ФЕДЕРАЛЬНОЕ АГЕНТСТВО

ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

ПРЕДВАРИТЕЛЬНЫЙ НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

ПНСТ 835— 2023 (ISO/IEC TS 4213:2022)

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

Оценка эффективности моделей и алгоритмов машинного обучения в задаче классификации

(ISO/IEC TS 4213:2022, Information technology — Artificial intelligence — Assessment of machine learning classification performance, MOD)

Издание официальное

Москва
Российский институт стандартизации
2023

Предисловие

1 ПОДГОТОВЛЕН Федеральным государственным автономным образовательным учреждением высшего образования «Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ) на основе собственного перевода на русский язык англоязычной версии документа, указанного в пункте 4

2 ВНЕСЕН Техническим комитетом по стандартизации ТК 164 «Искусственный интеллект»

3 УТВЕРЖДЕН И ВВЕДЕН В ДЕЙСТВИЕ Приказом Федерального агентства по техническому регулированию и метрологии от 16 ноября 2023 г. № 59-пнст

4 Настоящий стандарт является модифицированным по отношению к международному документу ISO/IEC TS 4213:2022 «Информационные технологии. Искусственный интеллект. Оценка эффективности классификации машинного обучения» (ISO/IEC TS 4213:2022 «Information technology — Artificial intelligence — Assessment of machine learning classification performance», MOD) путем изменения отдельных фраз (слов, значений показателей, ссылок), которые выделены в тексте курсивом.

Внесение указанных технических отклонений направлено на учет особенностей российской национальной стандартизации.

Сведения о соответствии ссылочных национальных стандартов международным стандартам, использованным в качестве ссылочных в примененном международном стандарте, приведены в дополнительном приложении ДА.

Наименование настоящего стандарта изменено относительно наименования указанного международного документа для увязки с наименованиями, принятыми в существующем комплексе национальных стандартов Российской Федерации

Правила применения настоящего стандарта и проведения его мониторинга установлены в ГОСТ Р 1.16—2011 (разделы 5 и 6).

Федеральное агентство по техническому регулированию и метрологии собирает сведения о практическом применении настоящего стандарта. Данные сведения, а также замечания и предложения по содержанию стандарта можно направить не позднее чем за 4 мес до истечения срока его действия разработчику настоящего стандарта по адресу: info@tc164.ru и/или в Федеральное агентство по техническому регулированию и метрологии по адресу: 123112 Москва, Пресненская набережная, д. 10, стр. 2.

В случае отмены настоящего стандарта соответствующая информация будет опубликована в ежемесячном информационном указателе «Национальные стандарты» и также будет размещена на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет (www.rst.gov.ru)

Содержание

Введение

Ввиду того, что исследователи научных, коммерческих и правительственных организаций продолжают совершенствовать модели машинного обучения, при оценке эффективности моделей и алгоритмов машинного обучения в задаче классификации необходимо применять последовательные подходы и методы.

О достижениях в области машинного обучения часто сообщают с точки зрения повышения эффективности по сравнению с современным состоянием или релевантным базовым уровнем. Выбор подходящего показателя для оценки эффективности моделей и алгоритмов машинного обучения в задаче классификации зависит от вариантов использования и ограничений предметной области. Кроме того, выбранный показатель может отличаться от показателя, используемого во время обучения. Эффективность моделей и алгоритмов машинного обучения в задаче классификации может быть представлена следующими примерами:

- новая модель обеспечивает долю правильных исходов классификации набора данных 97,8 %, тогда как наиболее современная модель демонстрирует лишь долю правильных исходов 96,2 %;
- новая модель обеспечивает долю правильных исходов классификации, эквивалентную современной, но требует гораздо меньше данных для обучения, чем современные подходы;
- новая модель генерирует выводы в 100 раз быстрее, чем современные модели, при сохранении эквивалентной доли правильных исходов.

Для определения корректности таких утверждений необходимо учитывать такие факторы оценки эффективности моделей и алгоритмов машинного обучения в задаче классификации, как спецификация модели, состав набора данных и оценка результатов. В настоящем стандарте описаны подходы и методы для обеспечения релевантности, легитимности и обобщаемости получаемых оценок эффективности моделей и алгоритмов машинного обучения в задаче классификации.

Заинтересованные в развитии искусственного интеллекта (ИИ) стороны, перечисленные в [1], могут в своих целях использовать подходы и методы, описанные в настоящем стандарте. Например, разработчики ИИ могут использовать эти подходы и методы при оценке моделей и алгоритмов машинного обучения.

Для оценки эффективности моделей и алгоритмов машинного обучения применяются методологические средства контроля в целях достижения достоверности и репрезентативности результатов. Примеры таких средств контроля включают создание вычислительной среды, выбор и подготовку набора данных, а также ограничение утечки данных, которая потенциально может привести к искажению результатов классификации. Пункт 5 посвящен этой теме.

Построение выводов об эффективности моделей и алгоритмов машинного обучения исключительно на основе доли правильных исходов может быть некорректным в зависимости от характеристик обучающих и вводных данных. В том случае, если классификатор восприимчив к наиболее частотным классам в выборке, несбалансированные данные для обучения могут приводить впоследствии к завышению показателя доли правильных исходов и отражать лишь априорную вероятность наиболее частотного из классов. Дополнительные оценки, отражающие неочевидные аспекты эффективности моделей и алгоритмов машинного обучения в задаче классификации, такие как макроусредненные показатели, в ряде случаев более информативны. Кроме того, различные типы классификации машинного обучения, такие как бинарная (binary classification), мультиклассовая (multi-class) и классификация с несколькими метками (multi-label), связаны с конкретными показателями эффективности. В дополнение к этим показателям могут иметь значение другие аспекты оценки моделей и алгоритмов в задаче классификации, такие как скорость работы, производительность и энергопотребление. Пункт 6 посвящен этим темам.

Разнородные распределения данных для обучения могут приводить к сложностям в эксплуатации. Для определения условий, при которых эффективность моделей и алгоритмов машинного обучения в задаче классификации существенно различается, проводятся статистические тесты значимости. При разработке моделей и алгоритмов машинного обучения используются специальные методологии обучения, проверки и тестирования для учета ряда потенциальных сценариев. Пункт 7 посвящен этим темам.

В приложении А приведен расчет эффективности для задачи многоклассовой классификации с примерами положительной и отрицательной классификации. В приложении В приведена ROC-кривая, полученная на основе данных, представленных в приложении А.

В приложении С приведена сводная информация об эталонных тестах моделей и алгоритмов машинного обучения для решения задачи классификации.

В приложении D представлены данные о причинно-специфической смертности с поправкой на случайность как пример использования модели машинного обучения в задаче классификации. Иные вопросы, связанные с бенчмарк-тестированием, применением или вариантами использования, не рассматриваются.

ПРЕДВАРИТЕЛЬНЫЙ НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

Оценка эффективности моделей и алгоритмов машинного обучения в задаче классификации

Artificial intelligence. Evaluating the effectiveness of machine learning models and algorithms in problem classification

**Срок действия — с 2024—01—01
до 2027—01—01**

1 Область применения

Настоящий стандарт устанавливает методологию оценки эффективности моделей и алгоритмов машинного обучения в задаче классификации.

2 Нормативные ссылки

В настоящем стандарте использованы нормативные ссылки на следующие стандарты:

ПНСТ 838—2023 (ИСО/МЭК 23053:2022) Искусственный интеллект. Структура описания систем искусственного интеллекта, использующих машинное обучение

ПНСТ 839—2023 (ISO/IEC TR 24027:2021) Искусственный интеллект. Смещенность в системах искусственного интеллекта и при принятии решений с помощью искусственного интеллекта

Примечание — При пользовании настоящим стандартом целесообразно проверить действие ссылочных стандартов в информационной системе общего пользования — на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет или по ежегодному информационному указателю «Национальные стандарты», который опубликован по состоянию на 1 января текущего года, и по выпускам ежемесячного информационного указателя «Национальные стандарты» за текущий год. Если заменен ссылочный стандарт, на который дана недатированная ссылка, то рекомендуется использовать действующую версию этого стандарта с учетом всех внесенных в данную версию изменений. Если заменен ссылочный стандарт, на который дана датированная ссылка, то рекомендуется использовать версию этого стандарта с указанным выше годом утверждения (принятия). Если после утверждения настоящего стандарта в ссылочный стандарт, на который дана датированная ссылка, внесено изменение, затрагивающее положение, на которое дана ссылка, то это положение рекомендуется применять без учета данного изменения. Если ссылочный стандарт отменен без замены, то положение, в котором дана ссылка на него, рекомендуется применять в части, не затрагивающей эту ссылку.

3 Термины и определения

В настоящем стандарте применены термины по [*1*] и *ПНСТ 838*, а также следующие термины с соответствующими определениями.

3.1 Классификация и связанные термины

3.1.1

классификация (classification): Метод отнесения элементов определенного типа (объектов или документов) к различным классам и подклассам в соответствии с их характеристиками.

[[*2*], пункт 3.1]

Издание официальное

3.1.2 **классификатор** (classifier): Обученная модель и ассоциированный с ней механизм для осуществления классификации (см. 3.1.1).

3.2 Показатели и связанные термины

3.2.1 **оценка** (evaluation): Процесс сравнения прогнозов классификации (см. 3.1.1), выполненных моделью на данных, с разметкой данных.

3.2.2 **ложноотрицательный образец;** *пропуск; ошибка второго рода* (false negative; miss; type II error), F_N: Образец, ошибочно классифицированный как отрицательный.

3.2.3 **ложноположительный образец F_P;** *ложная тревога; ошибка первого рода* (false positive; false alarm; type I error): Образец, ошибочно классифицированный как положительный.

3.2.4 **истинно положительный образец T_P** (true positive): Образец, верно классифицированный как положительный.

3.2.5 **истинно отрицательный образец T_N** (true negative): Образец, верно классифицированный как отрицательный.

3.2.6 **доля правильных исходов** (accuracy): Количество верно классифицированных образцов, разделенное на количество всех классифицированных образцов.

Примечание — Пример расчета

$$a = (T_p + T_N)/(T_p + F_p + T_N + F_N). \quad (1)$$

3.2.7 **матрица ошибок** (confusion matrix): Матрица, используемая для регистрации количества правильно и неправильно классифицированных (см. 3.1.1) образцов.

3.2.8 **F_1-число;** *F-число; F-мера, F_1-мера* (F_1 score; *F*-score; *F*-measure; F_1-measure): Среднее гармоническое значение точности (см. 3.2.9) и полноты (см. 3.2.10).

Примечание — Пример расчета

$$F_1 = 2T_P/(2T_P + F_P + F_N). \quad (2)$$

3.2.9 **точность;** *прогностическая ценность положительного результата* (precision; positive predictive value): Количество образцов, верно классифицированных как положительные, разделенное на количество образцов, классифицированных как положительные.

