Аннотация

Анализ электроэнцефалографии (ЭЭГ) является важным инструментом в нейробиологии, нейронной инженерии (например, интерфейсы мозг-компьютер (ИМК)), а также имеет промышленное применение. С развитием этих областей появляется всё большая потребность в анализе ЭЭГ данных. Получение информации об электрической активности зон головного мозга и о том, каким образом эта активность влияет на классификацию типа решаемой человеком задачи является важным шагом на пути к тому, чтобы извлекать из ЭЭГ больше информации о функционировании мозга. Для достижения данной цели на имеющихся ЭЭГ данных мы рассмотрели каждый электрод в отдельности, чтобы ответить на следующие вопросы: (1) Как амплитуда биоэлектрической активности, регистрируемой каждым электродом с соответствующей ему зоны мозга, в отдельности влияет на результат классификации типа решаемой задачи? (2) Какие из них наибольшим образом влияют на результат?

Используя методы машинного обучения и проанализировав полученные результаты, было выдвинуто предположение о нецелесообразности использования в дальнейшем полученных результатов, рассматривая электроды таким образом.

Оценив значения p-value натренированных классификаторов, было рассмотрено два подхода, результаты которых согласуются друг с другом, и установлено, что нету такого электрода, по которому бы получилось сделать предсказание с точностью свыше 60%. Поэтому была исключена возможность определения активности человека по сигналу от одного электрода.

Содержание

1 Введение

Электроэнцефалография (ЭЭГ) — метод исследования головного мозга, основанный на регистрации его электрических потенциалов. ЭЭГ измеряет колебания напряжения в результате ионного тока в нейронах головного мозга. Клинически электроэнцефалограмма является графическим изображением спонтанной электрической активности мозга в течение определенного периода времени, записанной с нескольких электродов мозга или поверхности скальпа [1], (т.е. каждый электрод соответствует определённой области мозга).

ЭЭГ широко используется в исследованиях, связанных с нейронной инженерией, неврологией и биомедицинской инженерией (например, интерфейсы мозг-компьютер [2], анализ сна [3], обнаружение приступов эпилепсии [4]) из-за относительно низкой финансовой стоимости. Классификация этих сигналов является важным шагом на пути к тому, чтобы сделать использование ЭЭГ более практичным в применении и менее зависимым от подготовленных специалистов. Типичный процесс подготовки классификации ЭЭГ включает в себя удаление глазодвигательных и мышечных артефактов, отбор признаков и классификацию.

На самом базовом уровне набор данных ЭЭГ состоит из объектов — векторов действительных значений, которые которые представляют генерируемые мозгом потенциалы, снятые с кожи головы. Размерность каждого такого вектора, характеризуется числом электродов и количеством диапазонов спектра каждого электрода.

На данный момент существует большое количество информации о применении традиционных алгоритмов машинного обучения для распознавания типа решаемой человеком задачи (например, [5], [6]), в которых для предсказания учитываются данные всех электродов в совокупности. В то же время, исследований, в которых электроды рассматриваются по отдельности, не достаточно для того, чтобы располагать полной информацией об электрической активности мозга.

В данной работе электроды были рассмотрены отдельно друг от друга, чтобы ответить на следующие вопросы: (1) Как амплитуда биоэлектрической активности, регистрируемой каждым электродом с соответствующей ему зоны мозга, в отдельности влияет на результат классификации типа решаемой задачи? (2) Какие из них наибольшим образом влияют на результат? (3) Можно ли получить результат, точность которого свыше 60%, рассматривая таким образом электроды?

Для получения ответов на вышеперечисленные вопросы были использованы стратегии для классификации ЭЭГ с использованием линейных методов машинного обучения, а также методы предварительной обработки данных ЭЭГ. Полученная информация может послужить отправной точкой для начального этапа проектирования архитектуры в будущих приложениях машинного обучения для классификации ЭЭГ.

2 Использованные данные

2.1 Теоретическое введение

Тест Стернберга — классический эксперимент, проведенный в 1966 году психологом Солом Стернбергом, позволивший сделать вывод о том, что информация извлекается из кратковременной памяти путём последовательного исчерпывающего сканирования. Оригинальные и модифицированные схемы теста (Sternberg item recognition paradigm, SIRP), описанного в статье [7], используются для изучения особенностей кратковременной и рабочей памяти.

Оригинальный тест состоял из 24 тренировочных и 144 тестовых проб. В каждой пробе участнику эксперимента предъявлялся произвольный набор цифр, который требовалось запомнить. Длина набора варьировалась от 1 до 6 цифр, каждая из которых предъявлялась отдельно в течение 1-2 секунды. После этого следовала пауза длиной в 2 секунды, а за ней контрольная цифра. Испытуемые должны были потянуть один из двух рычагов в качестве ответа «Да, это одна из запомненных цифр» или «Нет, это новая цифра» (требовавшиеся с одинаковой вероятностью), после чего контрольный стимул исчезал, а загоравшаяся лампочка давала обратную связь о правильности ответа. В конце испытуемых также просили произнести запомненную последовательность (схема экспериментальной установки изображена на рис. 1).

Подробный сценарий каждого теста:

1. предупреждающий сигнал;
2. запоминаемый стимул-образец;
3. отсрочка;
4. предупреждающий сигнал;
5. контрольный стимул;
6. ответ испытуемого;
7. обратная связь о правильности выполнения задания;
8. просьба к испытуемому вспомнить стимул-образец, после чего начинается следующая проба.

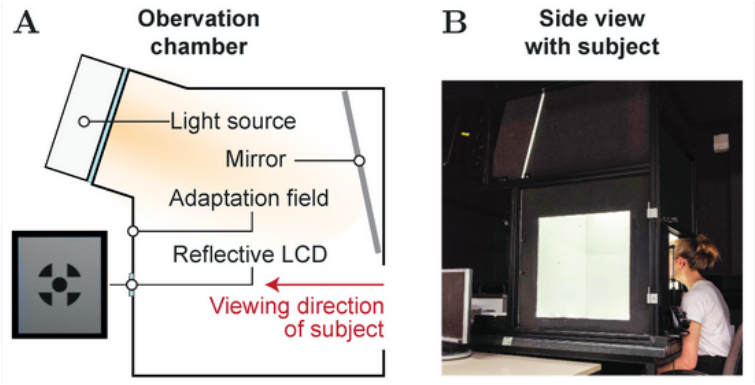

Рисунок 1 — Экспериментальная установка. (A) Схематическое изображение камеры наблюдения и направления обзора. (B) Изображение вида сбоку камеры наблюдения [8]

Парадигма Стернберга была успешно использована в рамках изучения индивидуальных различий в процессах памяти у здоровых участников [9], [10], в исследованиях посвященных изучению дефицитарности и изменений кратковременной памяти при старении [11], в исследованиях шизофрении и болезни Альцгеймера [12], депрессии [13], множественного склероза [14], в исследованиях изучающих воздействия различных медикаментов на процессы памяти [15].

