УДК 004.89:16

НАРУШЕНИЕ ЗАКОНА ИСКЛЮЧЕННОГО ТРЕТЬЕГО В SOFTMAX-КЛАССИФИКАТОРАХ КАК СИСТЕМНАЯ ПРОБЛЕМА ЛОГИЧЕСКОЙ НЕПРОТИВОРЕЧИВОСТИ ИИ

Охотенко А.Л., магистрант

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Шаталова В. В. – канд. техн. наук, доцент

В статье рассматривается имплицитное противоречие между архитектурой softmax-классификаторов и требованиями классической двузначной логики. Показано, что интерпретация вероятностного выхода как истинностного значения приводит к кажущемуся, но системно значимому нарушению закона исключенного третьего. Анализируются последствия этого разрыва для построения дедуктивных выводов на основе предсказаний нейросетевых моделей. Делается вывод о необходимости отказа от бинарной семантики при интеграции вероятностных классификаторов в логически непротиворечивые системы искусственного интеллекта.

Введение. Современные системы искусственного интеллекта, основанные на глубоких нейронных сетях, демонстрируют высокую эффективность в задачах классификации. Стандартным выходным слоем для многоклассовой классификации является softmax-функция, преобразующая действительные значения в вероятностное распределение. Однако при попытке интеграции таких классификаторов в логические дедуктивные системы возникает фундаментальная проблема: выход softmax не удовлетворяет закону исключенного третьего, который является базовым принципом классической логики.

Закон исключенного третьего утверждает, что для любого высказывания P истинно либо P, либо его отрицание $\neg P$, и третьего не дано [1]. В контексте классификации это означает, что объект либо принадлежит заданному классу, либо не принадлежит. Softmax же всегда выдает распределение, в котором ни одна вероятность, как правило, не равна ни единице, ни нулю. Это создает ситуацию, когда формальный логический вывод становится невозможным без дополнительных соглашений, а сам ИИ оказывается логически несостоятельным, если к нему применять критерии истинности.

Цель работы – показать, что нарушение закона исключенного третьего в softmax-классификаторах является не ошибкой реализации, а системным свойством, порождающим логическую непротиворечивость на уровне интерпретации. В отличие от распространенных замечаний о вероятностной природе softmax, здесь утверждается, что проблема носит не количественный, а категориальный характер.

Основная часть. Для классического понимания истинности необходимо, чтобы каждое высказывание имело одно из двух значений: истина или ложь. В контексте классификатора высказывание об объекте формулируется как «объект принадлежит данному классу». Softmax же выдаёт для каждого класса число строго между нулём и единицей, и сумма этих чисел по всем классам равна единице.

Из этого следуют два важных обстоятельства. Во-первых, ни один класс никогда не получает значения, равного в точности единице. Во-вторых, ни один класс никогда не получает значения, равного в точности нулю. Это прямое следствие того, как устроена функция: она всегда даёт положительные результаты, и если один результат приближается к единице, остальные становятся очень маленькими, но всё же ненулевыми.

Если попытаться интерпретировать выход softmax как истинностное значение высказывания о принадлежности к классу, то окажется, что ни само высказывание, ни его отрицание не являются истинными в двузначном смысле. Исходное высказывание получает значение, отличное от единицы, а его отрицание — значение, отличное от единицы, поскольку отрицание интерпретируется как дополнение до единицы, которое также лежит строго между нулём и единицей. Таким образом, мы сталкиваемся с ситуацией, когда третье значение не только существует, но и заполняет весь интервал между ложью и истиной.

Это прямое нарушение закона исключённого третьего. Более того, нарушается и более фундаментальный принцип бивалентности — утверждение о том, что каждое высказывание либо истинно, либо ложно. Softmax в принципе не способен породить бивалентную оценку, если только входные сигналы не являются бесконечно большими или бесконечно малыми, что в реальных вычислительных системах недостижимо.

