Интерпретируемые решающие правила классификации на основе дерева секущих гиперплоскостей

Чернавин П. Ф., к.э.н., доцент,
Уральский федеральный университет, Екатеринбург, Россия

В статье анализируются различия между интерпретируемым и объяснимым искусственным интеллектом. Два этих понятия рассматриваются автором как различные, хотя и взаимодополняющие подходы к решению задач машинного обучения. На основании проведенного анализа автором предлагается представлять решающие правила для задач классификации в виде древовидной структуры, но в ее узлах использовать не отдельные входные признаки, а секущие гиперплоскости. Такой подход позволяет оценивать информативность каждого входного признака в отдельности и всей совокупности признаков в каждом узле и получать интерпретируемое решающее правило в целом для всей модели с высокими метриками качества классификации.

Введение

Знания и информацию часто структурируют на основе пирамиды Аккофа-Кэмпбелла (DIKW) [15]. В данной статье пирамида DIKW используется как концептуальная рамка, позволяющая структурировать требования к интерпретируемости моделей, а не как формальная методологическая основа построения интерпретируемых решающих правил. Данная пирамида в более удобном для дальнейших рассуждений виде представлена в Таблице 1.

Таблица 1. Пирамида Аккофа-Кэмпбелла в табличном виде

№	Уровень пирамиды	Определение	В чем суть	На какой вопрос отвечает
1	Сигналы	Материально воплощенные сообщения	Любые измеряемые импульсы	«Что происходит в среде?»
2	Данные	Собранные сигналы	Набор фактов или наблюдений без интерпретации	«Какие значения мы получили?»
3	Информация	Данные, обработанные для повышения вероятности правильного решения	Структурированные данные с выявленными связями, контекстом и значениями	«Что означает это?»
4	Знания	Механизмы использования информации – инструкции, алгоритмы, рецепты	Способы применения информации для достижения результата	«Как это использовать?»
5	Понимание	Понимание причино-следственных связей	Осознание «почему» и умение прогнозировать последствия	«Почему это так?»

Несмотря на значительные успехи в развитии искусственного интеллекта (ИИ) в последнее десятилетие, в настоящий момент достаточно хорошо освоены только четыре первых уровня пирамиды. То есть, существуют различные методы анализа данных, позволяющие на основе прецедентов создать решающие правила (РП), которые могут прогнозировать вероятность конечного результата при определенном наборе и значениях входных признаков. Такие РП даже в лучшем случае являются только знаниями, на основе которых можно создавать инструкции, рецепты, компьютерные системы, но без пояснения причин, почему следует делать именно так. То есть, многие РП не дают ответа на вопрос «почему?». В ряде случаев, этого вполне достаточно, но если речь идет о вынесении сурового судебного приговора, удалении жизненно важного органа или уничтожении самолета нарушителя воздушного пространства, то лицу, принимающему это ответственное решение в большинстве случаев, захочется иметь аргументацию, почему все же это следует делать. Поэтому в настоящий момент возникло новое направление - интерпретируемый ИИ и требования об обязательной интерпретируемости ИИ уже вносятся в законодательство ряда стран. В 2023 году представители 28 стран подписали «Блетчлиевскую декларацию», которая призывает к международному сотрудничеству и установлению глобальных стандартов для ИИ, уделяя особое внимание этическим аспектам и интерпретируемости [10]. Несмотря на необязательный характер, подобные декларации формируют практику, оказывающую влияние на требования регуляторов и надзорных органов.

Интерпретируемый и объяснимый ИИ

Отметим, что в данной статье акцент будет сделан на задачах классификации. Следует различать понятия интерпретируемость моделей и решающих правил ИИ и их объяснимость. Данным проблемам посвящено большое количество научных работ как у нас в стране [1-7], так и за рубежом[8-21] . Подробные обзоры различных подходов к этим вопросам приводятся в [1,3, 17]. Основные различия, по мнению автора статьи, между интерпретируемым и объяснимым ИИ приведены в таблице 2.