2.2 Процедура снятия ЭЭГ данных

В настоящем исследования для получения ЭЭГ данных использовался "Тест Стернберга". В эксперименте на голову участника исследования одевается специальная шапочка со специальными металлическими электродами, которые регистрирует биоэлектрическую активность мозга (см. рис. 2). Каждый электрод с какой-то частотой (обычно 500-1000 Гц) регистрирует амплитуду колебаний электромагнитной активности изменения электрического потенциала с поверхности головы.

Затем участнику эксперимента последовательно предъявляются наборы цифр. В каждом наборе цифры от 1 до 9 (без повторений одной цифры дважды) были представлены в случайной последовательности. При этом размеры наборов могут отличаться (в оригинальных исследованиях Стернберга, как было описано выше, это наборы объемом от 1 до 6 цифр). В зависимости от того, какой длины последовательность была предъявлена участнику, определялся тип решаемой задачи: лёгкая, средняя, повышенной сложности и тяжёлая (3, 4, 5 или 6 и более цифр для запоминания соответственно) — всего 4 типа задачи. Цифры показываются одна за другой, участнику необходимо запомнить их последовательность. После контрольного сигнала (появление определенной цифры на экране) участнику необходимо как можно быстрее ответить, присутствовала ли цифра в предъявленном до этого наборе. После этого участника просят вспомнить порядок представления цифр, для того чтобы убедиться, что он действительно запомнил последовательность.

Рисунок 2 — Расположение электродов на поверхности головы

Перед предъявлением набора стимулов участникам эксперимента предъявляется фиксационный крест на 1 секунду. Предъявление набора стимулов начинается через 1.2 секунды после предъявления фиксационного креста и длится 1.5 секунды. Через 2 секунды после окончания предъявления набора предъявляется тестовый стимул (задача участника - сказать был ли тестовый стимул в наборе). Тестовый стимул предъявляется на 2 секунды. С момента начала предъявления тестового стимула участник может давать ответ.

В исследовании принимал участие 101 человек. Каждому человеку предлагалось для решения порядка 30 задач на каждый из 4 типов.

3 Формат данных

3.1 Проблемы при регистрации ЭЭГ

После того, как ЭЭГ снято, необходимо отфильтровать сигнал. Сигнал плохо виден на фоне шума, который создают различные артефакты.

При регистрации ЭЭГ артефактом является любая активность, не связанная с электрической активностью мозга. Все артефакты при регистрации ЭЭГ могут быть разделены на две группы:

1. артефакты, связанные с аппаратурой, внешние помехи физической природы;

2. физиологические артефакты, регистрируемые от больного.

Наиболее частыми являются артефакты, связанные с регистрацией потенциалов, возникающих при моргании и движении глаз, миографических потенциалов при мышечном напряжении.

В связи с этим возникает проблема обработки записи ЭЭГ — необходимо удалить артефакты, чтобы получить более точные показатели.

3.2 Предварительная обработка данных

Для удаления артефактов был использованный автоматический метод, реализованный для среды обработки ЭЭГ (MNE) [16]. После предварительной очистки данные ЭЭГ были отфильтрованы в диапазоне частот 1-30 Гц. Далее с помощью метода Уэлча для стандартных узких частотных диапазонов ЭЭГ

- δ : 1-4 Гц
- θ : 4-8 Гц
- α_1 : 8-10 Гц
- α_2 : 10-13 Гц
- β_1 : 13-20 Гц
- β_2 : 20-30 Гц

спектральная мощность сигнала была проанализирована отдельно для тестов Стернберга разной сложности (3, 4, 5 или 6 цифр и более для запоминания).

После удаления артефактов данные были получены в виде набора .csv файлов. Каждый файл содержал в себе данные, относящиеся к конкретному человеку и типу задачи (см. главу 2.2),то есть всего 404 файла. Данные из файлов были объединены в один набор данных, пригодных для обработки алгоритмами машинного обучения.

3.3 Вид полученных данных

Полученный набор данных устроен следующим образом (см. рис. 3)

	1_1	1_2	1_3	1_4	...	63_5	63_6	name	task
0	2.148821e-12	1.762882e-12	1.459278e-13	1.587917e-13	...	6.211839e-13	8.285990e-13	chcon_s_100	0
1	1.270260e-12	8.164477e-13	5.903729e-13	3.056763e-13	...	1.079865e-12	2.844522e-12	chcon_s_100	0
2	7.709517e-13	1.717117e-13	3.039963e-13	1.663584e-13	...	1.407080e-12	4.519942e-12	chcon_s_100	0
3	2.800220e-12	3.611335e-13	9.828645e-13	3.723583e-13	...	6.247119e-13	1.532856e-12	chcon_s_100	0
4	9.507999e-13	4.833313e-13	1.694255e-12	1.434799e-13	...	1.444978e-12	2.585149e-12	chcon_s_100	0
...	...	...	...	...	...	...	...	...	...
10638	2.783871e-12	4.492069e-13	6.107854e-13	2.169733e-13	...	5.649186e-14	1.532088e-13	mcon_str_41	3
10639	1.394714e-12	2.106615e-12	2.131392e-12	1.251115e-12	...	1.013557e-12	8.060742e-13	mcon_str_41	3
10640	3.272053e-12	2.410372e-12	3.519692e-12	3.088923e-13	...	3.932170e-13	6.883666e-13	mcon_str_41	3
10641	1.226257e-11	3.028533e-12	6.798046e-13	3.032247e-13	...	5.681097e-13	6.363103e-12	mcon_str_41	3
10642	8.328733e-11	2.666764e-12	6.353130e-12	8.844715e-13	...	9.100197e-13	1.183650e-12	mcon_str_41	3

10643 rows × 380 columns

Рисунок 3 — Полученный набор данных ЭЭГ

Каждая строка нашей выборки имеет размерность $63 \cdot 6 + 1 + 1 = 380$. В строке первые $63 \cdot 6 = 378$ ячеек есть амплитуды колебаний электромагнитной активности, которые регистрировались каждым из 63 электродов (каждый электрод в свою очередь характеризуется шестью частотными диапазонами спектра). Также есть столбец "$name$", содержащий идентификатор участника эксперимента, и столбец "$task$" с соответствующим типом задачи, которую решал участник во время снятия ЭЭГ данных. Каждому из 4 типов задачи (лёгкая, средняя, повышенной сложности и трудная) соответствует число от 0 до 3.