Примечание — Пример расчета

$$p = T_P/(T_P + F_P). \quad (3)$$

3.2.10 **полнота;** *чувствительность; процент попаданий; доля истинно положительных образцов* (recall; sensitivity; hit rate; true positive rate): Количество образцов, верно классифицированных как положительные, разделенное на количество всех положительных образцов.

Примечание — Пример расчета

$$r = T_P/(T_P + F_N). \quad (4)$$

3.2.11 **специфичность;** *избирательность; доля истинно отрицательных образцов* (specificity; selectivity; true negative rate): Количество образцов, верно классифицированных как отрицательные, деленное на количество всех отрицательных образцов.

Примечание — Пример расчета

$$s = T_N/(T_N + F_P). \quad (5)$$

3.2.12 **доля ложноположительных образцов;** *отрицательный побочный эффект* (false positive rate; fall-out): Количество образцов, неверно классифицированных как положительные, деленное на количество всех отрицательных образцов.

Примечание — Пример расчета

$$F_{P,R} = F_P/(F_P + T_N). \quad (6)$$

3.2.13 **кумулятивная кривая отклика;** *gain-кривая* (cumulative response curve; gain chart): Графический метод отображения доли истинно положительных образцов (см. 3.2.10) и процентного соотношения положительных прогнозов для общих данных для различных пороговых значений.

3.2.14 **lift-кривая** (lift curve): Графический метод отображения по оси ординат доли истинно положительных образцов (см. 3.2.10) между моделью и случайным классификатором, а по оси абсцисс процентного соотношения положительных прогнозов для общих данных для различных пороговых значений.

3.2.15 **кривая точности полноты;** PRC (precision recall curve; PRC): Графический метод отображения полноты (см. 3.2.10) и точности (см. 3.2.9) для различных пороговых значений.

Примечание — Для отражения производительности при работе с несбалансированными данными более подходящей является PRC, а не ROC-кривая (рабочая характеристика приемника).

3.2.16 **кривая рабочей характеристики приемника;** *ROC-кривая* (receiver operating characteristic curve; ROC curve): Графический метод отображения доли истинно положительных образцов (см. 3.2.10) и ложноположительных образцов (см. 3.2.12) для различных пороговых значений.

3.2.17 **кросс-валидация** (cross-validation): Метод оценки эффективности метода машинного обучения с использованием одного набора данных.

Примечание — Кросс-валидация обычно используется для выбора гиперпараметров перед обучением конечной модели.

3.2.18 **наиболее встречаемый класс** (majority class): Класс, встречающийся в наборе данных наиболее часто.

4 Сокращения

ИИ — искусственный интеллект;
ANOVA — дисперсионный анализ;
AUPRC — площадь под кривой точности полноты;
AUROC — площадь под ROC-кривой;
ЦПТ — центральная предельная теорема;
ЦП — центральный процессор;
CRC — кумулятивная кривая отклика;
FC — полносвязный слой;
FDR — частота ложных обнаружений;
IoU — пересечение над объединением;
GPU — графический процессор;
ROC — рабочая характеристика приемника.

5 Общие принципы

5.1 Обобщенный процесс оценки эффективности моделей и алгоритмов машинного обучения в задаче классификации

Обобщенный процесс оценки эффективности моделей и алгоритмов машинного обучения в задаче классификации показан на рисунке 1.

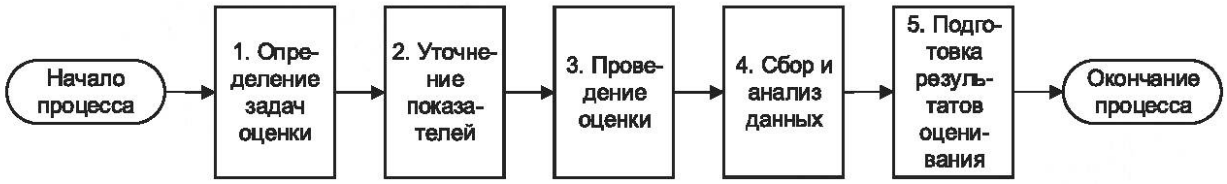


Рисунок 1 — Обобщенный процесс оценки эффективности моделей и алгоритмов машинного обучения в задаче классификации

Шаг 1. Определение задач оценки. Определите соответствующую задачу или задачи классификации для проведения оценки.

Шаг 2. Определение показателей. В зависимости от задачи классификации определите необходимый показатель или показатели.

Шаг 3. Проведение оценки. Создайте план оценки, подготовьте среду для проведения оценки, включив в нее программное и аппаратное обеспечение, подготовьте и обработайте наборы данных.

Шаг 4. Сбор и анализ данных. В соответствии с определенными показателями соберите выходные данные модели, такие как прогнозы классификации для каждой выборки.

Шаг 5. Подготовка результатов оценивания. Подготовьте результаты оценивания на основе определенных показателей и другой имеющейся информации.

5.2 Цель оценки эффективности моделей и алгоритмов машинного обучения в задаче классификации

Цель оценки и базовые требования могут существенно различаться на этапах «проектирование и разработка» и «проверка и валидация».

Целью оценки на этапе «проектирование и разработка» являются выбор архитектуры модели, конструирование и отбор значимых признаков, подбор и оптимизация гиперпараметров для достижения лучшей эффективности в задаче классификации. Целью оценки на этапе «проверка и валидация» является оценка эффективности обученной модели.

Оценку эффективности проводят в целях:

- оценки надежности прогнозов модели, ожидаемой частоты и масштаба ошибок;
- сравнения двух и более моделей для выбора одной из них;
- сравнения на отложенной выборке данных как в части самих примеров для обучения (out-of-sample), так и в части временного периода (out-of-time) для анализа эффективности работы на новых данных в условиях, приближенных к среде эксплуатации.

5.3 Критерии контроля при оценке эффективности моделей и алгоритмов машинного обучения в задаче классификации

5.3.1 Общие положения

При проведении оценки эффективности моделей и алгоритмов машинного обучения в задаче классификации следует применять последовательные подходы и методы для подтверждения релевантности, применимости и способности к обобщению. Особое внимание следует уделять при проведении сравнительной оценки нескольких моделей, алгоритмов или систем машинного обучения в задаче классификации.

5.3.2 Репрезентативность и несмещенность данных

За исключением случаев, предусмотренных постановкой цели, данные для обучения и тестирования должны быть максимально свободными от систематической ошибки выборки. То есть распределение признаков и классов в обучающих данных должно максимально соответствовать их распределению в реальном мире. Данные для обучения могут не соответствовать возможному сценарию применения модели. Например, в случае беспилотных автомобилей оценка эффективности моделей машинного обучения в задаче классификации, обученных на замкнутых трассах, а не на открытых дорогах, может быть вполне целесообразна для разработки прототипов систем. Данные, используемые для тестирования модели и алгоритма машинного обучения, должны отражать предполагаемое использование системы.

Данные могут быть искаженными, неполными, устаревшими, непропорциональными или содержать исторические погрешности. Такие нежелательные смещения могут порождать смещения в данных для обучения и снижать эффективность обучения модели. Если среда эксплуатации модели машинного обучения сложна и многогранна, ограниченные данные для обучения могут не отражать весь диапазон возможных входных данных. Кроме того, данные для обучения для конкретной задачи, как правило, не позволяют применять без доработки модель для других задач. Следует проявлять особую осторожность при разделении несбалансированных данных на обучающие и тестовые выборки в целях обеспечения сохранения одинаковых распределений между данными для обучения, данными для валидации и тестовыми данными.

Причиной смещенности при сборе данных может быть как организационный процесс сбора данных, так и предпочтения лиц или организаций, ответственных за сбор данных. Смещение меток может возникнуть при некорректном определении категорий (например, похожие изображения могут быть аннотированы разными метками, в то время как из-за изменчивости внутри класса одни и те же метки могут быть присвоены визуально различным изображениям). Для получения дополнительной информации о смещенности в системах ИИ см. *ПНСТ 839.*

5.3.3 Первичная обработка

Особое внимание следует уделить первичной обработке данных и ее влиянию на оценку эффективности, особенно в случае сравнительной оценки. В зависимости от цели оценки непоследовательная первичная обработка данных может привести к необъективной интерпретации результатов. В частности, когда в результате первичной обработки данных одна модель преобладает над другой, различия в их эффективности не следует обосновывать нижестоящими алгоритмами. Примерами первичной обработки могут служить удаление выбросов, восстановление неполных данных или фильтрация шума.

5.3.4 Данные для обучения

Особое внимание следует уделить выбору данных для обучения и валидации, а также тому, как выбор влияет на оценку эффективности, особенно в случае сравнительной оценки. В зависимости от цели оценки использование разных обучающих данных может привести к необъективной интерпретации результатов. В частности, в таких случаях любые различия в эффективности следует отнести к комбинации алгоритма и обучающих данных, а не только к самому алгоритму.

При сравнении моделей обучающие данные, используемые для построения соответствующих моделей, могут различаться. Допускается использование двух моделей, обученных на разных обучающих данных, для сравнения их друг с другом на одних и тех же тестовых данных.

5.3.5 Тестовые и валидационные данные

Данные, используемые для тестирования модели машинного обучения, должны совпадать для всех сравниваемых моделей машинного обучения. Данные тестирования и валидации не должны содержать выборки, которые пересекаются с данными для обучения.

5.3.6 Кросс-валидация

Кросс-валидация — это метод оценки эффективности метода машинного обучения с использованием одного набора данных.

Набор данных делится на k сегментов, где один сегмент используется для тестирования, а остальные для обучения. Этот процесс повторяется k раз, в каждом из которых используется новый сегмент в качестве тестового набора. Случай, когда k равно N, размеру набора данных, называется кросс-валидацией с исключением одного элемента. Когда k меньше N, это называется k-кратной кросс-валидацией.

Особый интерес представляет сравнение эффективности различных методов кросс-валидации при фиксированном наборе переменных. При этом модели, эффективность которых сопоставляется, не должны использовать разные методы кросс-валидации (например, нецелесообразно сравнивать результаты кросс-валидации модели А с результатами однократной валидации на отложенной выборке модели В).

Кросс-валидация используется главным образом для подбора оптимальных значений гиперпараметров путем сравнения их общего влияния в различных моделях. По этой причине после окончания такой проверки модель переобучают на полном наборе данных с использованием гиперпараметров, которые в среднем дают лучший результат. Однако кросс-валидация не позволяет оценить эффективность этой окончательной модели, а экстраполяция эффективности на основе результатов перекрестной проверки является грубым приближением и не дает гарантии достоверности.