Таблица 2. Сравнение интерпретируемого и объяснимого ИИ

Критерий	Интерпретируемый ИИ	Объяснимый ИИ
Что это	Модель понятна сама по себе	Модель требует внешнего объяснения
Примеры моделей	Линейная регрессия, линейное разделение множеств, дерево решений, классификация	Глубокие нейросети, ансамбли (XGBoost, Random Forest)

	выпуклыми оболочками	
Средства понимания	Анализ коэффициентов, структуры модели, геометрическая интерпретация. Дополнительных программ для анализа не требуется	Внешние программы анализа локальной и глобальной интерпретируемости (LIME, SHAP, Grad-CAM, Anchors и др.). В [1] приведены 22 варианта программ
Прозрачность	Полная (white-box) на всех этапах	Частичная (black-box с объяснением)
Применение	Критические области (медицина, финансы, право)	Сложные модели, где важна точность, но нужно объяснение

Таким образом, *интерпретируемость* — это понятность самой модели. Объяснимость означает, что даже если модель сложная и "чёрный ящик" , мы можем построить внешние средства объяснения её решений. В настоящий момент, есть два взаимосвязанных, но самостоятельных направления: Интерпретируемый ИИ (Interpretable AI) и Объяснимый ИИ (Explainable AI). Однако, общепринятое научным сообществом разграничения этих понятий отсутствует. Есть точки зрения, что интерпретируемый ИИ является частью объяснимого [1,17], есть противоположные утверждения [18,21] и есть мнения, что четких различий между этими понятиями нет [8]. Отметим, что в ряде моментов мнения сторонников различных взглядов сходятся:

1. Качество решающих правил в обоих случаях оцениваются отдельными стандартными метриками AUC ROC, Accuracy, Precision и т.п. Данные метрики оценивают предсказательную способность модели, но не являются достаточным условием для ее использования в регулируемых процессах принятия решений
2. Объяснение или интерпретация должны быть понятны специалисту в конкретной области.

Если в первом случае, все понятно, то второй порождает большое количество вопросов так как чем можно измерить степень понимания человеком? Видимо только из общих соображений и с его слов.

С точки зрения автора интерпретируемость измеряется структурными метриками модели. Эти метрики описывают насколько легко понять саму модель. Поэтому их можно вычислить просто по структуре модели и если модель соответствует этим метрикам, то она интерпретируема и дает интерпретируемое решающее правило. Данные метрики приведены в таблице 3 и нацелены на анализ модели.

Таблица 3. Метрики интерпретируемости

№	Метрика	Интерпретация
1	Число информативных входных признаков	Чем меньше входных признаков, тем выше интерпретируемость
2	Размер модели	Чем меньше глубина дерева, листьев на дереве, членов ансамбля линейных разделителей и т.п., тем проще интерпретировать
3	Соблюдение ожидаемых зависимостей	Зависимость отклика от признака не противоречит здравому смыслу. Например, растет доходность, значит растет риск. Увеличивается количество лейкоцитов, увеличивается вероятность заболевания и т.п.
4	Стабильность объяснений	Чем меньше изменяются объяснения при малых изменениях, тем лучше
5	Понимание структуры входных данных в пространстве информативных признаков	Формализованное разбиение входных данных на отдельные области упрощает восприятие и последующее применение решающего правила. Чем меньше будет таких областей, тем лучше

Если модель, этим метрикам не соответствует, то она не интерпретируема, но может дать объяснимое решение. Можно сказать, что *объяснимость* — это создание интерпретаций для не интерпретируемых решений. Поэтому анализируются решения для отдельных примеров (локальная объяснимость) и, если на основе такого анализа, можно создать некоторое целостное объяснение, то получается глобальная объяснимость. Таким образом, объяснимость измеряется качеством и стабильностью объяснений результатов, выданных “черным ящиком”. Следует отметить, что предложенные метрики не претендуют на полноту, однако позволяют формализовать интерпретируемость как свойство структуры модели, а не как субъективное качество объяснения. Последний пункт Таблицы №3 требует отдельных пояснений.

