**МЕТРИКИ РЕГРЕССИИ**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **НАЗВАНИЕ** | **ФОРМУЛА** | **ИНТЕРПРЕТАЦИЯ И ПРИМЕНЕНИЕ** | **ДОСТОИНСТВА** | **НЕДОСТАТКИ** | **ФУНКЦИЯ В МОДУЛЕ METRICS БИБЛИОТЕКИ SKLEARN** |
| **MAE** |  | Помогает оценить абсолютную ошибку: насколько в среднем число в предсказании разошлось с реальным числом. | Удобно интерпретировать. Измеряется в тех же единицах, что и целевой признак. Несильно искажается при наличии выбросов. | Не поможет, если необходимо сравнить модели, предсказывающие одно и то же по разным признакам. | mean\_absolute\_error() |
| **MAPE** |  | Помогает абстрагироваться от конкретных чисел и оценить абсолютную ошибку в процентах. | Легко интерпретировать. Используется в задачах, где неизвестно, какое значение метрики считать приемлемым. | Плохо подходит для задач, в которых важны конкретные единицы измерений. Лучше использовать в паре с *MAE*, чтобы знать абсолютную ошибку и её значение в процентах. | mean\_absolute\_percentage\_error() |
| **MSE** |  | Интерпретации нет. Используется в задачах, где критически важны большие ошибки, например при предсказании координат полёта. | Каждая ошибка вносит свой квадратичный штраф, большие расхождения между предсказанием и истиной увеличивают штраф. | Измеряется в квадратах единиц, поэтому менее доступна для понимания. Искажается при наличии выбросов. Не поможет, если нужно сравнить модели, предсказывающие одно и то же по разным признакам. | mean\_squared\_error() |
| **RMSE** |  | Можно трактовать как стандартное отклонение предсказаний от истинных ответов. Используется в тех же задачах, что и *MSE*. | Имеет те же преимущества, что и *MSE*, но более удобна для понимания (измеряется в тех же единицах, что и целевая переменная). | Не поможет, если нужно сравнить модели, предсказывающие одно и то же по разным признакам. | Отдельная функция отсутствует, но можно извлечь корень из результата функции mean\_squared\_error(). |
| R2 |  | Помогает понять, какую долю разнообразия (дисперсии) смогла уловить модель в данных. Позволяет сравнить, насколько модель лучше, чем простое предсказание средним. | Можно сравнивать модели, обученные на разных признаках. Легко оценить качество модели: измеряется от  до 1. Удовлетворительным показателем считается показатель выше 0.5. | Чувствительна к добавлению новых данных. Чувствительна к выбросам, так как основана на *MSE*. | r2\_score() |

**СРАВНЕНИЕ АНАЛИТИЧЕСКОГО И ЧИСЛЕННОГО РЕШЕНИЙ**

У вас наверняка возник вопрос: что лучше использовать — *LinearRegression* (аналитическое решение через метод наименьших квадратов) или *SGDRegressor* (численное решение через стохастический градиентный спуск)?

Приведём сравнение двух реализаций в виде таблицы:

|  |  |  |
| --- | --- | --- |
| **Показатель сравнения/Наименование в sklearn** | **LinearRegression** | **SGDRegressor** |
| **Метод решения и его сходимость к истинному минимуму** | Аналитический — метод наименьших квадратов. Это главное преимущество метода: есть формула => подставили значения => совершили вычисления.  Аналитический метод по определению является сходящимся, так как опирается на условие минимума функции. | Численный — метод стохастического градиентного спуска. Поиск минимума осуществляется итерациями.  Сходимость зависит от множества факторов: темпа обучения, характера функции потерь, критерия остановки. |
| **Функция потерь** | Средний квадрат ошибки (*MSE*) | Любая гладкая функция, главное — чтобы она была дифференцируемой во всех точках. Функции потерь, доступные в *sklearn*, можно увидеть [здесь](https://scikit-learn.ru/1-5-stochastic-gradient-descent/#mathematical-formulation). Каждая функция потерь предназначена для конкретной задачи. |
| **Сложность алгоритма и время обучения** | Кубическая сложность из-за вычисления обратной матрицы.  Время обучения кубически возрастает, что критически сказывается на наборах данных с большим количеством признаков. | Линейная сложность, простые математические операции умножения и сложения.  Время обучения линейно возрастает с количеством признаков. |
| **Возможность дообучения по новым данным** | Отсутствует. Все данные должны быть поданы в модель заранее. Новый вызов fit() приведёт к новой настройке параметров. | Есть возможность дообучить модель на новых данных в режиме реального времени ([инкрементальное обучение](https://coderzcolumn.com/tutorials/machine-learning/scikit-learn-incremental-learning-for-large-datasets)). Повторный вызов fit() уточняет уже существующие параметры модели. |
| **Чувствительность к разному масштабу факторов** | Низкая, стандартизация (нормализация) факторов желательна только на большом количестве признаков в данных. | Обязательная стандартизация (нормализация) факторов при наличии разных масштабов из-за особенностей сходимости. |
| **Подбор внешних параметров** | Внешних параметров нет. | Для поисков лучшего решения, возможно, придётся подбирать параметры: начальный темп обучения, режим обучения и т. д. Правильную реализацию подбора параметров мы обсудим в отдельном модуле. |