Всего 101 участником эксперимента было решено от 22 до 32 задач каждого из 4 типов сложности. На каждого человека приходится всего порядка 80–120 задач. Суммарное количество объектов в выборке составляет 10643.

Рассмотрим объект выборки более подробно (см. рис. 4):

- Красным цветом выделены амплитуды колебаний электромагнитной активности электродов (каждый из 63 характеризуется шестью частотными диапазонами). В каждой ячейке значением является вещественное число;
- Зелёным цветом — идентификатор участника;
- Синим цветом — тип задачи (число от 0 до 3).

	1_1	1_2	1_3	1_4	...	63_5	63_6	name	task
0	2.148821e-12	1.762882e-12	1.459278e-13	1.587917e-13	...	6.211839e-13	8.285990e-13	chcon_s_100	0

Рисунок 4 — Детальный вид объекта выборки

4 Постановка задачи

4.1 Задача обучения классификатора

Итак, у нас есть обучающая выборка, которой соответствует пара объект-ответ. Объект описывается 63 вещественными признаками — амплитудами колебаний электромагнитной активности (см. главу 3.3, выделено красным цветом), а ответы — это числа от 0 до 3, соответствующие типу решаемой задачи.

$$\textbf{Обучающая выборка: } X^l = (x_i, y_i)_{i=1}^l, \quad x_i \in \mathbb{R}^n, \quad y_i \in \{0, 1, 2, 3\}, \quad (1)$$

где $\quad l = 10643, \quad n = 63.$

Перед нами задача обучения с учителем, задача классификации.

4.2 Цели и задачи

В данной работе мы будем рассматривать электроды не все сразу как это уже исследовалось во многих работах (например, [5],[6],[3], [4]), а по отдельности.

Выдвигаем следующую гипотезу: рассматривая таким образом электроды, получим результат (качество предсказания классификатора), точность которого свыше 60% — хотим получить ответы на следующие вопросы: (1) как амплитуда биоэлектрической активности, регистрируемой электродом с конкретной зоны мозга, в отдельности влияет на результат классификации типа решаемой задачи? (2) Какие из них наибольшим образом влияют на результат?

Проанализировав полученные результаты сделаем выводы об истинности выдвинутой гипотезы, а, следовательно, и достоверности полученных результатов.

5 Ход работы

Для того, чтобы рассмотреть каждый электрод по отдельности, необходимо разделить наши исходные данные перед тем, как применять алгоритмы машинного обучения. Процесс разделения исходных данных будет описан далее.

5.1 Разделение данных

Разобьём наш исходный набор данных на "подвыборки в каждой из которых будет информация, относящаяся только к одному участнику эксперимента. Всего принимал участие в эксперименте 101 человек, следовательно, и подвыборок будет 101 (по одной на каждого человека). На рис. 5 показан пример одной из подвыборок.

	1_1	1_2	1_3	1_4	...	63_5	63_6	name	task
0	1.492692e-12	1.238635e-12	1.910639e-13	1.391849e-13	...	9.200580e-13	4.572670e-13	chcon_s_102	0
1	7.049684e-13	1.342876e-12	2.752966e-13	4.697597e-13	...	2.386464e-13	1.629742e-13	chcon_s_102	0
2	6.013291e-13	1.099664e-12	3.951522e-13	2.993900e-13	...	5.496934e-13	3.740934e-13	chcon_s_102	0
3	1.983854e-13	4.378199e-13	5.245689e-13	1.822007e-13	...	1.099374e-12	2.168006e-13	chcon_s_102	0
4	1.369758e-12	4.479227e-13	7.826958e-13	1.221268e-12	...	9.774529e-13	1.985916e-13	chcon_s_102	0
...	...	...	...	...	...	...	...	...	...
110	3.250180e-12	6.762252e-13	8.742837e-13	2.789760e-13	...	9.660850e-13	8.965652e-13	chcon_s_102	3
111	3.005538e-12	7.622600e-13	3.951518e-13	5.211876e-13	...	3.029778e-13	1.062439e-12	chcon_s_102	3
112	2.683009e-12	2.795308e-12	2.816187e-13	7.334988e-13	...	7.614454e-13	1.369562e-12	chcon_s_102	3
113	7.558109e-12	3.800672e-13	7.696249e-13	7.126857e-13	...	4.882033e-13	7.098585e-14	chcon_s_102	3
114	6.721495e-12	2.818428e-12	5.192166e-13	4.371117e-13	...	6.159956e-13	2.257836e-13	chcon_s_102	3

115 rows × 380 columns

Рисунок 5 — Пример набора данных для одного человека

Затем необходимо разделить полученные выборки так, чтобы в них содержалась информация только об одном электроде. Каждый электрод характеризуется шестью частотными диапазонами ЭЭГ (см. главу 3.1). То есть в получившемся наборе данных (см. рис. 6) для одного электрода будет шесть колонок (не считая колонки "task"целевых переменных с типом сложности задачи).

	1_1	1_2	1_3	1_4	1_5	1_6	task	name
0	1.492692e-12	1.238635e-12	1.910639e-13	1.391849e-13	2.909224e-13	1.895839e-13	0	chcon_s_102
1	7.049684e-13	1.342876e-12	2.752966e-13	4.697597e-13	1.565714e-13	2.686569e-13	0	chcon_s_102
2	6.013291e-13	1.099664e-12	3.951522e-13	2.993900e-13	3.229311e-13	4.628919e-13	0	chcon_s_102
3	1.983854e-13	4.378199e-13	5.245689e-13	1.822007e-13	3.682875e-12	6.127147e-13	0	chcon_s_102
4	1.369758e-12	4.479227e-13	7.826958e-13	1.221268e-12	5.726069e-13	3.406408e-13	0	chcon_s_102
...	...	...	...	...	...	...	...	...
110	3.250180e-12	6.762252e-13	8.742837e-13	2.789760e-13	8.287411e-13	1.660736e-13	3	chcon_s_102
111	3.005538e-12	7.622600e-13	3.951518e-13	5.211876e-13	1.039477e-12	1.069287e-12	3	chcon_s_102
112	2.683009e-12	2.795308e-12	2.816187e-13	7.334988e-13	6.672073e-13	1.255715e-13	3	chcon_s_102
113	7.558109e-12	3.800672e-13	7.696249e-13	7.126857e-13	4.592769e-13	3.354001e-13	3	chcon_s_102
114	6.721495e-12	2.818428e-12	5.192166e-13	4.371117e-13	1.037650e-13	4.216510e-13	3	chcon_s_102

115 rows × 8 columns

Рисунок 6 — Пример получившегося набора данных для первого электрода (колонка "$name$" добавлена для наглядности).