Интерпретируемость, метрики качества решающего правила, структура данных

Обычно принято считать, что есть некоторое противоречие между интерпретируемостью решения и его точностью [3,8,13,17]. Достаточно часто утверждается, что сложно интерпретируемые методы, такие как нейронные сети или случайный лес имеют метрики качества решающего правила более высокие, чем, например, решения, построенные на основе линейных разделителей или ближайших соседей. На самом деле, все зависит от структуры данных. Например, если множества линейно разделимы (Рисунок 1), то использовать даже просто дерево решений в большинстве случаев бессмысленно, тем более случайный лес. Если же множества вообще нельзя структурировать (Рисунок 2), то любой из методов не даст интерпретируемого решения, даже если метрики качества классификации будут высокими. В последнем случае, наилучшие результаты дадут методы ближайших соседей или потенциальных функций, но их нельзя называть интерпретируемыми, так как они не дают общей картины (не соответствуют Таблице №3). Это объяснимые методы потому, что они дают только локальную объяснимость, на основе которой вряд ли можно достичь пятого уровня пирамиды Аккофа-Кэмпбелла.

Рисунок 1. Линейно разделимые множества

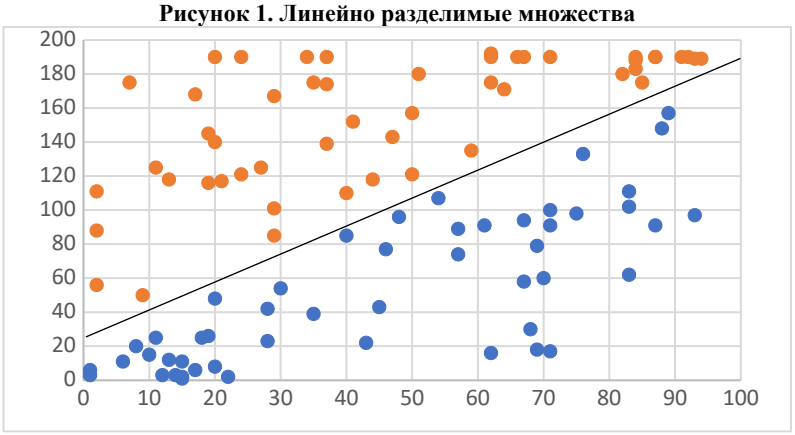
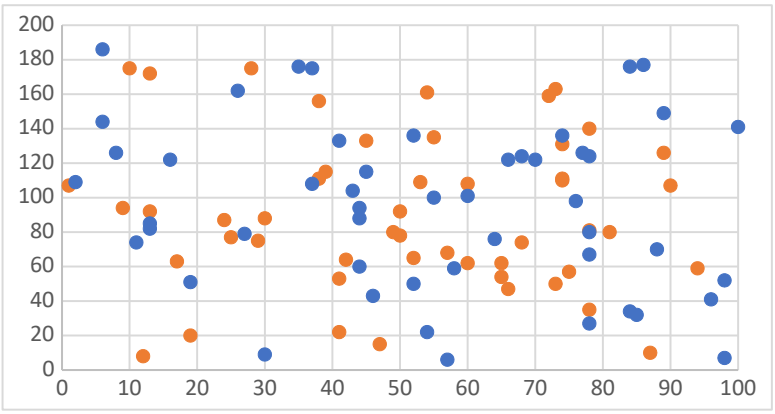


Рисунок 2. Множества без структуры



Попробуем на основании всего выше сказанного сформулировать требования к хорошо интерпретируемому алгоритму классификации:

1. высокие метрики интерпретируемости;
2. высокие заданные метрики качества классификации;
3. высокая локальная и глобальная объяснимость;
4. оценка информативности (направления и степени влияния) каждого признака;
5. понимание можно ли входные данные структурировать.

Причем, если принятие решения можно разбить на этапы (узлы), то сформулированные требования должны выполняться в каждом узле. По мнению автора, одним из таких алгоритмов может быть, алгоритм построения дерева секущих гиперплоскостей.

