

Оценивающее правило (Scoring rule)

Разделы: [Метрики](#)

В машинном обучении и теории принятия решений правило, используемое для оценки качества работы вероятностных предсказательных моделей, выходом которых является вероятность некоторого события, например принадлежности объекта к классу в задачах классификации.

В основе идеи использования таких правил лежит начисление баллов модели за каждое предсказание: чем выше качество предсказания, тем больше начисляемый балл. Отсюда и название метода: от английского *scoring* — начисление баллов. К кредитному скорингу данный метод не имеет прямого отношения.

Обозначим начисляемый балл s (от *score*), предсказанную вероятность p , а признак события e (от *event*). Тогда для набора из n взаимоисключающих событий (например, принадлежности объекта к классу), можно задать вектор $\mathbf{p} = (p_1, p_2, \dots, p_n)$, который содержит вероятности, присваиваемые соответствующим выходам модели (например, классам).

При этом должно выполняться условие:

$$\sum_{i=1}^n p_i = 1.$$

Поскольку события взаимоисключающие и в совокупности исчерпывающие, только один из возможных вариантов окажется правильным. Тогда вектор $\mathbf{e} = (e_1, e_2, \dots, e_n)$ указывает, какой из взаимоисключающих вариантов оказался правильным: соответствующий элемент вектора принимает значение 1, а остальные 0.

Рассмотрим пример. Пусть модель бинарной классификации должна классифицировать ситуацию по курсу валюты на завтра. При этом выход модели (предсказанная вероятность p) лежит в диапазоне от 0 до 1. Соответственно, значение 1 указывает на то, что целевое событие (изменение курса) точно произойдет, а 0 — что точно не произойдет.

Промежуточные значения можно интерпретировать как степень уверенности «предсказателя» в том или ином исходе.

Вероятность можно интерпретировать как степень уверенности в том, что событие произойдет. Например, величина $p = 0.8$ указывает на высокую уверенность, что курс изменится, а $p = 0.2$ на низкую (или высокую уверенность, что не изменится). Таким образом, для каждого предсказания фиксируются две величины: вероятностное значение, формируемое моделью (изменяется от 0 до 1) и фактический исход (1 — событие произошло, 0 — не произошло).

Теперь рассмотрим четыре ситуации:

1. Модель выдала предсказание $p = 1$ (100% уверенность, что целевое событие произойдет), и событие действительно произошло, т. е. $e = 1$. В этом случае «предсказатель» выдал правильный ответ с наивысшей степенью уверенности. Это наилучший вариант, и за такое предсказание должно быть начислено максимальное количество баллов (например, $s = 1$).
2. Модель выдала предсказание $p = 1$ (100% уверенность, что событие произойдет), но событие не произошло, т. е. $e = 0$. Здесь «предсказатель» оказался абсолютно уверен в своем ответе, но он оказался неверным. Это наихудший случай, и за такое предсказание начисляется минимальное количество баллов (скажем, $s = 0$).
3. Модель выдала предсказание $p = 0.7$ (70% уверенность, что событие произойдет), и событие действительно произошло, т. е. $e = 1$. Это можно интерпретировать как верный прогноз с высокой, но не абсолютной уверенностью. За него модели начисляют высокий балл, но меньший, чем в случае полной уверенности (например, $s = 0.7$).
4. Модель выдала предсказание $p = 0.7$ (70% уверенность, что событие произойдет), но событие не произошло, т. е. $e = 0$. В этом случае модель ошиблась, но ее уверенность была не абсолютной. За такое предсказание начисляется низкий балл, но выше, чем в случае полной уверенности в неправильном ответе (например, $s = 0.3$).

Именно правило, по которому производится начисление таких баллов и называется **оценивающим**. Оно, в частности, будет зависеть от того, какое событие считается целевым. Например, если предсказание $p = 0$ рассматривать как полную уверенность в том, что событие не произойдет (т.е. тоже $e = 0$), то максимальное количество баллов будет начисляться уже для этого предсказания.

Для наглядности поясним сказанное в таблице:

Предсказание, p	Факт	Баллы, s
1	1	1
1	0	0
0	0	1
0	1	0
0.7	1	0.7
0.7	0	0.3
0.3	1	0.3
0.3	0	0.7

Правило, приведенное в таблице, является «наивным». В случае, если фактическое событие соответствует предсказанному ($e = 1$), начисляемое количество баллов равно его предсказанной вероятности, то есть $s = p$. В противном случае $s = 1 - p$.

Поскольку начисленный балл оценивает качество предсказания модели на каждом наблюдении, его можно рассматривать как функцию потерь и использовать для повышения производительности модели. Кроме того, оценивающие правила позволяют сравнивать модели и выбирать те из них, для которых получены лучшие балльные оценки (то есть минимизировать функцию потерь).

На практике применяют оценивающие правила, называемые «правильными» (proper scoring rules). Такое правило обеспечивает максимальную балльную оценку (или минимальные потери) в случае, когда распределение предсказанных вероятностей соответствует распределению фактических событий, а частота предсказанных событий совпадает с частотой их реального наступления.

Например, можно зафиксировать за длительный период, сколько раз предсказанная вероятность $p = 0.65$ для изменения курса валюты соответствовала фактическому изменению. Если доля правильно предсказанных изменений совпадает с долей положительных предсказаний и при этом правило обеспечивает максимальную балльную оценку (или минимальные потери), то оно считается верным. В противном случае говорят, что модель «плохо откалибрована».