Итого 63 получившихся набора данных для каждого человека. Всего таких наборов $101 \cdot 63 = 6363$. Количество строк в таком наборе от человека к человеку варьируется от 80 до 120 (см. главу 3.3).

Опишем формально получившийся набор данных. Получится выражение, аналогичное выражению (1), но изменится размерность n и l:

$$\textbf{Обучающая выборка: } X^l = (x_i, y_i)_{i=1}^l, \quad x_i \in \mathbb{R}^n, \quad y_i \in \{0, 1, 2, 3\}, \quad (2)$$

где $\quad l \in [80, 120], \quad n = 6.$

Итак, у нас есть обучающие выборки (по 63 на каждого участника эксперимента), каждой из которых соответствует пара объект-ответ. Объект описывается шестью характеризующими электрод вещественными признаками — амплитудами колебаний электромагнитной активности, а ответы — это числа от 0 до 3, соответствующие типу решаемой задачи.

Учитывая небольшой размер данных (80-120 строк), с которыми нужно будет работать, возьмём простой и широко используемый в задачах классификации метод — логистическую регрессию.

5.2 Логистическая регрессия

Логистическая регрессия применяется для предсказания вероятности возникновения некоторого события по значениям множества признаков [17]. Для этого вводится целевая переменная y.

5.2.1 Описание

1. *Случай бинарной классификации.*

В этом случае зависимая переменная y, принимающая одно из двух значений— это числа 0 и 1, и множество независимых переменных (также называемых признаками, предикторами или регрессорами) — вещественных x_1,x_2, $\cdots$,x_n, на основе значений которых требуется вычислить вероятность принятия того или иного значения целевой переменной. Как и в случае линейной регрессии, для простоты записи вводится фиктивный признак $x_0 = 1$.
Делается предположение о том, что вероятность наступления события $y = 1$ равна:

$$\mathbb{P}\{y = 1 \mid x\} = f(z),$$

где $z = \theta^T x = \theta_0 + \theta_1 x_1 + \ldots + \theta_n x_n$, x и θ — векторы-столбцы значений независимых переменных 1, $x_1, \ldots, x_n$ и параметров (коэффициентов регрессии) — вещественных чисел θ_0,...,θ_n, соответственно, а $f(z)$ — так называемая логистическая функция (иногда также называемая сигмоидой или логит-функцией): $f(z) = \frac{1}{1+e^{-z}}$.

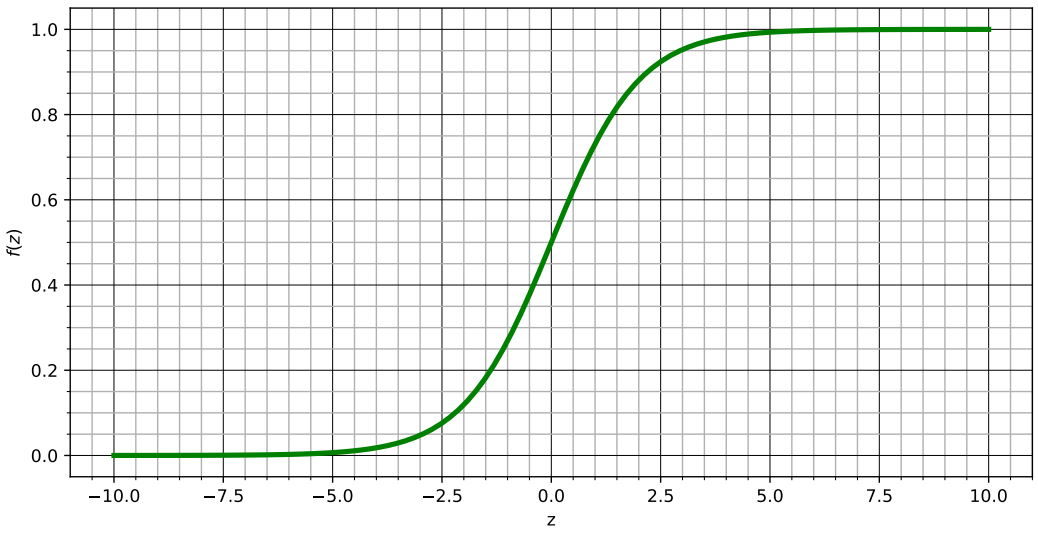

Рисунок 7 — Логистическая функция (сигмоида): $f(x) = \frac{1}{1+e^{-x}}$.

Так как y принимает только значения 0 и 1, то вероятность принять значение 0 равна:

$$\mathbb{P}\{y = 0 \mid x\} = 1 - f(z) = 1 - f(\theta^T x).$$

Перепишем функцию распределения y при заданном x в следующем виде:

$$\mathbb{P}\{y \mid x\} = f(\theta^T x)^y (1 - f(\theta^T x))^{1-y}, \quad y \in \{0,1\}.$$

Это есть распределение Бернулли с параметром, равным $f(\theta^T x)$.

2. *Случай мультиклассовой классификации (наш случай).*

В отличие от бинарной классификации, где результат принимал значения 0 или 1, т.е. $\in \{0, 1\}$, в данном случае результатом будет один из множества классов (которых теперь больше, чем два).
В нашем случае $n = |y| = 4, y \in \{0, 1, 2, 3\}$.
В данном случае задача мультиклассовой классификации делится на 4 задачи бинарной классификации. В каждой такой задаче бинарной классификации предсказывается вероятность того, что y относится к заданному классу.

$$\begin{aligned} y &\in \{0, 1, 2, 3\} \\ f^{(0)}(\theta^T x) &= \mathbb{P}\{y = 0 \mid x; \theta\} \\ f^{(1)}(\theta^T x) &= \mathbb{P}\{y = 1 \mid x; \theta\} \\ f^{(2)}(\theta^T x) &= \mathbb{P}\{y = 2 \mid x; \theta\} \\ f^{(3)}(\theta^T x) &= \mathbb{P}\{y = 3 \mid x; \theta\} \\ Prediction &= \max_i (f^{(i)}(\theta^T x)) \end{aligned}$$

В случае мультиклассовой классификации выбирается рассматриваемый класс, все остальные классы при этом «объединяются» в другой класс, отличный от выбранного. Это проделывается для каждого класса с применением для каждого случая бинарной логистической регрессии [17].