Если попытаться интерпретировать выход softmax как истинностное значение высказывания о принадлежности к классу, то окажется, что ни само высказывание, ни его отрицание не являются истинными в двузначном смысле. Исходное высказывание получает значение, отличное от единицы, а его отрицание – значение, отличное от единицы, поскольку отрицание интерпретируется как дополнение

до единицы, которое также лежит строго между нулём и единицей. Таким образом, мы сталкиваемся с ситуацией, когда третье значение не только существует, но и заполняет весь интервал между ложью и истиной.Softmax как формальная конструкция, несовместимая с бинарной истинностью. Пусть задан классификатор с *K* классами. Softmax-функция имеет вид, представленный формулой 1:

$$\sigma(z_i) = \frac{exp\ (z_i)}{\sum_{i=1}^{K} \exp(z_j)}, \quad (1)$$

где *I = 1, … , K.*

Сумма всех $\sigma(z_i)$ тождественно равна 1. Каждое $\sigma(z_i)$ лежит в интервале (0,1). Строгое неравенство достигается при любых конечных z_i, что исключает значения 0 и 1. Исключением является лишь асимптотический случай бесконечно больших аргументов, не реализуемый в реальных вычислениях [2].

Следовательно, для любого объекта x и любого класса C_i softmax-классификатор выдает значение $p_i = P(C_i|x)$ такое, что 0 < p_i < 1. Если попытаться интерпретировать p_i как истинностное значение высказывания «*x* принадлежит C_i», то окажется, что ни это высказывание, ни его отрицание не являются истинными в двузначном смысле. Для отрицания «*x* не принадлежит C_i» истинностным значением должно быть 1 − p_i, но оно также лежит строго между 0 и 1.

Формально: для любого i имеет место формула 2:

$$\neg(p_i = 1) \wedge \neg(p_j = 0) \quad (2)$$

Это прямо противоречит закону исключенного третьего, который требует $p_i = 1 \vee p_j = 0$. Возникает ситуация tertium datur – третье дано, и оно заполняет весь интервал (0,1).

Часто возражение формулируется так: softmax не обязан подчиняться логике высказываний, поскольку он выражает степень уверенности, а не истину [3]. Однако это возражение не снимает проблемы, а лишь переименовывает ее. В любой системе ИИ, которая используется для автоматического принятия решений, требуется переход от числового выхода к бинарному действию: отклонить заявку или одобрить, поставить диагноз или нет. Этот переход осуществляется через пороговое правило (argmax или сравнение с порогом). Но введение порога есть внешняя по отношению к модели операция, не выводимая из самой модели.

Более того, если один и тот же классификатор применяется в двух различных выводах (например, в цепочке правил «если кошка, то млекопитающее» и «если пушистое, то кошка»), то вероятности не являются транзитивными. *Из p(C|B)* = 0.6 и *p(B|A)* = 0.7 нельзя получить p(C|A) классическими логическими средствами. Это разрушает композиционность – ключевое свойство логических систем.

Таким образом, нарушение закона исключенного третьего в softmax ведет к системной проблеме: классификатор не может быть одновременно вероятностно корректным и логически непротиворечивым в классическом смысле.

Попытка сохранить логическую непротиворечивость приводит к отказу от информации о неопределенности. Если округлить выход softmax до бинарного вектора (единица на argmax, нули на остальных), то закон исключенного третьего восстанавливается. Однако при этом теряется различие между уверенностью 0.51 и 0.99, что критично во многих приложениях (медицина, автопилоты, финансы).

Обратная стратегия – сохранить вероятности, но заменить классическую логику на вероятностную. Это решает проблему противоречия с ЗИТ, но порождает новую: вероятностная логика не обладает свойством монотонности, и вывод становится зависимым от порядка поступления свидетельств. Кроме того, истинность в вероятностной логике не является истинностью в обычном смысле, что делает такую систему ИИ логически непрозрачной для человека.

Следовательно, перед разработчиком стоит жесткий выбор: либо жертвовать логической непротиворечивостью (в классическом понимании), либо жертвовать полнотой использования выходной информации softmax. Ни один из путей не является безупречным.

Рассмотрим простой пример. Имеется softmax-классификатор с тремя классами: A, B, C. Для объекта *x* получены вероятности: A=0.45, B=0.35, C=0.20. Классификатор выбирает A.