В настоящее время разработано множество правильных оценивающих правил. Однако наиболее популярными являются логарифмическое правило Гуда, правило Брайера и квадратичное правило.

Логарифмическое правило (Logarithmic score). Представляет собой функцию оценки, которая использует формулу:

$$S_{\log(p)} = \ln(p_i).$$

Например, если модель сделает предсказание вероятности изменения курса валюты $p = 0.7$, а того, что он останется постоянным — $p = 0.3$, то если курс по факту изменится, то будет начислено $S(p, e) = \ln(0.7) = -0.36$, а если не изменится, то $S(p, e) = \ln(0.3) = -1.2$.

Конечно, отрицательные значения оценок могут выглядеть необычно, но, тем не менее, лучшим предсказаниям все равно соответствуют более высокие баллы.

Функция потерь для логарифмического правила будет иметь вид:

$$S_{\log}(p, e) = - \sum_{i=1}^n e_i \ln(p_i),$$

где n — число предсказаний, сделанных моделью.

Правило Брайера (Brier score) было предложено в 1950 году американским статистиком Гленном Брайером и задается выражением:

$$S_{brier}(p, e) = \frac{1}{n} \sum_{i=1}^n (p_i - e_i)^2.$$

По сути, это не что иное, как среднеквадратическая ошибка предсказания для бинарных исходов. Ее можно интерпретировать следующим образом:

- если $p = 1$ и $e = 1$, то получим $s = (1 - 1)^2 = 0$, что соответствует наилучшей оценке;

- если $p = 1$ и $e = 0$, то получим $s = (1 - 0)^2 = 1$, что соответствует наихудшей возможной оценке;
- если $p = 0.7$ и $e = 1$, получим $s = (0.7 - 1)^2 = 0.09$;
- если $p = 0.7$ и $e = 0$, получим $s = (0.7 - 0)^2 = 0.49$;
- если $p = 0.5$, т.е. предсказана вероятность события 0.5, то оценка Брайера всегда будет $s = (0.5 - 1)^2 = (0.5 - 0)^2 = 0.25$ независимо от того, произошло событие или нет.

Таким образом, оценка Брайера лежит в диапазоне от 0 до 1, и чем ближе значение к 1, тем хуже точность соответствующего предсказания.

Оценка Брайера в том виде, как мы ее рассмотрели, пригодна только для случая с бинарным исходом. Однако существует ее обобщение для произвольного числа исходов (классов) R :

$$S_{brier}(p, e) = \frac{1}{n} \sum_{i=1}^N \sum_{t=1}^R (p_{it} - e_{it})^2.$$

Например, если расширить задачу о предсказании котировок валют до трех возможных исходов: курс не изменится, увеличится и уменьшится, то $R = 3$.

Квадратичное правило задает оценку в соответствии с формулой:

$$S_{quad}(p, e) = \sum_{i=1}^N p_i(2e_i - p_i).$$

Несложно увидеть, что значение балльной оценки в этом случае изменяется от -1 до 1. Действительно, если $p = 1$ и $e = 1$, то получим $s = 1 \cdot (1 \cdot 2 - 1) = 1$, что соответствует наивысшей оценке. Наихудшей оценке соответствует $s = 1 \cdot (1 \cdot 0 - 1) = -1$. Иными словами, чем ближе оценка к 1, тем точнее предсказание с точки зрения квадратичного правила.

Оценивающие правила позволяют не только определять точность предсказаний, но и сравнивать предсказательные модели (например, классификаторы).

Рассмотрим ситуацию, когда две модели имеют одинаковую точность, то есть делают правильные предсказания в одинаковом числе случаев. Однако одна из них чаще предсказывает верные результаты, когда предсказанная вероятность высока (например, $p = 0.7 - 0.9$, высокая уверенность), и реже — при низких значениях вероятности ($p = 0.1 - 0.3$, низкая уверенность).

Вторая модель, напротив, делает правильные предсказания независимо от уровня уверенности. Очевидно, что первая модель предпочтительнее, так как ее предсказания более согласованы с уровнями уверенности.

Аналитики используют различные методы оценивания. При этом можно выделить несколько ключевых свойств, которые делают одни методы более предпочтительными, чем другие:

- **Вознаграждение точности:** правильные предсказания с высокой уверенностью должны получать больше баллов, а ошибки — наказываться сильнее.
- **Интуитивность:** чем выше балл, тем лучше предсказание.

- **Отсутствие искажений:** итоговая оценка модели не должна зависеть от нескольких случайных удачных предсказаний.
- **Ограничение верхней границы:** предотвращает аномальные значения баллов и излишнюю концентрацию модели на «выгодных» примерах.
- **Наличие нижней границы:** защищает от чрезмерного штрафа за одно неудачное предсказание.
- **Простое правило знаков:** положительный балл — за правильное предсказание, отрицательный — за ошибку.
- **Учет уверенности:** уверенное предсказание оценивается выше, но и штраф за ошибку в такой ситуации больше.
- **Неопределенность равна нулю:** предсказание с максимальной неопределенностью (например, вероятность 0.5) не должно давать баллов.
- **Устойчивость:** небольшие изменения вероятности не должны сильно влиять на итоговую оценку.

Таким образом, оценивающие правила предоставляют широкие возможности для оценки качества и сравнения предсказательных моделей. Однако их использование требует учета ряда особенностей — игнорирование этих факторов может привести к искаженному представлению о точности и надежности моделей.