5.2.2 Подбор параметров

Чтобы подобрать параметров θ_0,...,θ_n необходимо составить обучающую выборку (2), состоящую из наборов значений признаков и соответствующих им значений целевой переменной y. Формально, это множество пар $(x^{(1)}, y^{(1)})$, ...,$(x^{(l)}, y^{(l)})$, где $x^{(i)} \in \mathbb{R}^n$ — вектор значений признаков, а $y^{(i)} \in$

$\{0,1,2,3\}$ — соответствующее им значение y. Каждая такая пара называется обучающей выборкой. Обычно используется метод максимального правдоподобия, согласно которому выбираются параметры θ, максимизирующие значение функции правдоподобия на обучающей выборке:

$$\hat{\theta} = arg\max_{\theta} L(\theta) = arg\max_{\theta} \prod_{i=1}^{m} \mathbb{P}\{y = y^{(i)} \mid x = x^{(i)}\}.$$

Максимизация функции правдоподобия эквивалентна максимизации её логарифма [17]:

$$\ln L(\theta) = \sum_{i=1}^{m} \log \mathbb{P}\{y = y^{(i)} \mid x = x^{(i)}\} = \sum_{i=1}^{m} y^{(i)} \ln f(\theta^T x^{(i)}) +$$
$$+(1 - y^{(i)}) \ln(1 - f(\theta^T x^{(i)})), \quad \text{где } \theta^T x^{(i)} = \theta_0 + \theta_1 x_1^{(i)} + \cdots + \theta_n x_n^{(i)}.$$

Как один из вариантов для максимизации этой функции может быть применён метод градиентного спуска. Он заключается в проведении следующих итераций, начиная с некоторого начального значения параметров θ:

$$\theta := \theta + \alpha \nabla \ln L(\theta) = \theta + \alpha \sum_{i=1}^{m} (y^{(i)} - f(\theta^T x^{(i)})) x^{(i)}, \alpha > 0.$$

5.2.3 Преимущества

- Простота внедрения;
- Малый объём вычислений при классификации, высокая скорость работы, а затраты по памяти малы;
- Можно легко обновить модель для поглощения новых данных;
- Удобно оценивать вероятности возникновения некоторого события по значениям множества признаков (пригодится нам далее при вычислении p-value);

5.2.4 Использованная реализация

Чтобы применить логистическую регрессию на наших наборах данных (рис. 6), воспользуемся готовой реализацией алгоритма, которую предоставляет **scikit-learn** [18].

Scikit-learn — это Python-модуль для машинного обучения, построенный поверх *SciPy* [19] и распространяемый по лицензии *BSD* [20]. В модуле Scikit-learn реализован класс $sklearn.linear\_model.LogisticRegression$, который и предоставляет весь необходимый функционал.

5.3 Нормировка признаков

Так как значения признаков имеют малые порядки (от 10^{-14} до 10^{-10}) и сильно различаются между собой по величине, то, в виду чувствительности алгоритма к диапазону изменений входных переменных и чтобы избежать ухудшения результатов обучения, перед использованием логистической регрессии необходимо провести *нормировку признаков.*

В машинном обучении нормировкой признаков называют метод предобработки числовых признаков в обучающей выборке, чтобы привести их к определённой общей шкале (в нашем случае к отрезку [0, 1]) без потери информации о различии диапазонов [17].

Необходимость нормализации вызвана тем, что разные признаки обучающей выборки могут быть разных масштабов и изменяться в разных диапазонах.

В этом случае возникает нарушение баланса между влиянием входных переменных, представленных в разных масштабах, на выходную переменную. Т.е. это влияние обусловлено не реальной зависимостью, а изменением масштаба. В результате, обучаемая модель выявит некорректные зависимости.

В модуле содержится несколько реализаций нормировок признаков, иллюстрации которых приведены ниже:

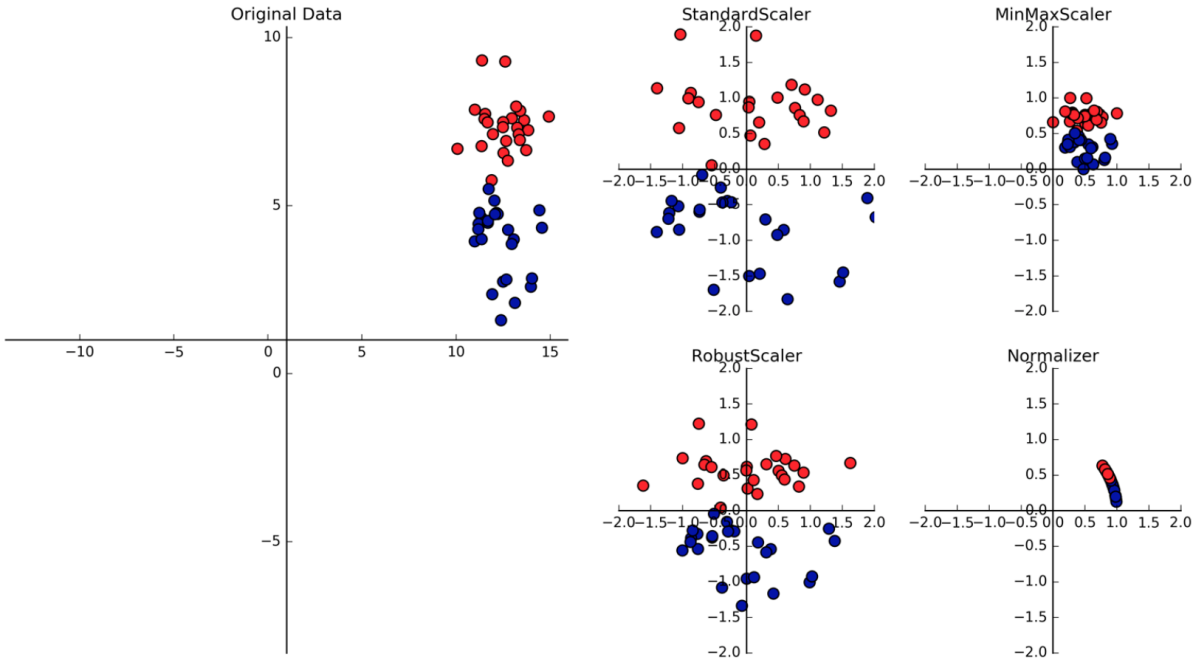

Рисунок 8 — Иллюстрация нормировки признаков в зависимости от различных реализаций (MinMaxScaler, StandardScaler, RobustScaler, Normalizer), содержащихся в модуле scikit-learn [18]