Также кросс-валидация может использоваться для сравнительной оценки алгоритмов машинного обучения без последующего обучения окончательной модели. Считают, что один алгоритм превосходит по эффективности другой алгоритм, если в среднем полученные с помощью него модели показывают лучшие показатели эффективности.

5.3.7 Контроль «утечки данных»

Утечка данных происходит, когда алгоритм машинного обучения использует информацию, отсутствующую в обучающих данных, для создания модели машинного обучения.

Утечка данных часто происходит, когда обучающие данные включают информацию, недоступную на этапе эксплуатации. При оценке утечка данных может приводить к завышению доли правильных исходов модели машинного обучения в задаче классификации. Модель, обученная в таких условиях, как правило, не способна к эффективному обобщению.

Оценка должна проводиться таким образом, чтобы предотвратить утечку информации между обучающими и тестовыми данными.

Пример — Модель машинного обучения может быть разработана для классификации носителей и не носителей испанского языка с использованием нескольких аудиопримеров по каждому субъекту. Некоторые особенности, такие как произношение гласных, потенциально полезны для этого типа классификации. Однако такие признаки также могут быть использованы для идентификации конкретного говорящего. Модель может использовать данные, идентифицирующие конкретного субъекта для точной классификации тестовых данных, даже если эта информация недоступна в производственных системах. Решением здесь может стать включение примеров аудиозаписей одного и того же субъекта строго в один из наборов данных: либо в обучающий, либо в тестовый.

5.3.8 Ограничение канальных эффектов

Канальный эффект — это характеристика данных, отражающая не содержание самих данных, а способ их сбора, т.е. по какому каналу были собраны данные. Канальные эффекты могут приводить

к извлечению алгоритмами классификации нерелевантных характеристик из обучающих данных, что, в свою очередь, может привести к снижению эффективности алгоритма машинного обучения в задаче классификации.

Среди прочих факторов причиной канальных эффектов могут быть механизмы получения данных, первичная обработка данных, индивидуальные особенности человека, собирающего данные, и условия окружающей среды, в которых данные были получены.

Канальные эффекты должны быть нивелированы максимально, насколько это возможно. Управление канальными эффектами при сборе данных для обучения способствует повышению эффективности. Управление канальными эффектами в тестовых данных позволяет проводить более качественную оценку.

Примечание — Одним из методов уменьшения канального эффекта может быть сбалансированное распределение каналов для каждого класса в данных.

Отчеты должны содержать описание известных канальных эффектов, введенных в обучающие данные. Влияние канальных эффектов должно учитываться при тестировании статистической значимости (см. раздел 7).

Пример — Система машинного зрения может различать изображения кошек и собак. Однако, если все изображения «кошки» имеют высокое разрешение, а все изображения «собаки» — низкое разрешение, классификатор машинного обучения может научиться классифицировать изображения на основе разрешения, а не содержания.

5.3.9 Эталонные данные

Эталонные данные представляют собой значение целевой переменной для определенного элемента размеченных входных данных. Чистота эталонных данных может повлиять на оценку эффективности в задаче классификации. Если целевая переменная представляет собой результат экспертной разметки, то при оценке эффективности в задаче классификации необходимо гармонизировать между собой разметки отдельных экспертов.

Коллективное согласие по агрегированным эталонным данным может быть установлено количественно с помощью оценки согласия, такой как коэффициент Коэна.

В некоторых областях (например, в медицине) различия между разметчиками могут быть значительными, особенно в задачах, в которых задействовано коллективное согласие.

5.3.10 Алгоритмы, гиперпараметры и параметры в машинном обучении

Большинство алгоритмов машинного обучения имеют настроечные характеристики (известные как гиперпараметры), которые значимо влияют на процесс обучения. Алгоритмы машинного обучения используют гиперпараметры и обучающие данные для подбора оптимальных значений параметров. Способы вычисления значений этих параметров могут отличаться. Например, генеративные алгоритмы могут оптимизировать параметры таким образом, чтобы вероятность получения обучающих данных была максимальной, тогда как дискриминационные алгоритмы могут оптимизировать параметры, чтобы максимизировать долю правильных исходов классификации.

При оценивании следует указывать типы гиперпараметров для всех алгоритмов машинного обучения, а также значения гиперпараметров для каждой модели машинного обучения.

При сравнении моделей машинного обучения следует учитывать предвзятость выбора гиперпараметров. Кроме того, разные алгоритмы машинного обучения могут иметь разное количество гиперпараметров с разными возможностями настройки. Ввиду этого степень переобучения в процессе обучения может различаться в зависимости от алгоритмов машинного обучения.

Эта характеристика особенно заметна в глубоком обучении с его многочисленными комбинациями архитектур, функций активации, скорости обучения и параметров регуляризации. Информация из тестового набора не должна использоваться при настройке гиперпараметров, так как это может привести к чрезмерно оптимистичной оценке эффективности. В случае, когда для такой настройки требуется информация о метках, она обычно извлекается из отдельного набора данных, называемого валидационным набором, который не пересекается с тестовым набором.

Эта проблема может быть решена с помощью таких подходов, как вложенная кросс-валидация. В процессе обучения внешний цикл оценивает точность прогнозирования, а внутренний цикл корректирует гиперпараметры отдельных моделей. Таким образом, методы позволяют выбирать оптимальные настройки для построения прогностических моделей во внешнем цикле.

В приложении С приведена сводная информация о выбранных эталонных тестах моделей и алгоритмов машинного обучения в задаче классификации, включая параметры модели и значения, связанные с эффективностью по отношению к различным наборам данных.

5.3.11 Среда оценки

Требования к среде оценки:

- среда оценки не должна меняться в процессе оценивания;
- оборудование и программное обеспечение не подлежат замене в процессе оценивания;
- оценивание моделей машинного обучения должно проводиться в схожей тестовой среде.

Среда оценки должна отвечать минимальным требованиям, предъявляемым к среде целевой модели машинного обучения. Это повышает полезность результатов оценки для параметров эффективности, зависящих от среды, таких как время и стоимость обработки. Если невозможно воссоздать необходимые минимальные требования к среде реального применения, среда оценки может быть спроектирована таким образом, чтобы имитировать реальное применение. В таких случаях следует проанализировать потенциальное воздействие окружающей среды на результаты оценки. Например, результаты по времени и стоимости обработки могут не всегда отражать эффективность, характерную для реального применения.

5.3.12 Ускорение

Для всех моделей машинного обучения, подлежащих оценке, должна быть использована одна и та же тестовая среда. Любое ускорение в процессе обучения или тестирования должно быть зафиксировано. Каждый тестируемый метод машинного обучения должен быть оптимизирован для использования ускорения, когда оно доступно и целесообразно.

Ускорители могут представлять собой специализированное оборудование, GPU, специализированные интегральные схемы или наборы инструкций, встроенные в ЦП. Другие примеры ускорения включают разреженность, сокращение архитектуры модели и другие виды оптимизации, направленные на уменьшение требуемой памяти. Ускорители можно применять как к простой функции (например, к общему умножению матриц), так и к сложной функции (например, к полной функции ResNET).

5.3.13 Релевантные модели базового уровня

Модель базового уровня может быть необходима при сравнении эффективности моделей и алгоритмов машинного обучения в задаче классификации. В качестве моделей базового уровня часто выбирают т. н. «наивные» классификаторы, например те, которые всегда прогнозируют наиболее частотный класс. Сравнение с наивными классификаторами полезно учитывать для интерпретации показателей, но они, конечно, не должны являться единственной точкой сравнения.

5.3.14 Контекст эффективности моделей и алгоритмов машинного обучения в задаче классификации

При оценке эффективности моделей и алгоритмов машинного обучения важно учитывать систему в целом (включая компоненты и подсистемы), в которой будет развернута модель машинного обучения. Кроме того, контекст, такой как переменные среды, может определять ограничения эффективности моделей и алгоритмов машинного обучения в задаче классификации. Чтобы сделать окончательный вывод об эффективности модели машинного обучения, может потребоваться оценка с использованием нескольких контекстных наборов данных.

6 Статические показатели эффективности

6.1 Общие положения

Задачи классификации в машинном обучении можно разделить на бинарную, мультиклассовую и классификацию с несколькими метками. Оценка эффективности для каждой категории приведена в настоящем разделе.

Классы обычно представлены данными из дискретного и неупорядоченного набора, так что задача не может быть формализована как задача регрессии. Например, медицинский диагноз по набору симптомов (инсульт, передозировка наркотиками, судороги) не имеет порядка значений классов и непрерывного перехода от одного класса к другому.

6.2 Базовые элементы для расчета показателей

6.2.1 Общие положения

Некоторые показатели классификации рассчитывают с помощью следующих переменных:

- количество истинно положительных образцов T_P;
- количество истинно отрицательных образцов T_N;
- количество ложноположительных образцов F_P;
- количество ложноотрицательных образцов T_N.

6.2.2 Матрица ошибок

В большинстве случаев матрица ошибок представляет собой истинные классы в столбцах и прогнозируемые классы в строках. Несмотря на то, что матрицы ошибок используют для расчета нескольких широко используемых показателей эффективности классификации, их также можно использовать для расчета специализированных показателей. См. приложение D, в котором приведен пример одного из таких показателей, доля причинно-специфической смертности с поправкой на случайность (cause-specific mortality fraction, CSMF).

6.2.3 Доля правильных исходов

Доля правильных исходов не может быть использована для выражения сравнительной эффективности моделей, за исключением случаев достаточной сбалансированности классов.

6.2.4 Точность, полнота и специфичность

С увеличением точности оценки результата фиксируется больше истинно положительных значений, а ложноотрицательные значения не учитываются. Точность класса рассчитывают по формуле

$$p = \frac{T_P}{T_P + F_P}. \quad (7)$$

С увеличением полноты фиксируется больше истинно положительных значений, а ложноположительные значения не учитываются. Полноту класса рассчитывают по формуле

$$r = \frac{T_P}{T_P + F_N}. \quad (8)$$

С увеличением специфичности фиксируется больше истинно отрицательных значений, а ложноотрицательные значения не учитываются. Специфичность класса рассчитывают по формуле

$$s = \frac{T_N}{T_N + F_P}. \quad (9)$$

6.2.5 F_1-мера

F_1-меру рассчитывают по формуле

$$F_1 = \frac{2T_P}{2T_P + F_P + F_N}. \quad (10)$$

6.2.6 F_β

Точность и полнота не всегда одинаково важны в приложениях систем ИИ, так как как ложноположительные значения могут иметь значительно большую стоимость по сравнению с ложноотрицательными, и наоборот. Такие случаи можно учесть, обобщив показатель F_1 до показателя F_β.