С точки зрения классической логики, утверждается A, но модель одновременно утверждает (в вероятностном смысле), что *P(¬A)* = 0.55, то есть отрицание A более вероятно, чем само A. Если принять порог 0.5 для истинности, то одновременно истинны «не-A» и «A», что есть классическое противоречие. Если не принимать порог, то невозможно высказать ни одного логического суждения.

Такая ситуация не является исключительной – она нормальна для softmax. Это означает, что любой классификатор данного типа при работе в реальном диапазоне значений (без насыщения) систематически порождает логически противоречивую интерпретацию, если настаивать на бинарной истинности.

Из проведённого анализа следует, что интеграция softmax-классификатора в систему, требующую дедуктивного вывода, невозможна без одного из следующих решений:

1. Отказ от классической логики в пользу нечёткой, вероятностной или многозначной. Это требует пересмотра всего механизма вывода.
2. Преобразование softmax-выхода в бинарный вектор с потерей информации о неуверенности. Это допустимо только в задачах, где цена ошибки низка.
3. Введение мета-уровня, который оперирует не высказываниями, а распределениями, и выводит распределения, а не истинностные значения (байесовские сети, вероятностное программирование).

Нарушение закона исключённого третьего перестаёт быть проблемой, если явно признать, что ИИ на основе softmax не порождает логических суждений. Он порождает числовые оценки, которые лишь затем, по внешним правилам, конвертируются в суждения. Ответственность за логическую непротиворечивость лежит не на модели, а на правиле конверсии.

Заключение. Проведенный анализ показывает, что нарушение закона исключенного третьего в softmax-классификаторах представляет собой не случайный дефект, а фундаментальное следствие вероятностной природы их выходов. При попытке интерпретировать эти выходы как истинностные значения возникает системное противоречие: ни само высказывание о принадлежности классу, ни его отрицание не могут быть признаны истинными в двузначной логике.

Это противоречие делает классический логический вывод поверх softmax-классификаторов принципиально некорректным без дополнительных мета-правил, которые всегда произвольны. Следовательно, любые гибридные системы, сочетающие нейросетевую классификацию и логический вывод, должны либо явно отказаться от закона исключенного третьего, перейдя к многозначным или вероятностным логикам, либо ограничить область применения классификаторов задачами без дедуктивных цепочек.

Практическим следствием является необходимость пересмотра критериев логической корректности для современных ИИ. Требование непротиворечивости не может быть механически перенесено с формальных систем на нейросетевые компоненты. Вместо этого следует разрабатывать композитные архитектуры, где логический вывод оперирует не с «сырыми» вероятностями, а с их преобразованиями, сохраняющими непротиворечивость в смысле неклассической логики.

Список использованных источников:

1.Закон исключённого третьего [Электронный ресурс]. – Режим доступа: https://gtmarket.ru/concepts/6974. – Дата доступа: 10.02.2026.

2. Функция Softmax [Электронный ресурс]. – Режим доступа: https://aiew.ru/glossary/softmax/. – Дата доступа: 15.03.2026.

3. Why Softmax Lies in Production: Better Uncertainty from Logits (Without Changing Your Model) [Электронный ресурс]. – Режим доступа: https://medium.com/@muhibuddin12/why-softmax-lies-in-production-better-uncertainty-from-logits-without-changing-your-model-bdad0df9996b. – Дата доступа: 31.03.2026.

UDC 004.89:16

VIOLATION OF THE LAW OF THE EXCLUDED MIDDLE IN SOFTMAX CLASSIFIERS AS A SYSTEMIC PROBLEM OF LOGICAL CONSISTENCY OF AI

Okhotenko A.L., master's student

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Shatalova V.V. – Cand. of Sci., Associate Professor of the Department of IPiE

Annotation. This article examines the implicit contradiction between the architecture of softmax classifiers and the requirements of classical two-valued logic. It is shown that interpreting the probabilistic output as a truth value leads to an apparent but systemically significant violation of the law of excluded middle. The implications of this gap for deductive inference based on the predictions of neural network models are analyzed. A conclusion is reached regarding the need to abandon binary semantics when integrating probabilistic classifiers into logically consistent artificial intelligence systems.

Keywords. Law of excluded middle, softmax, logical consistency, classifier, uncertainty, semantic gap.