5.3.1 Минимаксная реализация

Существует несколько реализаций методов нормировки признаков в модуле scikit-learn. В данной работе была использована минимаксная реализация [18].

```
class sklearn.preprocessing.MinMaxScaler(feature_range=(0, 1), *, copy=True, clip=False)
```

Рисунок 9 — Минимаксная реализация модуля scikit-learn

Данная минимаксная нормировка реализуется по формуле:

$$\hat{x} = \frac{x - \min(x)}{\max(x) - \min(x)},$$

где x является исходным значение признака, а $\hat{x}$ – преобразованное к заданному диапазону (к отрезку $[0, 1]$) значение.

5.3.2 Применение

Итак, воспользуемся готовой реализацией и проведём нормировку признаков.

	1_1	1_2	1_3	1_4	1_5	1_6	task	name
0	0.012118	0.238420	0.105437	0.043330	0.060459	0.037924	0	chcon_s_102
1	0.004845	0.260210	0.169407	0.190699	0.023245	0.056127	0	chcon_s_102
2	0.003889	0.209371	0.260431	0.114749	0.069325	0.100841	0	chcon_s_102
3	0.000169	0.071025	0.358716	0.062507	1.000000	0.135331	0	chcon_s_102
4	0.010983	0.073137	0.554749	0.525719	0.138483	0.072698	0	chcon_s_102
...	...	...	...	...	...	...	...	...
110	0.028343	0.120859	0.624305	0.105649	0.209430	0.032512	3	chcon_s_102
111	0.026084	0.138843	0.260431	0.213626	0.267802	0.240435	3	chcon_s_102
112	0.023107	0.563813	0.174209	0.308273	0.164686	0.023188	3	chcon_s_102
113	0.068113	0.058953	0.544822	0.298995	0.107091	0.071492	3	chcon_s_102
114	0.060390	0.568646	0.354651	0.176145	0.008618	0.091347	3	chcon_s_102

115 rows × 8 columns

Рисунок 10 — Пример одного из наборов данных после нормировки признаков. (можно сравнить с рисунком 6)

5.4 Применение логистической регрессии

На обучающихся выборках типа (2) используем логистическую регрессию, чтобы узнать как амплитуда биоэлектрической активности, регистрируемой электродом с конкретной зоны мозга, в отдельности влияет на результат классификации типа решаемой задачи и выявить какие из них больше всего влияют на результат.

5.5 Метрика оценки качества

Для оценки качества построенных линейных моделей используем метрику *accuracy score*.

Перед тем, как перейти к описанию самой метрике, введём матрицу ошибок (confusion matrix) для описания метрики в терминах ошибок классификации. Пусть у нас есть два класса и алгоритм, предсказывающий принадлежность каждого объекта к одному из классов, тогда матрица ошибок классификации будет выглядеть следующим образом:

	$y = 1$	$y = 0$
$\hat{y} = 1$	True Positive (TP)	False Positive (FP)
$\hat{y} = 0$	False Negative (FN)	True Negative (TN)

Рисунок 11 — Матрица ошибок классификации

Здесь $\hat{y}$ — это ответ алгоритма на рассматриваемом объекте, а y — истинная метка класса на этом объекте. Ошибки классификации бывают двух видов: False Negative (FN) или ошибка 1-го рода и False Positive (FP) или ошибка 2-го рода [17].
Accuracy score — метрика, показывающая долю верных ответов алгоритма:

$$accuracy\ score = \frac{TN + TP}{FP + FN + TP + TN}$$

6 Обработка результатов

6.1 Оценки качества предсказания

Допустим, что можно получить качество предсказания классификатора свыше 60%, рассматривая электроды по отдельности.

Проанализируем полученные данные.
В результате применения линейного метода машинного обучения (логистической регрессии) на обучающих выборках типа (2) получаем следующие результаты, которые изобразим в виде столбчатой диаграммы:

