
labs(colour="Approach:")+
scale_x_continuous(breaks=c(1:11))
ggsave("graph_err.pdf", width=14, height=10)

impres<-data.frame(err=c(e01,e02,e1,e2,e3,e4,e5,e6))
impres$num <- 1:40
impres$type[impres$num%%5==1]<-"Apparent"
impres$type[impres$num%%5==2]<-"Leave-one-out"
impres$type[impres$num%%5==3]<-"LOO Bootstrap"
impres$type[impres$num%%5==4]<-"Bootstrap 632"
impres$type[impres$num%%5==0]<-"Bootstrap 632+"
impres$number <- impres$num%/%5+1
impres$number[impres$num%%5==0] <- 1:8
impres<-impres[,-2]
ggplot(impres,aes(x=number,y=err,col=type))+
geom_line(size=1.7,alpha=0.7)+
geom_point(size=6,alpha=0.75)+
```

```
xlab("Group number")+
ylab("Estimation of error rate")+
labs(colour="Approach:")+
scale_x_continuous(breaks=c(1:8))
ggsave("experim_graph.pdf", width=14, height=10)
```