Значения β больше 1 указывают на преобладающую важность полноты над точностью, что требует минимизации ложноотрицательных значений. Значения β меньше 1 указывают на преобладание важности точности над полнотой, что требует минимизации ложноположительных значений. F_β рассчитывают по формуле

$$F_\beta = \left(1+\beta^2\right)\frac{p \cdot r}{r + \beta^2 \cdot p}, \quad (11)$$

где p — точность прогнозирования класса;

β — параметр преобладания полноты над точностью;

r — полнота прогнозирования класса.

Веса для точности и полноты могут прямо применяться с помощью пар значений (α, β) в соответствии с формулой

$$F(\alpha p, \beta r) = \frac{(\alpha + \beta) p \cdot r}{\alpha r + \beta p}. \quad (12)$$

6.2.7 Дивергенция Кульбака-Лейблера

Дивергенция Кульбака-Лейблера, D_{KL} — мера количественной оценки различий между эталонным и эмпирическим распределениями. Хотя она часто используется как функция потерь, она также используется в качестве показателя для оценки набора данных. В таких случаях рассматриваемое распределение касается встречаемости каждой метки в наборе данных.

Общая формула для всех классов:

$$D_{KL}(p,q) = \sum p(x)\log\left(\frac{p(x)}{q(x)}\right), \quad (13)$$

где $p(x)$ — истинное распределение меток классов;

$q(x)$ — оценочное распределение по классификатору.

6.3 Бинарная классификация

6.3.1 Общие положения

В бинарной классификации каждый пример помечается как один из двух взаимоисключающих классов. Эти классы часто являются «положительными» или «отрицательными» по отношению к категоризации. Два класса обычно не упорядочены.

Пример — Программное обеспечение, применяемое для моделей и алгоритмов машинного обучения в задаче классификации, учится помечать электронную почту как «спам» или «не спам» на основе ввода данных получателем электронной почты.

Одноклассовая классификация (также известная как унарная) похожа на бинарную классификацию. В унарной классификации существует единственный целевой класс, а также класс-выброс, состоящий из всего, что не входит в целевой класс. В отличие от бинарной классификации этот класс выброса не моделируется. Одноклассовые классификаторы обычно используются, когда обучающие данные сильно не сбалансированы, например, когда обучающие данные существуют только для одного целевого класса. Таким образом, данные одного класса используются для построения модели.

6.3.2 Матрица ошибок для бинарной классификации

В бинарной классификации один класс является положительным, а другой отрицательным, поэтому расчет матрицы ошибок эквивалентен вычислению истинно положительных и ложноположительных и отрицательных результатов. В таблице 1 представлена форма матрицы ошибок для бинарного классификатора.

Таблица 1 — Элементы матрицы ошибок для бинарной классификации

		Истинные классы	
		положительные	отрицательные
Прогнозируемые классы	положительные	истинно положительные T_P	ложноположительные F_P
	отрицательные	ложноотрицательные F_N	истинно отрицательные T_N

6.3.3 Доля правильных исходов для бинарной классификации

Для бинарной классификации долю правильных исходов вычисляют по формуле

$$a = \frac{(T_P + T_N)}{(T_P + F_P + T_N + F_N)}. \quad (14)$$

6.3.4 Точность, полнота, специфичность, F_1-мера и F_β для бинарной классификации

Для бинарной классификации точность, полнота, F_1-мера и F_β используются для вычисления показателей положительного класса.

6.3.5 Дивергенция Кульбака-Лейблера для бинарной классификации

Для бинарной классификации дивергенцию Кульбака-Лейблера рассчитывают по формуле

$$D_{KL} = \frac{(T_P + F_N)\cdot\log\frac{(T_P + F_N)}{(T_P + F_P)} + (T_N + F_P)\cdot\log\frac{(T_N + F_P)}{(T_N + F_N)}}{N}, \quad (15)$$

где N — количество образцов.

6.3.6 ROC-кривая и площадь под кривой

ROC-кривая — это графический метод отображения истинно положительных и ложноположительных значений по нескольким пороговым значениям из бинарного классификатора.

Для выражения эффективности по всем пороговым значениям необходимо рассчитать AUROC. Более высокий показатель AUROC указывает на более надежную работу в диапазоне от 0 (наихудший) до 1 (наилучший). Классификаторы, которые работают не лучше случайного, будут иметь AUROC 0.5.

AUROC подходит для ранжированных прогнозов, а также для случаев, когда частота ложноположительных и истинно положительных результатов примерно одинакова. AUROC не подходит для несбалансированных данных, т.к. не учитывает пропорцию ложных и истинных срабатываний.

В приложении В приведена ROC-кривая, полученная на основе выходных данных бинарной классификации.

6.3.7 Кривая точности полноты и площадь под кривой точности полноты

Кривая точности полноты PRC — это графический метод отображения полноты и точности по нескольким пороговым значениям. График отображает полноту по оси абсцисс в зависимости от точности по оси ординат.

Для выражения эффективности по всем пороговым значениям необходимо рассчитать площадь под кривой PR (AUPRC). Более высокий AUPRC свидетельствует о более надежной работе. Чем ближе PRC к правому верхнему углу, тем лучше показатели классификатора.

AUPRC информативен, когда положительный класс относительно важнее, а также в случае несбалансированных данных.

6.3.8 CRC

CRC, также называемая кривой прироста (gain curve) или графиком прироста (gain chart), представляет собой графический метод отображения истинно положительных значений и доли положительных прогнозов от общих данных по нескольким пороговым значениям.

Для выражения эффективности по всем пороговым значениям необходимо рассчитать площадь под CRC. Более высокие значения указывают на более устойчивую производительность в диапазоне от 0 (наихудший) до 1 (наилучший). Классификаторы, которые работают не лучше случайного, будут иметь значение 0,5.

CRC является альтернативой ROC. CRC позволяет получить результаты для определенной целевой части набора данных.

6.3.9 Lift-кривая

Lift-кривая это графический метод отображения отношения истинной положительной доли между моделью и случайным классификатором по оси ординат, и процента положительных прогнозов от общих данных по нескольким пороговым значениям по оси абсцисс. Таким образом, lift-кривая напрямую связана с gain-кривой.

Lift-кривая отражает преимущество, обеспечиваемое классификатором, по сравнению со случайным угадыванием для различных процентных долей выборки, отсортированной по убыванию прогнозной вероятности целевого класса.

6.4 Мультиклассовая классификация

6.4.1 Общие положения

В мультиклассовой классификации каждый образец помечен как один из трех или более взаимоисключающих классов.

Пример — Программное обеспечение, используемое для моделей и алгоритмов машинного обучения в задаче классификации, учится классифицировать изображения по классам «собака», «кошка» или «другое» на основе меток, присвоенных человеком.

6.4.2 Доля правильных исходов для мультиклассовой классификации

Доля правильных исходов является характерным оценочным показателем для мультиклассового классификатора. В этом случае она соответствует сумме верных срабатываний для каждого класса, деленной на общее количество элементов во всех классах. Для каждого класса доля правильных исходов может быть вычислена и в условиях нескольких классов, что эквивалентно полноте определения этого класса.

6.4.3 Макроусредненное, средневзвешенное и микроусредненное

Некоторые показатели мультиклассовой классификации основаны на усреднении показателей, рассчитанных для каждого класса: точность, полнота, специфичность и F_1-мера. Эффективность нескольких классов может быть выражена через несколько макроусредненных, средневзвешенных и микроусредненных значений. Для выбора соответствующего многоклассового подхода к характеристикам должна быть представлена основа.

Двухклассовые концепции положительного и отрицательного значений можно обобщить для многоклассовой задачи, рассматривая выборки в целевом классе как положительные, а выборки во всех других классах как отрицательные. Например, ложные срабатывания по отношению к классу *i* — это образцы, принадлежащие другим классам, которые ошибочно отнесены к классу *i*.

Для иллюстрации F_1 представим макроусредненное F_1, средневзвешенное F_1 и микроусредненное F_1 значения следующим образом:

- макроусредненное значение F_1 — это среднее значение F_1 для каждого класса без учета числа наблюдений в каждом классе;
- средневзвешенное значение F_1 указывает на дисбаланс классов, где взвешенное значение F_1-меры для каждого класса позволяет получить данные о размере каждого класса. Вычисления производятся путем умножения F_1-меры для каждого класса на количество наблюдений данного класса, а затем деления на общее количество наблюдений;
- микроусредненное значение F_1 объединяет точность и полноту по всем классам. Вычисление производится путем сложения истинно положительных и ложноположительных и отрицательных значений для всех классов с последующим вычислением F_1 из этой суммы.

Три подхода используются для трех разных сценариев. Макроусредненное F_1 (M_{AC,F_1}) подходит для проверки эффективности в каждом классе независимо от встречаемости классов.

Микроусредненное F_1 (M_{IC,F_1}) подходит для проверки, насколько хорошо модель прогнозирует для заданного наблюдения независимо от его класса.

Средневзвешенное F_1 (W_{TD,F_1}) фокусируется на эффективности определения наиболее встречаемого класса, что, как правило, усиливает дисбаланс по сравнению с микроусредненным значением.

Формулы для эффективности моделей и алгоритмов машинного обучения в задаче многоклассовой классификации приведены ниже.

Допустим, T_{Pi} — количество наблюдений, верно классифицированных как класс i.

Допустим, F_{Pi} — количество наблюдений, неверно классифицированных как класс i.

Допустим, F_{Ni} — количество наблюдений из класса i, неверно классифицированных как другой класс.

Допустим, L — общее количество классов.

Допустим, N — общее количество наблюдений во всех классах.

$$M_{AC,F_1} = \frac{1}{L}\sum_{i=1}^{L}\frac{2\cdot T_{Pi}}{2\cdot T_{Pi} + F_{Pi} + F_{Ni}}, \tag{16}$$

$$W_{TD,F_1} = \sum_{i=1}^{L}\frac{T_{Pi} + F_{Ni}}{N}\cdot\frac{2\cdot T_{Pi}}{2\cdot T_{Pi} + F_{Pi} + F_{Ni}}, \tag{17}$$

$$M_{IC,F_1} = \frac{\sum_{i=1}^{L} 2\cdot T_{Pi}}{\sum_{i=1}^{L} 2\cdot T_{Pi} + F_{Pi} + F_{Ni}}. \tag{18}$$

Тот же подход можно использовать для расчета микроусредненных, макроусредненных и средневзвешенных результатов для точности, полноты и специфичности. Однако для этих показателей наиболее типичным подходом является макроусреднение, поскольку большинство других вариантов эквивалентны другим показателям.

В приложении А приведен пример эффективности для задачи мультиклассовой классификации, а также переход от матрицы ошибок для бинарной классификации к мультиклассовой классификации.

6.4.4 Показатели разницы распределений или расстояния между распределениями

Другой способ оценки эффективности моделей и алгоритмов машинного обучения в задаче мультиклассовой классификации учитывает разницу в распределении классов между размеченными данными и прогнозируемыми метками, используя следующие допущения:

- пусть T — общее количество наблюдений;
- T_i — количество наблюдений для каждого класса;
- P_i — прогнозируемое количество наблюдений для каждого класса.