Дерево секущих гиперплоскостей

Древовидная структура алгоритма обычно хорошо воспринимается практическими специалистами из различных областей, если имеет мало узлов и ветвей [3,9,17]. Линейное разделение множеств гиперплоскостью, тоже хорошо интерпретируется потому, что по знакам коэффициентов гиперплоскости перед входными признаками можно понять направление влияния конкретного признака, а по значениям, если входные данные нормированы, понять степень этого влияния на выходной результат. Объединение этих двух подходов при наличии во входных данных даже сложной структуры (Рисунок 3) позволит построить решающее правило с гораздо меньшим количеством узлов, чем при классическом построении дерева решений, так как последнее является частным случаем дерева секущих гиперплоскостей. Дадим графическую иллюстрацию последовательности действий на условном примере.

Рисунок 3. Иллюстрация последовательности действий на условном примере

Графическая иллюстрация упрощает объяснение решающего правила заказчику исследования. Конечно, дать полноценную графическую иллюстрацию для n-мерного пространства достаточно затруднительно, но обычно вполне достаточно объяснения, что множество наблюдений в пространстве признаков можно разбить на отдельные области, если вместо линий использовать гиперплоскости, каждая из которых имеет конкретную формулу. Одним из способов построения такого решающего правила может быть следующая модель.

Модель линейного программирования с частично булевыми переменными для построения дерева секущих гиперплоскостей

Построение решающего правила делается итерационным способом. На каждой итерации отдельно для каждого класса определяются коэффициенты гиперплоскости, дающей максимальную долю правильных отсечений. Из двух гиперплоскостей в качестве секущей выбирается дающая максимальную долю правильных отсечений. Отсеченные наблюдения исключаются из дальнейшего рассмотрения и процесс отсечения повторяется до тех пор, пока доля неразделенных наблюдений станет меньше заданной величины или будет выполнено число заданных итераций. Данный процесс всегда сходится если нет абсолютно одинаковых наблюдений, одновременно принадлежащих двум классам [4].

Далее будет использована следующая система обозначений:

Для описания входных данных

J_1 и J_2– множества наблюдений, которые необходимо разделить;

I – множество признаков, которыми описывается наблюдение;

P_{ij} - i-ый признак j-ого наблюдения;

T - множество итераций для определения коэффициентов разделяющих гиперплоскостей;

Переменные определяемые на каждой итерации

a_{i1}^t - коэффициент t-ой гиперплоскости для i-ого признака для отсечения класса1;

b_1^t- свободный член t-ой гиперплоскости для отсечения класса1 ;

z_{j1}^t- булева переменная фиксирует расположение j-ого наблюдения направление относительно t-ой гиперплоскости для отсечения класса1;

a_{i2}^t - коэффициент t-ой гиперплоскости для i-ого признака для отсечения класса2;

b_2^t- свободный член t-ой гиперплоскости для отсечения класса2 ;

z_{j2}^t- булева переменная фиксирует расположение j-ого наблюдения направление относительно t-ой гиперплоскости для отсечения класса2;

N_1^tи N_2^t – число неправильно классифицированных наблюдений каждого класса на t-ой итерации;

d_1^tи d_2^t – доля правильно классифицированных каждого класса на t-ой итерации

На каждой итерации выполняются следующие действия:

Определение коэффициентов секущей гиперплоскости для класса1

$$\sum_{i \in I} P_{ij} * a_{i1}^t + b_1^t + L * z_{j1}^t \geq E \qquad j \in J_1 \ , \ t \in$$

$$\sum_{i \in I} P_{ij} * a_{i1}^t + b_1^t \leq - E \qquad j \in$$

$$\sum_{j \in J_1} z_{j1}^t = N_1^t$$

$$min\, N_1^t$$

Определение коэффициентов секущей гиперплоскости для класса2

$$\sum_{i \in I} P_{ij} * a_{i2}^t + b_2^t \geq E$$

$$\sum_{i \in I} P_{ij} * a_{i2}^t + b_2^t - L * z_{j2}^t \leq - E \qquad j \in J_2 \ , \ t \in$$