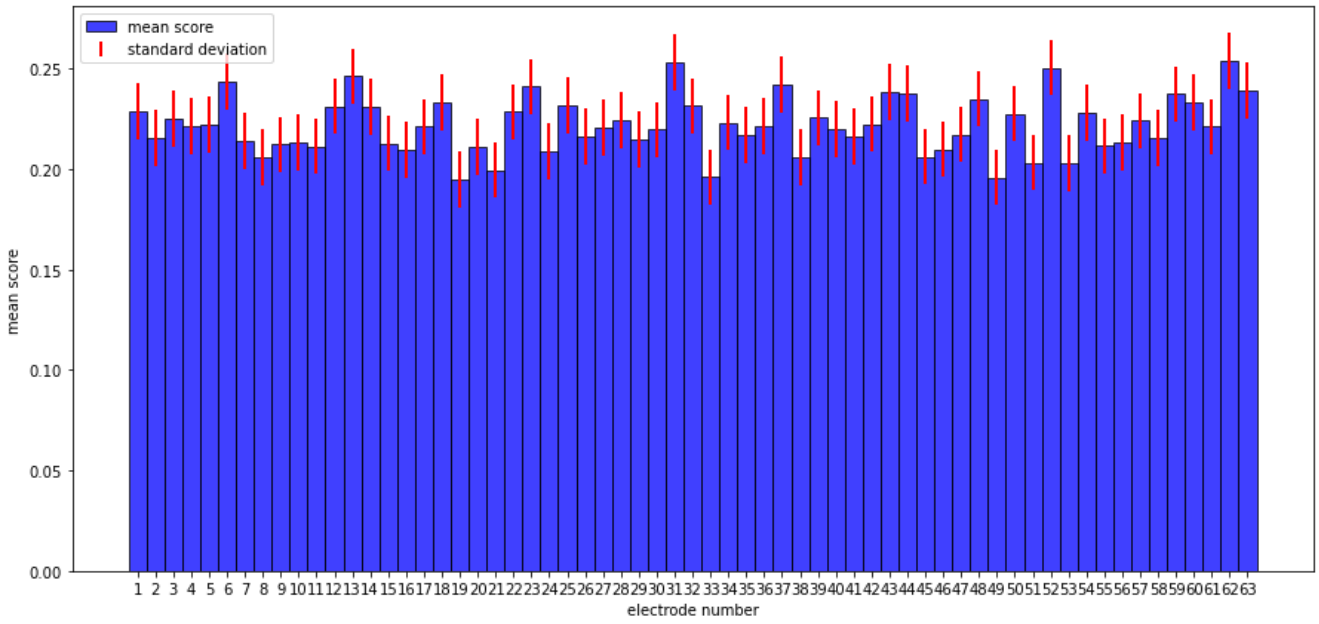

Рисунок 12 — Оценки качества предсказания натренированных классификаторов. По оси абсцисс — номера электродов, по оси ординат — средние по всем участникам эксперимента оценки качества предсказания.

Опираясь на изображённую столбчатую диаграмму, можем сформировать представление о том, как амплитуда биоэлектрической активности, регистрируемой электродом с конкретной зоны мозга, в отдельности влияет на результат классификации типа решаемой задачи и какие из них наибольшим образом влияют на результат.

Однако посмотрим на диапазон средних значений оценок качества предсказания натренированных классификаторов — варьируется от ~ 0.2 до ~ 0.25. Что интерпретируется следующим образом: доля правильно предсказанных ответов алгоритмом составляет 20–25%. Данное утверждение заставляет задуматься о целесообразности использования в дальнейшем полученных результатов.

Покажем формально, что рассматривать электроды таким образом (то есть по отдельности) нецелесообразно.

6.2 P-value

P-value — величина, используемая при тестировании статистических гипотез. Фактически это вероятность ошибки при отклонении нулевой гипотезы (ошибки первого рода) [21].

Пусть X — множество объектов выборки, H_0 — некоторая нулевая гипотеза, а $T(X)$ — статистика, используемая при проверке гипотезы H_0. Предполагаем, что если H_0 истинна, то распределение статистики $T(X)$ известно.

Обозначив функцию распределения $F(t) = P(T < t)$, p-value определяется как: $P(t) = 2\min(P_0,P)$.

6.2.1 Применение к логистической регрессии

Когда мы используем логистическую регрессию, мы используем некоторые независимые переменные для прогнозирования зависимой переменной (см. главу 5.2). Таким образом, применяя логистическую регрессию, мы получаем коэффициенты для каждой независимой переменной, которые мы использовали для прогнозирования зависимой переменной.

Рассматривая нашу нулевую гипотезу (см. главу 6.3) мы предполагаем, что нет корреляции между признаками и целевыми переменными. В логистической регрессии мы предполагаем, что используемая независимая переменная (признак) является статистически незначимой (т.е. нет корреляции между признаками и целевыми переменными) для прогнозирования зависимой переменной или, проще говоря, её коэффициент корреляции будет равен 0.

Поэтому чем меньше p-value, тем более статистически значимой является рассматриваемая переменная для нашей модели логистической регрессии.

Допустим, мы получили p-value равное 0.03 или 3%. Тогда это означает что наши результаты случайны лишь на 3% и на столько же не зависят от данного эксперимента. Поэтому, если получим p-value $< 5\%$, то придём к выводу, что переменная является значимой и отвергнем нашу нулевую гипотезу в пользу альтернативной гипотезы.

6.2.2 Реализация

Для вычисления p-value воспользовались готовой реализацией statsmodels. Statsmodels — это python-модуль, который предоставляет классы и функции для оценки множества различных статистических моделей, а также для проведения статистических тестов и исследования статистических данных. Для каждой реализации доступен обширный список статистики результатов. Результаты проверяются на соответствие существующим статистическим пакетам, чтобы убедиться в их правильности. Пакет выпущен под лицензией Modified BSD с открытым исходным кодом [20], [22].

```
import statsmodels.tools as sm
import statsmodels.api as sd

x = np.arange(10)[:, np.newaxis]
y = np.array([0,0,0,1,0,0,1,1,1,1])

sd_model = sd.Logit(y, sm.add_constant(x)).fit()
sd_model.summary()
```

Рисунок 13 — Пример использования готовой реализации для вычисления p-value на языке Python

Dep. Variable:	y	**No. Observations:**	10
Model:	Logit	**Df Residuals:**	8
Method:	MLE	**Df Model:**	1
Date:	Sun, 19 Jun 2022	**Pseudo R-squ.:**	0.4856
Time:	23:09:12	**Log-Likelihood:**	-3.5656
converged:	True	**LL-Null:**	-6.9315
Covariance Type:	nonrobust	**LLR p-value:**	0.009472

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.9587	2.506	-1.580	0.114	-8.870	0.952
x1	0.8797	0.515	1.707	0.088	-0.130	1.890

Рисунок 14 — Вывод работы программы

6.3 Формулировка гипотезы

Рассматривая электроды отдельно друг от друга, можно получить результат (качество предсказания классификатора), точность которого свыше 60%.

6.4 Доказательство

6.4.1 Подход №1

В нашем случае наша гипотеза о том, что с помощью одного электрода можно получить результат, точность которого больше 60% (см. главу6.3), будет нулевой гипотезой. Тогда альтернативной гипотезой будет противоположная ей, о том, что нельзя получить результат с точностью больше 60%.

Предположим наша нулевая гипотеза истинна. Тогда вычислим p-value для каждой полученной предсказательной модели. Получим, что максимальное значение по всем p-value равно 4.8%, минимальное значение — 0.7%, а среднее — 4.2% (см. таблицу 1):

Таблица 1 — Результаты вычисления p-value

p-value	значение	%
max	0.048	4.8
min	0.007	0.7
mean	0.042	4.2

Учитывая, что среднее значение p-value составляет 0.042 или 4.2%, а p-value каждой модели не превышает 0.048 или 4.8% (что меньше 5%), поэтому мы можем отклонить выдвинутую гипотезу (нулевую гипотезу) с уровнем значимости 95% в пользу альтернативной.