Тогда дискретные распределения представляют собой $T_d = (T_1, T_2, ..., T_n)/T$ и $P_d = (P_1, P_2, ..., P_n)/T$ и может быть вычислена разница распределений.

Дивергенцию Кульбака-Лейблера, показатель разницы распределений, рассчитывают по формуле

$$D_{KL} = \Sigma(P_i)\cdot\ln\left(P_i/T_i\right). \tag{19}$$

Примечание — $0 \cdot \log(0/x) = 0$.

6.5 Классификация с несколькими метками

6.5.1 Общие положения

В задаче классификации с несколькими метками объект может быть отнесен к одному или более классам. При этом классы не исключают друг друга, а объекту может быть присвоено несколько меток одновременно. Более того, классы могут быть скоррелированными.

Оценка эффективности моделей или алгоритмов машинного обучения в задаче классификации с несколькими метками затрудняется тем, что объекту может быть присвоено несколько меток одновременно. Модель машинного обучения может предсказать подмножество всех правильных меток для выбранного объекта. Однако также возможен вариант, когда модель может правильно предсказать лишь некоторое количество меток, но неправильно предсказать другие метки для выбранного объекта.

Пример — Программное обучение, используемое для моделей и алгоритмов машинного обучения в задаче классификации с несколькими метками, обучается классифицировать текст на одну или более категории: «комментарий», «новость», «критикующий», «одобрительный», «дезинформация» на основании меток, сделанных человеком. Одному тексту может быть присвоено несколько меток: «критикующий», «комментарий», «дезинформация». Другому тексту может быть присвоена только одна метка «новость».

Существуют различные показатели оценки эффективности моделей и алгоритмов машинного обучения в задаче классификации с несколькими метками, такие как функция потерь Хемминга, коэффициент точного совпадения и индекс Жаккара. Некоторые показатели бинарной и мультиклассовой классификации также применимы, зачастую после адаптации. Коэффициент точного совпадения — пример адаптации доли правильных исходов, применяемой для бинарной классификации.

Критерии выбора показателей оценки эффективности в задаче классификации с несколькими метками должны быть указаны в отчете.

6.5.2 Функция потерь Хемминга

Функция потерь Хемминга представляет собой количество неправильно предсказанных меток, разделенное на общее количество меток. Низкие значения указывают на более высокую эффективность. Поскольку функция потерь Хемминга не различает между положительными и отрицательными ошибками, рекомендуется использовать ее в качестве общего показателя эффективности.

Функция потерь рассчитывается следующим образом.

Допустим, что N, L обозначают общее количество объектов и меток, присутствующих в наборе данных.

Допустим, что $\hat{I}_{i.j} = \hat{I}_{i.1}, \hat{I}_{i.2}, \ldots, \hat{I}_{i.L}$ обозначает значения предсказанных меток для $x_i \cdot \hat{I}_{i.j} = 1$, если метка j присутствует в ряде предсказанных меток для x_i, в ином случае $\hat{I}_{i.j} = 0$.

Аналогичным образом допустим, что $I_{i,j} = I_{i,1}, I_{i,2}, \ldots, I_{i,L}$ обозначает эталонные значения меток для $x_i \cdot I_{i,j} = 1$ если метка j присутствует в ряде эталонных меток для x_i, в ином случае $I_{i,j} = 0$.

Функция потерь Хемминга для объекта данных x_i может быть записана следующим образом:

$$H_L(x_i) = \frac{1}{L}\sum_{j=1}^{L} I\left(\hat{I}_{i.j} \neq I_{i,j}\right). \quad (20)$$

Значение функции общих потерь Хемминга представляет собой среднее значение функции потери Хемминга по всем объектам данных

$$H_L = \frac{1}{NL}\sum_{i=1}^{N}\sum_{j=1}^{L} I\left(\hat{I}_{i.j} \neq I_{i,j}\right). \quad (21)$$

6.5.3 Коэффициент точного совпадения

Коэффициент точного совпадения, или доля правильных исходов на уровне подмножеств, представляет собой процент объектов, для которых в точности предсказаны все метки. При выборке с правильными метками A, B и C коэффициент точного совпадения воспринимает предсказание с метками A и B (но не C) как ошибочное. Данный метод воспринимает частично верные предсказания как ошибочные. Данный упрощенный метод может быть использован для оценки первого порядка в случае высокой несбалансированности данных.

Коэффициент точного совпадения рассчитывается следующим образом.

Допустим, что N, L обозначают общее количество объектов и меток, присутствующих в наборе данных.

Допустим, что I обозначает индикаторную функцию.

Допустим, что $\hat{I}_i$ и I_i обозначают набор значений предсказанных и эталонных меток для объекта данных x_i соответственно.

Коэффициент точного совпадения может быть выражен следующим образом:

$$E_{MR} = \frac{1}{N}\sum_{i=1}^{N} I\left(\hat{I}_i = I_i\right). \quad (22)$$

6.5.4 Индекс Жаккара

Другим часто используемым показателем для оценки эффективности моделей и алгоритмов машинного обучения в задаче классификации с несколькими метками является индекс Жаккара, обычно называемый пересечением над объединением (Intersection over Union, I_{oU}).

I_{oU} рассчитывается следующим образом.

Допустим, что P обозначает наборы предсказанных меток.

Допустим, что T обозначает наборы истинных меток.

$$I_{oU} = \frac{|T \cap P|}{|T \cup P|}. \quad (23)$$

Данный показатель может применяться на уровне наборов данных (для агрегирования всех меток из всех объектов) или на уровне объектов (для вычисления и получения средних данных I_{oU}, относящихся к объекту).

6.5.5 Показатели разницы распределений или расстояния между распределениями

Подобно показателю разницы распределений для оценки эффективности моделей и алгоритмов машинного обучения в задаче мультиклассовой классификации при оценке эффективности моделей и алгоритмов в задаче классификации с несколькими метками учитывают разницу в распределении классов между размеченными данными и прогнозируемыми метками, используя следующие допущения:

- пусть T_i — количество объектов для каждого класса. Объекты могут принадлежать нескольким классам одновременно;
- T_t — общее количество меток, полученных всеми объектами;
- P_i — количество объектов, предсказанных для каждого класса;
- T_p — общее количество меток для всех предсказаний.

Тогда дискретные распределения представляют собой $T_d = (T_1, T_2, ..., T_n)/T_t$, $P_d = (P_1, P_2, ..., P_n)/T_p$, и может быть вычислена разница распределений.

Дивергенцию Кульбака-Лейблера, показатель разницы распределений, рассчитывают по формуле

$$D_{KL} = \sum(P_i) \cdot \ln\left(P_i / T_i\right). \quad (24)$$

Примечание — $0 \cdot \log(0/x) = 0$.

6.6 Вычислительная сложность

6.6.1 Общие положения

Дополнительные атрибуты эффективности моделей и алгоритмов машинного обучения в задаче классификации могут быть приняты во внимание в связи с внедрением и использованием систем машинного обучения для решения задачи классификации. В дополнение к доле правильных исходов, полученной из матриц ошибок, могут быть использованы следующие показатели: скорость работы, производительность, эффективность и энергопотребление. Один или несколько этих элементов могут выступать в роли дополнительных ограничений при совместной оптимизации по ним при разработке или оценке системы.

6.6.2 Скорость работы

В случаях с интерактивными, ориентированными на пользователя системами может возникнуть необходимость оценки скорости работы моделей и алгоритмов в задаче классификации. Время ожидания от ввода информации пользователем до вывода результата моделями и алгоритмами машинного обучения может быть определяющим фактором в оценке качества предоставляемого сервиса. Время ожидания должно попадать в границы, которые определяются требованиями к эксплуатации.

Пример — Оптимизация гиперпараметров может быть использована для улучшения доли правильных исходов модели, возможно с ее потенциальным усложнением. В некоторых случаях, получившаяся оптимизированная модель, которая не соответствует ожиданиям по скорости расчета, считается неприемлемой.

Скорость работы моделей и алгоритмов машинного обучения в задаче классификации определяется такими факторами, как скорость сохранения признаков, объем данных, который влияет на производительность модели, сложность модели и некорректная настройка вычислительных ресурсов.

Для удовлетворения требований к скорости работы моделей и алгоритмов машинного обучения в задаче классификации могут использоваться специализированные аппаратные средства и ускорители, тем самым увеличивая объем требуемых настроек. Задержки в очереди задач на обработку моделями или алгоритмами в неоднородном бизнес-процессе могут возникать из-за волатильности входящего потока.

Скорость работы моделей и алгоритмов машинного обучения в задаче классификации может быть выражена следующим образом

$$\frac{1}{N}\sum_{i=1}^{N}\left(T_{mr} - T_{di}\right), \tag{25}$$

где N — общее количество объектов в заданном наборе данных;

T_{mr} — время, за которое модель выдает прогноз для i-го объекта;

T_{di} — время, за которое i-й объект вводится в процесс применения модели.

6.6.3 Производительность

Производительность моделей и алгоритмов машинного обучения в задаче классификации представляет собой количество выводов за единицу времени, вычисляемых моделью или алгоритмом машинного обучения с учетом требований по скорости работы.

Производительность моделей и алгоритмов машинного обучения в задаче классификации может быть выражена следующим образом:

$$\frac{N_{cla}}{T_e - T_b}, \tag{26}$$

где N_{cla} — общее количество объектов, для которых модель или алгоритм выдает прогноз в заданный период времени с учетом требований по скорости обработки;

T_e — время, когда заканчивается вывод для объектов N_{cla};

T_b — время, когда начинаются объекты N_{cla}.

Примечание — Скорость работы может быть повышена, если это заложено разработчиком системы. К примеру, производительность устройства ИИ или системы ИИ может быть улучшена за счет выполнения классификации сгруппированным или пакетным способом обработки (так называемый батч-режим). Данный подход обеспечивает более эффективное использование вычислительных возможностей системы за счет повторного использования весов при выводе, особенно на многопоточных и больших вычислительных устройствах. Однако сгруппированные или пакетные операции могут привести к увеличению времени обработки для каждого отдельного объекта.

6.6.4 Эффективность

Эффективность моделей и алгоритмов машинного обучения в задаче классификации представляет собой степень экономного выполнения поставленной задачи. Целью является функция атрибутов доли верных исходов, как описано в 6.2—6.5.

Эффективность моделей и алгоритмов машинного обучения в задаче классификации является важным показателем прогресса в машинном обучении. Эффективным классификаторам требуется меньшее количество вычислений для обучения закономерностям в данных по сравнению с неэффективными классификаторами.