$$\sum_{j \in J_2} z_{j2}^t = N_2^t$$

$$min\, N_2^t$$

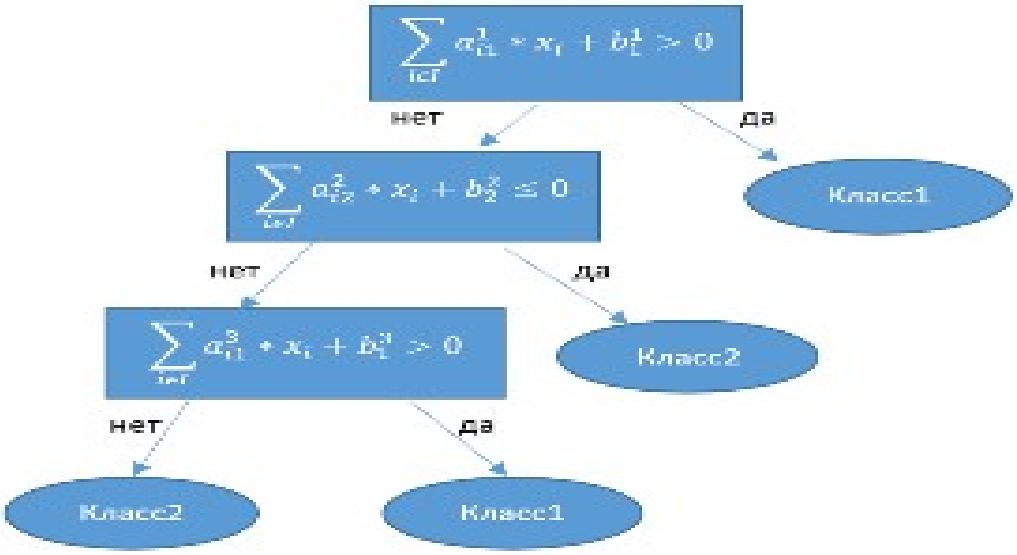
Определение долей правильной классификации на t-ой итерации

$$d_1^t = \frac{N_1^{t-1} - N_1^t}{N_1^{t-1}} \qquad N_1^0 \text{ - всего наблюдений в классе}$$

$$d_2^t = \frac{N_2^{t-1} - N_2^t}{N_2^{t-1}} \qquad N_2^0 \text{ - всего наблюдений в классе}$$

В качестве секущей гиперплоскости выбирается имеющая наибольшую долю правильной классификации. Количество неправильно классифицированных в это классе уменьшается, количество неправильно классифицированных в другом классе остается неизменным. Структура дерева решений для условного примера приведена на рисунке 4.

Рисунок 4. Дерево секущих плоскостей для условного примера



Следует отметить, что структура дерева секущих плоскостей может быть различной. Это зависит от структуры данных, используемых линейных разделителей и алгоритмов принятия решения в узлах. Для задач относительно небольшой размерности (несколько тысяч наблюдений) хорошие результаты дает приведенная выше модель. Более подробно такие модели объяснены в [5]. Для задач большой размерности вместо булевых переменных лучше использовать функцию Relu1 [6]. Первоначально оба подхода были апробированы и сравнены с другими методами машинного обучения на основе данных из репозитория UCI [22]. Результаты расчетов приведены в [6]. На практике расчеты на основе моделей линейного программирования с булевыми переменными использовались для решения задач медицинской диагностики, подбора технологических параметров для получения высококачественного агломерата и прогнозирования фондовых индексов. Так как классы были сильно несбалансированными, то оценка результаты по практическим задачам производилась по AUC ROC и результаты приведены в таблице 5.

Таблица 5. Результаты методов на тестовых выборках практических задач

Название задачи	Дерево секущих Гиперплоскостей	Дерево решений	RF	LD	SVM	LR	KNN	NB
Диагностика гипертензии	79.6	69.6	71.6	78.8	77.8	78.7	72.5	76.3
Диагностика гипотензии	74.3	67.5	73.3	73.9	70.4	71.9	71.1	62.2
Диагностика синдрома хронической усталости	74.1	67.0	77.3	73.3	75.0	72.7	65.3	74.3
Диагностика туберкулеза у детей	97.8	78.0	96.9	77.0	81.4	81.4	72.8	79.4
Диагностика туберкулеза у взрослых	96.1	80.9	76.3	66.6	63.5	68.3	66.1	60.4
Подбор технологических параметров при производстве металлургического агломерата	80.5	65.3	83.1	68.7	69.2	71.5	64.2	60.8
Прогнозирование индекса S&P500	85.2	47.2	46.1	45.5	47.3	46.2	47.1	51.1