По приведённой таблице можно выделить **примерные области применения каждого из методов:**

* + Если стоит задача одноразового обучения на всех данных, которые есть, и признаков немного (меньше 1 000), наш выбор — *LinearRegression*, так как МНК обеспечивает простое решение и гарантированную сходимость.
  + Если стоит задача непрерывного обучения модели в процессе её эксплуатации или количество признаков очень велико, наш выбор — *SGDRegressor* с возможностью корректировки параметров на новых данных.

Однако существенной разницы между двумя подходами нет, так как используется одна и та же модель. Наиболее распространённым является классический метод наименьших квадратов (*LinearRegression*), им в прикладных задачах пользуется большинство дата-сайентистов.

**МЕТРИКИ КЛАССИФИКАЦИИ**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **НАЗВАНИЕ** | **ФОРМУЛА** | **ИНТЕРПРЕТАЦИЯ И ПРИМЕНЕНИЕ** | **ДОСТОИНСТВА** | **НЕДОСТАТКИ** | **ФУНКЦИЯ В МОДУЛЕ METRICS БИБЛИОТЕКИ SKLEARN** |
| **ACCURACY (ДОСТОВЕРНОСТЬ)** |  | Доля правильных ответов среди всех ответов модели. Применяется в задачах, где классы сбалансированы. | Очень легко интерпретировать. Автоматически можно посчитать процент ошибок модели как . | Плохо показывает себя на сильно несбалансированных классах. | [accuracy\_score()](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html) |
| **PRECISION (ТОЧНОСТЬ)** |  | Способность модели отделять класс 1 от класса 0.  Используется в задачах, где важно минимальное количество ложноположительных срабатываний модели. | Можно использовать на несбалансированных выборках. | Вычисляется только для положительного класса — класса 1. Для класса 0 показатель необходимо вычислять отдельно. Не даёт представления о том, как много объектов положительного класса из общей совокупности нашёл алгоритм (как много ложноотрицательных срабатываний допустил алгоритм). | [precision\_score()](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html#sklearn.metrics.precision_score) |
| **RECALL (ПОЛНОТА)** |  | Способность модели находить класс 1.  Используется в задачах, где важно охватить как можно больше объектов положительного класса (1), то есть уменьшить количество ложноотрицательных срабатываний. | Можно использовать на несбалансированных выборках. | Вычисляется только для положительного класса — класса 1. Для класса 0 показатель необходимо вычислять отдельно. Не даёт представления о том, насколько точно модель находит объекты положительного класса (как много ложноположительных срабатываний). | [recall\_score()](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html#sklearn.metrics.recall_score) |
| **-МЕРА** |  | Нет бизнес-интерпретации.  Используется в задачах, где необходимо балансировать между *precision* и *recall*. | Даёт обобщённое представление о точности и полноте.  Максимум достигается, когда максимальны обе метрики, минимум — когда хотя бы одна из метрик равна 0.  При желании можно использовать обобщённый вариант — , чтобы управлять вкладом *precision* в общую метрику. | Отсутствие интерпретации не даёт интуитивного понимания человеку, не знакомому с этой метрикой. | [f1\_score()](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html#sklearn.metrics.f1_score) |