Пример — Объем вычислений, необходимый для обучения нейронных сетей до достижения определенного уровня эффективности в классификации ImageNet, становился меньше в два раза каждые 16 месяцев с 2012 года.

6.6.5 Энергопотребление

Важно учитывать, насколько производительность и применение системы ИИ будут ограничены пределами энергопотребления (такими, как производительность на ватт). Этот вопрос особенно важен, когда система ИИ широко распространена или активно используется.

Для многих сценариев использования удовлетворительная производительность при меньших энергетических затратах может быть желательнее или практичнее более высокой производительности. Во многих случаях это позволяет найти компромисс между потребляемой мощностью и производительностью.

Данное соотношение производительности на потребляемую мощность (ватт) может быть рассчитано путем измерения общей производительности P для заданного интервала T. Средняя производительность может быть представлена как P/T. Аналогично средняя потребляемая мощность для данного интервала T может быть представлена как E/T, где E — энергия, измеряемая в джоулях.

Таким образом, соотношение производительности на ватт может быть представлено как $(P/T)/(E/T) = P/E$.

Например, в сценарии использования с визуальным выводом, если производительность измеряется в количестве кадров, классифицируемых в секунду F/t, то производительность на ватт можно представить как соотношение кадров на джоули F_{PJ}. А соотношение джоулей на кадр J_{PF} может быть выражено следующим образом:

$$J_{\mathrm{PF}} = \frac{\int_{t=0}^{t=T} P(t)\,dt}{F}, \tag{27}$$

где F — общее количество выведенных кадров;

T — общее время, необходимое для анализа F кадров;

$P(t)$ — производительность в момент времени t.

В других случаях желательно проанализировать J_{PF} для верно классифицированных кадров.

Более обобщенная форма для различных случаев использования может быть представлена следующим образом:

$$J_{PI} = \frac{\int_{t=0}^{t=T} P(t)\,dt}{I}, \tag{28}$$

где J_{PI} — количество джоулей энергии, затраченных на вывод;

I — количество верно классифицированных выводов.

Примечание — Высокое энергопотребление повышает затраты на систему ИИ.

Пример — Тщательный учет энергопотребления может потребоваться при использовании в центрах обработки данных, поскольку внедрение решения ИИ ограничено доступными системами охлаждения и энергоснабжения.

Несмотря на то, что эти аспекты энергопотребления возникают на этапе применения модели, энергопотребление на этапе обучения также имеет значение.

7 Тесты статистической значимости

7.1 Общие положения

Различия между долей правильных исходов и другими показателями эффективности моделей и алгоритмов машинного обучения в задаче классификации могут зависеть от конкретных данных, на которых оценивались эти показатели. По этой причине тестовые данные должны быть проанализированы перед оценкой эффективности моделей и алгоритмов машинного обучения в задаче классификации. Этот анализ должен включать в себя количество доступных для оценки объектов и их распределение.

Этот вопрос представляет особую важность при проведении оценок в тех случаях, когда при тестировании используется относительно небольшой набор данных. Также он является актуальным в случаях, когда методология оценки основана на таких подходах, как кросс-валидация, которая использует несколько перестановок набора данных для обучения и тестирования модели. Рассматриваемые в настоящем разделе методы направлены на проверку того, могут ли различия в эффективности моделей быть случайностью.

Научное сообщество предлагает различные подходы к решению данного вопроса (например, в отношении обработки естественного языка). Тесты статистической значимости, рассматриваемые в настоящем разделе, являются одними из основных методов, применяемых на практике. Тем не менее

существует большое количество иных методов проверки, вследствие чего вопрос о том, какие методы являются оптимальными для тех или иных сценариев и областей применения, остается открытым.

Применение тестов на статистическую значимость должно быть отражено в результатах оценки эффективности моделей и алгоритмов машинного обучения в задаче классификации. Случаи, когда тесты на статистическую значимость не проводились и анализ не применялся, также должны быть отражены в отчете о валидации.

7.2 Метод проверки по парному *t*-критерию Стьюдента

Парный *t*-критерий Стьюдента сравнивает средние значения и стандартные отклонения двух групп с целью определения степени различий между группами. Парный *t*-критерий Стьюдента предполагает нормальное распределение; данный метод неприменим при оценке трех и более групп.

В случае моделей и алгоритмов машинного обучения для решения задачи классификации тест отвечает на вопрос о статистической значимости различий в значениях доли правильных исходов между моделями или алгоритмами.

Хотя парный *t*-критерий Стьюдента широко используется на практике, его не следует применять в оценках, основанных на перекрестных (*k-fold*) методах. Повторная выборка данных по нескольким обучающим наборам означает, что значения больше не являются независимыми, что противоречит основному предположению данного теста.

Более высокую робастность обеспечивает кросс-валидация 5x2, введенная Т. Диттерихом. Данный метод проверки был разработан для устранения проблемы зависимости от выборки и использует пять запусков двухкратной кросс-валидации. В каждом запуске данные делятся на тренировочные и тестовые. Модели обучаются и проверяются на каждом наборе, а производительность рассчитывается для каждой перестановки.

7.3 Дисперсионный анализ

При сравнении более двух групп можно использовать ANOVA, чтобы определить, равны ли средние значения более чем двух групп (т.е. являются ли различия в значениях долей правильных исходов между тремя и более моделями статистически значимыми). Дисперсионный анализ предполагает нормальное распределение и однородность дисперсии.

Дисперсионный анализ основан на межгрупповых и внутригрупповых среднеквадратичных значениях. Сумма квадратов разницы между группами выражает отклонение средних значений группы от общих средних значений. Сумма квадратов внутри группы основана на квадрате суммы значений, отцентрированных на среднем значении каждой группы. Это выражает разброс измерений внутри каждой группы.

Тестовой статистикой дисперсионного анализа служит F-статистика, которая рассчитывается как соотношение межгрупповых и внутригрупповых среднеквадратичных значений.

7.4 Метод проверки по критерию Краскела-Уоллиса

Критерий Краскела-Уоллиса представляет собой непараметрический, ранговый метод, который проверяет, происходят ли объекты из одного и того же распределения. Данный метод может быть использован для сравнения показателей трех и более независимых групп. Если значение критерия Краскела-Уоллиса выше критического значения распределения хи-квадрат, это означает, что эти группы имеют разное распределение.

7.5 Метод проверки по хи-квадрат

Метод проверки по хи-квадрат представляет собой метод определения соответствия наблюдаемых и ожидаемых частот для независимых категориальных переменных. Для метода проверки по хи-квадрат используют таблицу сопряженности для определения связи между двумя переменными, в результате чего появляется тестовая статистика с распределением хи-квадрат. Высокие значения хи-квадрат указывают на слабое совпадение наблюдаемых и ожидаемых частот.

7.6 Метод проверки по критерию знаковых рангов Уилкоксона

Метод проверки по критерию знаковых рангов Уилкоксона представляет собой непараметрическую альтернативу методу проверки по парному *t*-критерию Стьюдента, применяемую к ранжированным данным. Данный метод ранжирует различия в производительности двух классификаторов по положительным и отрицательным различиям для каждого набора данных. Цель данного метода — определить, будет ли случайно выбранное наблюдение больше или меньше, чем объект в распределении другого набора данных. Ранжированные значения обоих наборов перемежаются, чтобы выявить наличие кластеров на противоположных концах, что означает, что результаты не проходят проверку на значимость.

7.7 Точный тест Фишера

Точный тест Фишера представляет собой тест статистической значимости, используемый в анализе таблиц сопряженности, матриц, показывающих частотное распределение переменных. Он применим при анализе двух номинальных переменных (категориальных переменных без признаков ранга или порядка). Точный тест Фишера позволяет определить различия между пропорциями номинальных переменных, в частности в тестах с небольшим размером выборки.

7.8 ЦПТ

Согласно ЦПТ с увеличением размера выборки средние значения всех выборок принимают форму нормального распределения вокруг среднего значения популяции. Каждая выборка определяется как группа наблюдений, взятых из одного и того же базового распределения популяции, которая является частью более крупной поисковой популяции. Популяция определяется как пространство поиска со всеми возможными наблюдениями, полученными во время испытания. Инструменты на основе ЦПТ могут быть использованы для оценки вероятности того, что две выборки с разными значениями долей правильных исходов были взяты из популяции с разными исходными распределениями.

Путем оценки распределений ошибок для заданной модели или алгоритма метод ЦПТ также определяет, может ли ошибка классификации быть отнесена к шуму или отсутствию критического признака. Другими словами, если распределение ошибок для данной модели является ненормальным или асимметричным, это может быть связано с ошибочной моделью или отсутствующим признаком. Кроме того, независимые расчеты доли правильных исходов модели на нескольких выборках приблизительно соответствуют распределению навыков модели вокруг общей средней доли правильных исходов для данной задачи.

7.9 Тест Мак-Немара

Тест Мак-Немара представляет собой непараметрический тест, применяемый к парным номинальным данным, представленным в таблицах сопряженности. Он подходит для использования в случаях, когда обучение невозможно выполнить несколько раз, что исключает такие методы, как кросс-валидация.

Тест Мак-Немара предназначен для ответа на вопрос об однородности таблицы сопряженности. Для этого анализируются случаи, когда результаты моделей в задаче классификации различаются, например агрегированные результаты в случаях, когда модель машинного обучения А классифицирует объект как «собака», а модель Б классифицирует тот же объект как «не собака». Тест определяет различия между относительными долями ошибок моделей. Разновидности теста Мак-Немара можно использовать при малых значениях в таблице сопряженности.

7.10 Проблема множественных сравнений

7.10.1 Общие положения

Проблема множественных сравнений возникает, когда необходимо проверить ряд статистических гипотез одновременно. Например, если статистический тест выполняется на уровне α = 5 % и соответствующая нулевая гипотеза верна, существует только 5 %-ная вероятность ошибочного отклонения нулевой гипотезы. Тем не менее, если при проведении тестов все соответствующие нулевые гипотезы верны, ожидаемое количество ошибочных отклонений составляет 0,05 · n. Если тесты статистически независимы друг от друга, вероятность хотя бы одного ошибочного отклонения возрастает с увеличением числа тестов. Таким образом, коэффициент ошибок с поправкой на эффект множественных сравнений составляет

$$1 - (1 - \alpha)^m, \tag{29}$$

где m — количество тестов.

Проблема множественных сравнений возникает при сравнении классификаторов в случаях, когда один классификатор сравнивается в индивидуальном порядке с несколькими другими классификаторами или при сравнении нескольких гиперпараметризаций этих классификаторов.

7.10.2 Поправка Бонферрони

Поправка Бонферрони заключается в следующем.