Обозначения в таблице 5

RF – случайный лес

LD – линейный дискриминант

SVM – метод опорных векторов

LR – логистическая регрессия

KNN – метод ближайших соседей

NB – наивный байес

Выбранные практические задачи характеризуются различной структурой данных и выраженной несбалансированностью классов, что делает их показательными для оценки как качества классификации, так и интерпретируемости решающих правил. Анализ таблицы 5 показывает, что дерево секущих плоскостей на основе модели линейного программирования с частично булевыми переменными практически всегда дает результаты лучше, чем просто дерево решений, но иногда проигрывает случайному лесу. Конечно, эти результаты нельзя обобщать на все случаи, но можно сделать вывод, что в именно в этих задачах данные могут быть структурированы на основе линейных разделителей.

Заключение

Приведенный в данной статье подход к решению задач классификации отвечает ранее сформулированным требованиям к хорошо интерпретируемым моделям и на его основе можно получать легко интерпретируемые решающие правила. Графические интерпретации решающего правила дают понимание структуры входных данных, а формулы секущих гиперплоскостей позволяют оценить информативность признаков, направление и степень влияния на качество классификации. Исследования в данном направлении продолжаются. По мнению автора, перспективным является использование в узлах дерева более крупных структур (комитетов единогласия и выпуклых оболочек множеств). К числу недостатков метода следует отнести то, что он хорошо работает на задачах относительно небольшой размерности. Для увеличения размерности решаемых задач исследуется возможность использования других линейных разделителей.

Литература

1. Бирюков Д.Н., Дудкин А.С. Объяснимость и интерпретируемость – важные аспекты безопасности решений, принимаемых интеллектуальными системами (обзорная статья) // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 3. С. 373–386.
2. Волков Е.Н., Аверкин А.Н. Локальные объяснения для больших языковых моделей: краткий обзор методов. XXVII Международная конференция по мягким вычислениям и измерениям, 22-24 мая 2024, Санкт-Петербург, с. 239-242
3. Воронцов К.В. Интерпретируемость и объяснимость в машинном обучении [Электронный ресурс]. URL: https://www. MachineLearning.ru/wiki
4. Мазуров. Вл. Д. Метод комитетов в задачах оптимизации и классификации / Вл.Д. Мазуров - М.: Наука. - 1990. - 248 с.
5. Чернавин П.Ф., Гайнанов Д.Н., Панкращенко В.Н. Чернавин Ф.П., Чернавин Н.П. (2021) Машинное обучение на основе задач математического программирования // М.: Наука, 2021, 128 с.
6. Чернавин П.Ф., Чернавин Ф.П., Андросов Д.А., Чернавин Н.П. Дерево секущих плоскостей. Прикаспийский журнал: управление и высокие технологии 2023 № 2 (62) с. 18-25
7. Штукатуров С. Кризис машинного обучения в научных исследованиях: обладает ли научной ценностью эксперимент, результаты которого не удалось воспроизвести. [Электронный ресурс]. URL: troger.ru/translation/machine-learning-crisis