6.4.2 Подход №2

Используем другой подход. Пусть теперь нулевая гипотеза будет о том, что нельзя получить точность предсказания классификатора больше 60%. Вычислим для этого случая p-value.

Таблица 2 — Результаты вычисления p-value

p-value	значение	%
max	0.77	77
min	0.44	044
mean	0.56	56

Получив такое p-value нельзя отвергнуть нашу нулевую гипотезу. Данный результат согласуется с результатом, полученным в предыдущем подходе. Поэтому, рассмотрев несколько подходов, нельзя сказать, что есть хотя бы один электрод, по которому можно сделать предсказание с точностью свыше 60%.

7 Заключение

Получение информации об электрической активности зон головного мозга и о том, каким образом эта активность влияет на классификацию типа решаемой человеком задачи является важным шагом на пути к тому, чтобы извлекать из ЭЭГ больше информации о функционировании мозга.

В данной работе были рассмотрены данные для каждого электрода в отдельности, чтобы посмотреть на то, как активность каждой зоны мозга в отдельности влияет на предсказание типа решаемой задачи.

Проанализировав полученные результаты, мы рассмотрели два подхода, результаты которых согласуются друг с другом, и пришли к выводу, что нету такого электрода, по которому бы получилось сделать предсказание с точностью свыше 60%. Поэтому мы исключили возможность определить активность человека по сигналу от одного электрода.

Однако в связи с тем, что было протестировано достаточно большое количество электродов, то в действительности полученная двумя подходами оценка является консервативной и для более точных результатов можно было бы учесть «Look Elsewhere Effect» — увеличение вероятности обнаружить хотя бы один значимый сигнал при проверке большого числа независимых гипотез.

Полученные результаты говорят о нецелесообразности использования данных ЭЭГ при рассмотрении электродов отдельно друг от друга, по крайней мере применяя линейные методы машинного обучения, в частности логистическую регрессию.

Список литературы

[1] *H, Berger.* Ueber das Elektroenkephalogramm des Menschen / Berger H // *Arch. Psychiat. Nervenkr.* — 1929. — Pp. 527–570.

[2] *He Y Eguren D, Azorin J M.* Brain-machine interfaces for controlling lower-limb powered robotic systens / Azorin J M He Y, Eguren D // *J.Neural.* — 2018. — no. 15.

[3] *Motamedi-Fakhr S Moshrefi-Torbati M, Hill M Etc.* Signal processing techniques applied to human sleep EEG signals – a review / Hill M Etc. Motamedi-Fakhr S, Moshrefi-Torbati M // *Biomed. Signal Process. Control.* — 2014. — Pp. 21–33.

[4] *G, Chen.* Automatic EEG seizure detection using dual-tree complex wavelet-Fourier features / Chen G // *Expert Syst. Appl.* — 2014.

[5] *Hanjie Liu Jinren Zhang, Qingshan Liu Etc.* Minimum spanning tree based graph neural network for emotion classification using EEG / Qingshan Liu Etc. Hanjie Liu, Jinren Zhang // *Neural Networks.* — 2022. — Vol. 145. — Pp. 308–318.

[6] *Alexander Craik, Yongtian He Etc.* Deep learning for electroencephalogram (EEG) classification tasks: a review / Yongtian He Etc. Alexander Craik // *Journal of Neural Engineering.* — 2019. — Vol. 16, no. 031001.

[7] *Sternberg, Saul.* High-Speed Scanning in Human Memory / Saul Sternberg // *Science.* — 1966. — Vol. 153, no. 3736. — Pp. 652–654.

[8] *Julian Klabes Sebastian Babilon, Babak Zandi Etc.* The Sternberg Paradigm: Correcting Encoding Latencies in Visual and Auditory Test Designs / Babak Zandi Etc. Julian Klabes, Sebastian Babilon // *Vision.* — 2021. — Vol. 5(2), no. 5020021.

[9] *Etc., Jensen.* Oscillations in the alpha band 9-12 Hz increase with memory load during retention in a short-term memory task / Jensen Etc. // *Cereb Cortex.* — 2002. — no. 12(8).

[10] *Wolach.* The mode of short-term memory encoding as indicated by event-related potentials in a memory scanning task with distractions / Wolach, Pratt // *Clinical neurophysiology.* — 2001. — no. 112(1).

[11] *Etc., Bart Rypma.* Dissociating Age-related Changes in Cognitive Strategy and Neural Efficiency Using Event-related fMRI / Bart Rypma Etc. // *Cortex.* — 2005. — Vol. 41. — Pp. 582–594.

[12] *Corbin.* Is Sternberg's Memory Scanning Task Really a Short-Term Memory Task? / Corbin, Marquer // *Swiss Journal of Psychology.* — 2013. — Vol. 72(4), no. 181.

[13] *Slade, Blumhardt.* Working memory dysfunction in major depression: An event-related potential study / Blumhardt Slade, Sharma // *Clinical neurophysiology.* — 2000. — Vol. 111(9), no. 1531-43.

[14] *Catherine J Archibald 1 Xingchang Wei, James N Scott Etc.* Posterior fossa lesion volume and slowed information processing in multiple sclerosis / James N Scott Etc. Catherine J Archibald 1, Xingchang Wei // *Comparative Study.* — 2004. — Vol. 127(7), no. 1526-34.

[15] *Verster J C Volkerts E R, Verbaten M N.* Effects of alprazolam on driving ability, memory functioning and psychomotor performance: a randomized, placebo-controlled study. / Verbaten M N Verster J C, Volkerts E R // *Neuropsychopharmacology.* — 2002. — Vol. 27. — Pp. 260–269.

[16] *Mainak Jas Denis A. Engemann, Yousra Bekhti Etc.* Autoreject: Automated artifact rejection for MEG and EEG data / Yousra Bekhti Etc. Mainak Jas, Denis A. Engemann // *Neuroimage.* — 2017. — no. 159. — Pp. 417–429.

[17] *Etc., Andrew Ng.* Machine Learning Lectures / Andrew Ng Etc. — 2019.

[18] Документация scikit-learn. `https://scikit-learn.org`.

[19] Документация SciPy. `https://scipy.org`.

[20] The 3-Clause BSD License. `https://opensource.org/licenses/BSD-3-Clause`.

[21] *И., Кобзарь А.* Прикладная математическая статистика / Кобзарь А. И. // Прикладная математическая статистика. — Физматлит, 2006.

[22] Документация statsmodels.api. `https://www.statsmodels.org`.