Допустим, что $H_1,\ldots,H_m$ представляют собой семейство m нулевых гипотез, а $p_1,\ldots,p_m$ — соответствующие им p-значения.

p-значения располагают в порядке увеличения значений, $p'_1,\ldots,p'_m$, где соответствующие им гипотезы — это $H'_1, \ldots, H'_m$.

С учетом эмпирического уровня значимости α, k представляет собой минимальный показатель, при котором

$$p'_k > \frac{\alpha}{m+1-k}. \quad (30)$$

Необходимо отклонить нулевую гипотезу $H'_1,\ldots,H'_{k-1}$ и не отклонять другие.

7.10.3 FDR

FDR представляет собой метод корректировки p-значений каждого теста в многофакторном сравнительном исследовании. FDR описывает частоту, с которой данный тестовый набор неверно определяет результат теста значимости. FDR менее консервативен, чем поправка Бонферрони, и обладает более высокой способностью определения действительно значимых результатов.

8 Подготовка отчетов

В отчете должны быть представлены следующие сведения:

- источник, размер и состав данных для обучения;
- источник, размер и состав данных для тестирования;
- меры, принятые для анализа, учета и уменьшения смещения в данных для тестирования и обучения;
- методы, с помощью которых определяются истинные значения целевой переменной (метки класса) в тестовых и обучающих данных;
- надежность значений целевой переменной в тестовых и тренировочных данных и ее потенциальное влияние на статистическую значимость;
- количество истинных и ложноположительных случаев, правильно и неправильно классифицированных в разрезе репрезентативных подвыборок;
- тестовая среда, включающая аппаратные средства (CPU/GPU или другая архитектура обработки) и программное обеспечение (операционная система), используемые для генерации выводов, с указанием конкретных версий и поколений;
- длительность применения модели или другие показатели вычислительной эффективности.

Приложение А
(справочное)

Пример эффективности для задачи мультиклассовой классификации

А.1 Переход от необработанных результатов классификации к мультиклассовым результатам

Таблицы А.1—А.4 показывают переход от необработанных результатов классификации к результатам по конкретным классам и мультиклассовым результатам с использованием подходов, описанных в 6.4.2 и 6.4.3. В таблице А.1 показана матрица ошибок для классов A, B и C. Истинно положительные результаты отображены по диагонали и выделены жирным шрифтом.

Таблица А.1 — Матрица ошибок для мультиклассовой классификации с классами A, B, C

		Фактические		
		A	B	C
Предсказанные	A	**400**	150	14
	B	23	**3800**	144
	C	13	355	**65**

В таблице А.2 приведены условные значения истинно положительных, истинно отрицательных, ложноположительных и ложноотрицательных результатов для каждого класса. На основе этих показателей можно рассчитать точность, полноту и другие показатели.

Таблица А.2 — Условные истинно положительные, истинно отрицательные, ложноположительные и ложноотрицательные результаты для классов A, B, C

	A	B	C
T_P	400	3800	65
T_N	4364	492	4373
F_p	164	167	368
F_N	36	505	158

В таблице А.3 показаны бинарные доли правильных исходов, точность, полнота, специфичность и F_1 для каждого класса, рассчитанные из результатов, показанных в таблице А.2.

Таблица А.3 — Доля правильных исходов, точность, полнота, специфичность и F_1 для классов A, B, C

В процентах

	A	B	C
Доля правильных исходов	91,74	88,27	29,15
Бинарная доля правильных исходов	95,97	86,46	89,40
Точность	70,92	95,79	15,01
Полнота	91,74	88,27	29,15
Специфичность	96,38	74,66	92,24
F_1	80,00	91,88	19,82

В таблице А.3 показано, чем расчет доли правильных исходов, выраженной в процентах, отличается от применения показателей бинарной классификации. Общая доля правильных исходов для примера, приведенного в таблице А.3, составляет 85,92 %.

В таблице А.4 показаны микро-, макро- и взвешенные результаты мультиклассовой классификации, полученные из данных в таблицах А.2 и А.3 с использованием формулы, показанной в 6.4.2. В таблице А.4 также показано, чем расчет доли правильных исходов отличается от применения показателей бинарной классификации. Поскольк-

ку размер класса B много больше по сравнению с классами A и C, то целесообразным является использование взвешенных оценок эффективности.

Таблица А.4 — Макро-, микро- и взвешенные результаты мультиклассовой классификации

В процентах

	Макро	Взвешенные	Микро
Доля правильных исходов	90,61	87,43	90,61
Точность	60,57	89,98	85,92
Полнота	69,72	85,92	85,92
Специфичность	87,76	77,36	92,96
F_1	63,90	87,60	85,92

Приложение В
(справочное)

Пример ROC-кривой, полученной из результатов классификации

В.1 Переход от необработанных результатов бинарной классификации к ROC-кривой

Приведенные ниже рисунки показывают переход от необработанных результатов бинарной классификации к ROC-кривой, как описано в 6.3.4. Сначала результаты классификации сортируются в порядке убывания, как показано в таблице В.1.

Таблица В.1 — Пример результатов классификации для генерации ROC

Предсказанная вероятность	Истинный класс
1,00	Да
0,96	Да
0,94	Да
0,86	Да
…	…
0,03	Нет
0,03	Да
0,00	Нет

Используя порог отсечения по предсказанной вероятности, для каждой строки таблицы можно создать матрицу ошибок, подсчитывающую количество истинных и ложных положительных и отрицательных результатов. Затем доли T_P и F_P при данном пороге вероятности строятся в пространстве ROC. Данный процесс показан на рисунках В.1 — В.7, где крестиком обозначена неправильная классификация при данном пороге вероятности, а черным кругом отмечены правильные классификации:

a) построение графика фактических значений класса в сравнении с предсказанными вероятностями для каждой выборки (см. рисунок В.1);

b) порог = 0,99 (вертикальная линия), где доля T_P = 0,18 и доля F_P = 0 (см. рисунок В.2);

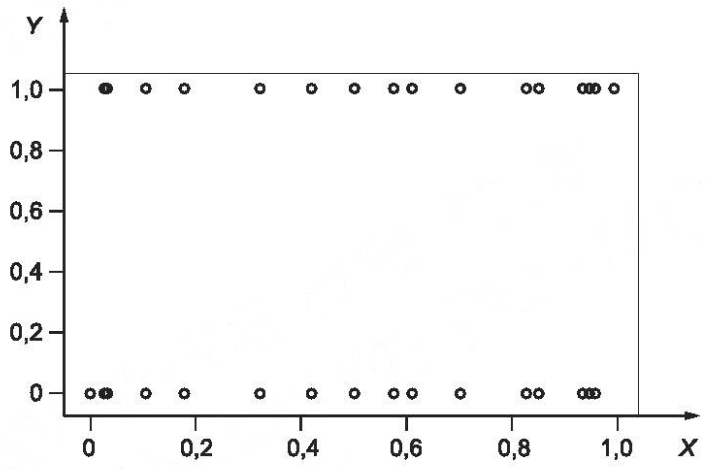
c) порог = 0,9, где доля T_P = 0,68 и доля F_P = 0,02 (см. рисунок В.3);

d) порог = 0,7, где доля T_P = 0,75 и доля F_P = 0,03 (см. рисунок В.4);

e) порог = 0,3, где доля T_P = 0,96 и доля F_P = 0,14 (см. рисунок В.5);

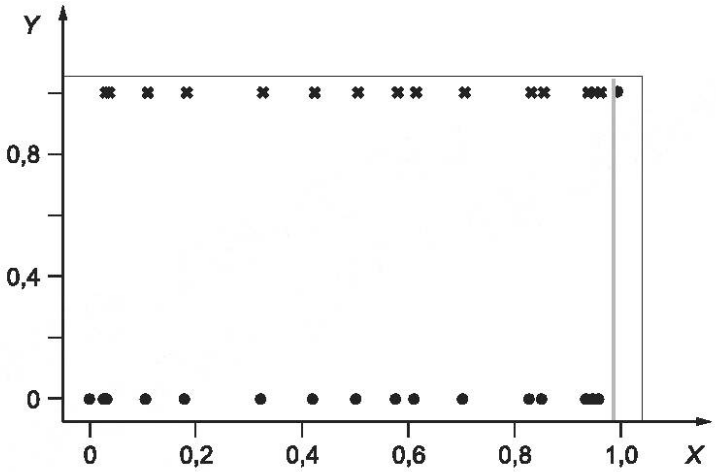
f) порог = 0,01, где доля T_P = 1 и доля F_P = 0,59 (см. рисунок В.6);

g) ROC-кривая, построенная по нескольким опорным точкам *y* (см. рисунок В.7).



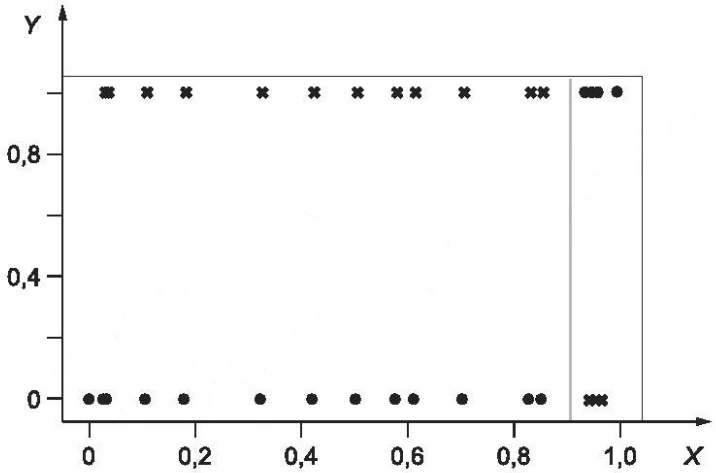
X — предсказанная вероятность; *Y* — фактические значения

Рисунок В.1 — Построение фактических классовых значений в сравнении с предсказанными вероятностями каждой выборки



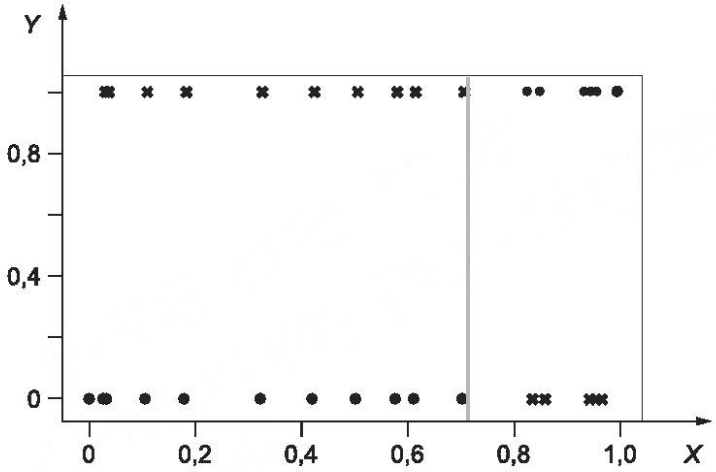
X — предсказанная вероятность; *Y* — фактические значения