8. Arya V., Bellamy R., Chen P.-Yu., Dhurandhar A., Hind M., et al. One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. 2019. [Электронный ресурс]. URL: https://doi.org/10.48550/arXiv.1909.03012
9. Bach F. Machine Learning Research Blog. 2025, [Электронный ресурс]. URL: francisbach.com
10. The Bletchley Declaration by Countries Attending the AI Safety Summit 1-2 November 2023. [Электронный ресурс]. URL: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023
11. Can we trust scientific discoveries made using machine learning? 2019. [Электронный ресурс]. URL: https://eurekalert.org/news-releases/611930
12. Gary Marcus. The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence. 2020, [Электронный ресурс] URL: https://arxiv.org/abs/2002.06177
13. Gilpin L.H., Bau D., Yuan B.Z., Bajwa A., Specter M., Kagal L. Explaining explanations: an overview of interpretability of machine learning. [Электронный ресурс] URL: https://doi.org/10.48550/arXiv.1806.00069
14. Gunning D., Stefik M., Choi J., Miller T., Stumpf S., Yang G.-Z. XAI — Explainable artificial intelligence // Science Robotics. 2019. [Электронный ресурс] URL: https://doi.org/10.1126/scirobotics.aay7120
15. Lankow, J. and Ritchie, J. and Crooks, R. Chapter 7. Data Visualization Interfaces // Infographics: The Power of Visual Storytelling. — Wiley, 2012. — 264 p
16. Miller T. Explanation in artificial intelligence: insights from the social sciences // Artificial Intelligence. 2019. V. 267. P. 1–38.
17. Molnar C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Christoph Molnar, 2025. 392 p.
18. Namatēvs I., Sudars K., Dobrājs A. Interpretability versus explainability: classification for understanding deep learning systems and models // Computer Assisted Methods in Engineering and Science. 2022. V. 29. N 4. P. 297–356.
19. Rahimi A. Recht B. Random features for Large-Scale Kernel Machine. [Электронный ресурс]. URL: https://dl.acm.org/doi/10.5555/2981562.2981710
20. Samek W., Montavon G., Vedaldi A., Hansen L.K., Muller K.-R. Explainable AI: interpreting, explaining and visualizing deep learning // Lecture Notes in Computer Science. 2019. V. 11700. 439 p
21. Yuan W., Liu P., Neubig G. Can we automate scientific reviewing? 2021. [Электронный ресурс]. URL: https://doi.org/10.48550/arXiv.2102.00176
22. UCI Machine Learning Repository. URL: https://archive.ics.uci.edu/dataset/.

References in Cyrillics

1. Biryukov D.N., Dudkin A.S. Explainability and interpretability as important aspects of safety of decisions made by intelligent systems: a review article. Scientific and Technical Journal of Information Technologies, Mechanics and Optics, 2025, vol. 25, no. 3, pp. 373–386.
2. Volkov E.N., Averkin A.N. Local explanations for large language models: a brief overview of methods. Proceedings of the XXVII International Conference on Soft Computing and Measurements, May 22–24, 2024, St. Petersburg, pp. 239–242.
3. Vorontsov K.V. Interpretability and explainability in machine learning. [Elektronnyj resurs]. URL:https://www.MachineLearning.ru/wiki
4. Mazurov V.D. Committee methods in optimization and classification problems. Moscow: Nauka, 1990, 248 p.
5. Chernavin P.F., Gainanov D.N., Pankrashchenko V.N., Chernavin F.P., Chernavin N.P. Mashinnoe obuchenie na osnove zadach matematicheskogo programmirovaniya // M.: Nauka, 2021, 128 s.
6. Chernavin P.F., Chernavin F.P., Androsov D.A., Chernavin N.P. Tree of secant planes. Caspian Journal: Management and High Technologies, 2023, no. 2 (62), pp. 18–25.
7. Shtukaturov S. The crisis of machine learning in scientific research: does an experiment have scientific value if its results cannot be reproduced? [Elektronnyj resurs]. URL: troger.ru/translation/machine-learning-crisis

Чернавин Павел Федорович,
Уральский федеральный университет (chernavin.p.f@gmail.com)

Ключевые слова
машинное обучение, задачи классификации, интерпретируемое решение, линейное разделение множеств, дерево решений, математическое программирование.

Chernavin Pavel, Interpretable decision rules based on a tree of secant hyperplanes

Keywords
machine learning, classification problems, interpretable decision, linearly separable sets, decision tree, mathematical programming

Abstract
The article analyzes the differences between interpretable and explainable artificial intelligence. These two concepts are considered by the author as distinct, yet complementary, approaches to solving machine learning problems.

Based on the conducted analysis, the author proposes representing decision rules for classification tasks in the form of a tree structure, where the nodes contain not individual input features but separating hyperplanes. This approach makes it possible to assess the informativeness of each individual input feature as well as of the entire set of features at each node, and to obtain an overall interpretable decision rule for the whole model while achieving high classification performance metrics.