Рисунок В.2 — Порог = 0,99, доля T_P = 0,18, доля F_P = 0



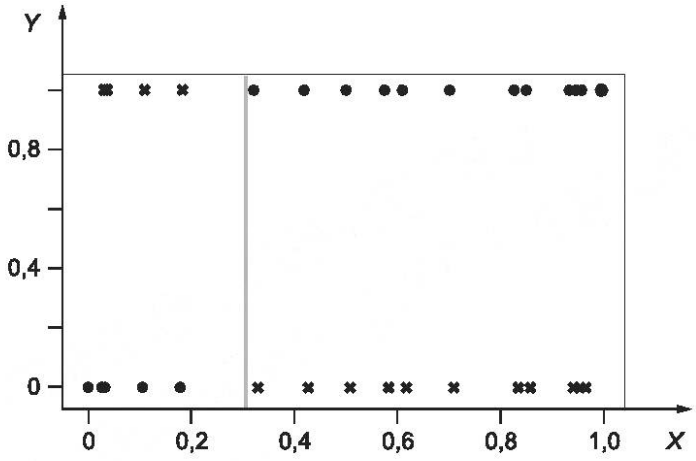
X — предсказанная вероятность; *Y* — фактические значения

Рисунок В.3 — Порог = 0,9, доля T_P = 0,68, доля F_P = 0,02



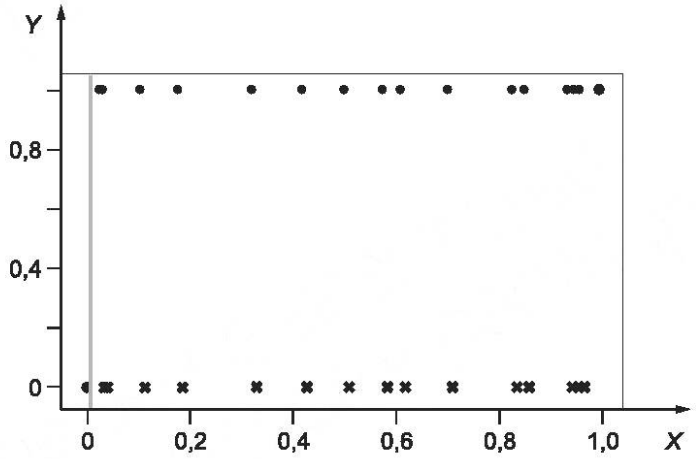
X — предсказанная вероятность; *Y* — фактические значения

Рисунок В.4 — Порог = 0,7, доля T_P = 0,75, доля F_P = 0,03



X — предсказанная вероятность; *Y* — фактические значения

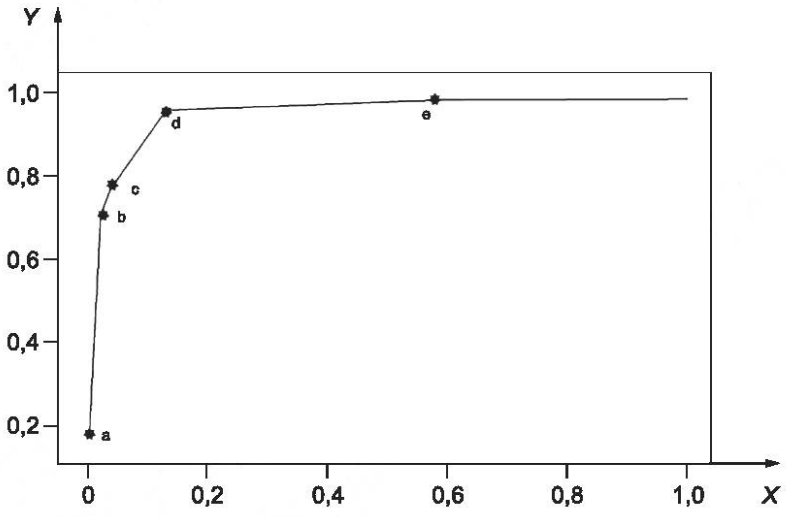
Рисунок В.5 — Порог = 0,3, доля T_P = 0,96, доля F_P = 0,14



X — предсказанная вероятность; *Y* — фактические значения

Рисунок В.6 — Порог = 0,01, доля T_P = 1, доля F_P = 0,59

Затем точки соединяются и образуют ROC-кривую.



X — предсказанная вероятность; Y — фактические значения; [a] Порог = 0,99; [b] Порог = 0,9; [c] Порог = 0,7; [d] Порог = 0,3; [e] Порог = 0,01

Рисунок В.7 — ROC-кривая, построенная в нескольких точках *y*

Приложение С
(справочное)

Сводная информация об эталонных тестах моделей и алгоритмов машинного обучения для решения задачи классификации

С.1 Примеры эталонных тестов моделей и алгоритмов машинного обучения для решения задачи классификации

В таблице С.1 приведена сводная информация по тестам моделей и алгоритмов машинного обучения для решения задачи классификации, в т. ч. слой свертки (conv) и FC.

Таблица С.1 — Сводная информация об эталонных тестах моделей и алгоритмов машинного обучения для решения задачи классификации

Набор данных	Модель	Доля правильных исходов, %	Ошибка по Топ 1, %	Ошибка по Топ 5, %	Глубина модели	Параметры	Параметры и значения модели
ImageNet	AlexNet	63,30	37,50	17,0	8 conv: 5 FC: 3	60M	- Размер батча: 128; - импульс: 0,9; - уменьшение весов: 0,0005; - темп обучения: 0,010
	VGG-19	74,50	25,50	8,0	19 conv: 16 FC: 3	144M	- Размер батча: 256; - импульс: 0,9; - уменьшение весов: 0,0005; - темп обучения: 0,010
	ResNet-50	77,15	22,85	6,7	50 conv: 49 FC: 1	20M	- Размер батча: 256; - импульс: 0,9; - уменьшение весов: 0,0001; - темп обучения: 0,100
	EfficientNet-B7	84,40	15,60	2,9	813	66M	- Импульс: 0,9; - уменьшение весов: 0,0001; - темп обучения: 0,256
MNIST	RMDL	99,82	—	—	—	—	—
	MCDNN	99,77	—	—	5 conv: 2 pooling: 2 FC: 1	—	—
	LeNet	99,30	—	—	6 conv: 3 Pooling: 2 FC: 1	—	—
CIFAR-10	EfficientNet-B7	98,90	—	—	—	64M	- Импульс: 0,9; - уменьшение весов: 0,0001; - темп обучения: 0,256
	ColorNet	98,46	—	—	—	19,0M	- Импульс: 0,9; - скорость затухания: 0,95; - темп обучения: 0,0001—0,0010
	DenseNet	96,54	—	—	—	—	- Размер батча: 6; - импульс: 0,9; - уменьшение весов: 0,0001; - темп обучения: 0,1000; - уровень исключения: 0,2

Окончание таблицы С.1

Набор данных	Модель	Доля правильных исходов, %	Ошибка по Топ 1, %	Ошибка по Топ 5, %	Глубина модели	Параметры	Параметры и значения модели
LFW	FaceNet	99,63	—	—	—	7,5M	- Темп обучения: 0,0500
	DeepID3	99,53	—	—	—	—	—
	DeepFace	97,50	—	—	—	—	- Размер батча: 128; - темп обучения: 0,0100

Приложение D
(справочное)

CSMF

D.1 Расчет доли правильных исходов CSMF

CSMF представляет собой долю внутрибольничной смертности по заданной причине, нормализованную по всем причинам, при том что основная причина смерти закодирована согласно Международной классификации болезней. Доля правильных исходов C_{SMF} представляет собой меру качества предсказаний на уровне популяции, которая выражает, насколько близко оцененные значения C_{SMF} приближаются к истине.

Дана матрица ошибок M, где $M_{i,j}$ представляет собой количество объектов данных с истинным классом i и предполагаемым (или предсказанным) классом j

$$C^t_{SMFi} = \frac{\sum_{k=1}^{K} M_{i,k}}{N}, \tag{D.1}$$

и

$$C^p_{SMFj} = \frac{\sum_{k=1}^{K} M_{k,j}}{N}, \tag{D.2}$$

где N — общее количество объектов данных,

$$N = \sum_{i}^{K}\sum_{j}^{K} M_{i,j}, \tag{D.3}$$

тогда доля правильных исходов CSMF без поправки на случайность (C_{SMFA}) с минимальным значением m от j составляет

$$C_{SMFA} = 1 - \frac{\left|C^t_{SMFi} - C^p_{SMFj}\right|}{2 \cdot \left(1 - m_J C^t_{SMFi}\right)}. \tag{D.4}$$

В случаях, когда возможно вычислить соответствующие коэффициенты, доля правильных исходов CSMF может быть использована.

Приложение ДА
(справочное)

Сведения о соответствии ссылочных национальных стандартов международным стандартам, использованным в качестве ссылочных в примененном международном стандарте

Таблица ДА.1

Обозначение ссылочного национального стандарта	Степень соответствия	Обозначение и наименование ссылочного международного стандарта
ПНСТ 838—2023 (ИСО/МЭК 23053:2022)	MOD	ISO/IEC 23053:2022 «Экосистема разработки систем искусственного интеллекта (ИИ) с использованием машинного обучения (МО)»

Примечание — В настоящей таблице использовано следующее условное обозначение степени соответствия стандарта:

- MOD — модифицированный стандарт.

Библиография

[1] *ИСО/МЭК 22989:2022* *Информационная технология. Искусственный интеллект. Концепции и терминология искусственного интеллекта (Information technology — Artificial intelligence — Artificial intelligence concepts and terminology)*

[2] ИСО 7200:2004 Техническая документация на продукцию. Поля данных в блоках наименований и заголовках документа (Technical product documentation — Data fields in title blocks and document headers)

УДК 004.01:006.354 ОКС 35.020

Ключевые слова: искусственный интеллект, оценка эффективности, модели машинного обучения, алгоритмы машинного обучения, классификация

Редактор *Е.Ю. Митрофанова*
Технический редактор *В.Н. Прусакова*
Корректор *Л.С. Лысенко*
Компьютерная верстка *Е.О. Асташина*

Сдано в набор 17.11.2023. Подписано в печать 07.12.2023. Формат 60×84⅛. Гарнитура Ариал.
Усл. печ. л. 3,72. Уч.-изд. л. 2,98.

Подготовлено на основе электронной версии, предоставленной разработчиком стандарта

Создано в единичном исполнении в ФГБУ «Институт стандартизации»
для комплектования Федерального информационного фонда стандартов,
117418 Москва, Нахимовский пр-т, д. 31, к. 2.
www.gostinfo.ru info@gostinfo.ru