Оценка качества моделей (метрики классификации и регрессии, подбор гиперпараметров)

30.09.2024

Метрики в задачах классификации

Перед переходом к самим метрикам необходимо ввести важную концепцию для описания этих метрик в терминах ошибок классификации — confusion matrix (матрица ошибок).

Допустим, что у нас есть два класса и алгоритм, предсказывающий принадлежность каждого объекта одному из классов, тогда матрица ошибок классификации будет выглядеть следующим образом:

Confusion matrix

	Y = 1	Y = 0
Y~ = 1	TRUE POSITIVE (TP)	FALSE POSITIVE (FP)
Y~ = 0	FALSE NEGATIVE (FN)	TRUE NEGATIVE (TN)

Здесь Y~ — это ответ алгоритма на объекте, а Y — истинная метка класса на этом объекте.

Таким образом, ошибки классификации бывают двух видов: False Negative (FN) и False Positive (FP).

Accuracy

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Интуитивно понятной, очевидной и почти неиспользуемой метрикой является *accuracy* — доля правильных ответов алгоритма:

Эта метрика **бесполезна** в задачах с неравными классами.

Precision и Recall

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

Для оценки качества работы алгоритма на каждом из классов по отдельности используются **precision** (точность) и **recall** (полнота).

Ошибки классификации бывают двух видов:

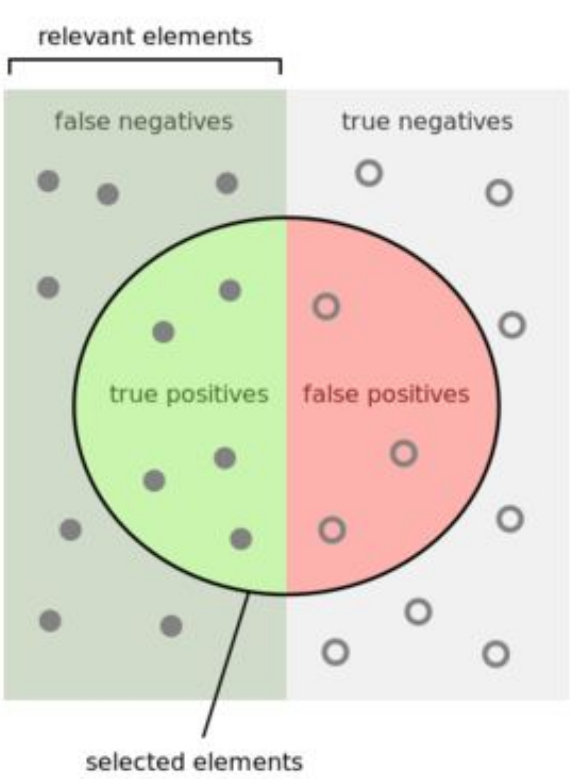
False Positive (I-го рода) и

False Negative (II-го рода).

Precision и Recall

Precision можно интерпретировать как **долю** объектов, названных классификатором **положительными** и при этом действительно **являющимися положительными**

Recall показывает, какую **долю** объектов **положительного класса** из всех объектов **положительного** класса нашел алгоритм.



Precision и Recall

Precision и recall **не зависят**, в отличие от accuracy, от соотношения классов и потому применимы в условиях несбалансированных выборок.

Часто в реальной практике стоит задача найти **оптимальный** (для заказчика) **баланс** между этими двумя метриками.

F1-мера

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

Существует несколько различных способов объединить precision и recall в агрегированный критерий качества. **F-мера** — среднее гармоническое precision и recall:

F-мера достигает максимума при полноте и точности, равными единице, и близка к нулю, если один из аргументов близок к нулю.

AUC-ROC

$$TPR = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{FP + TN}$$

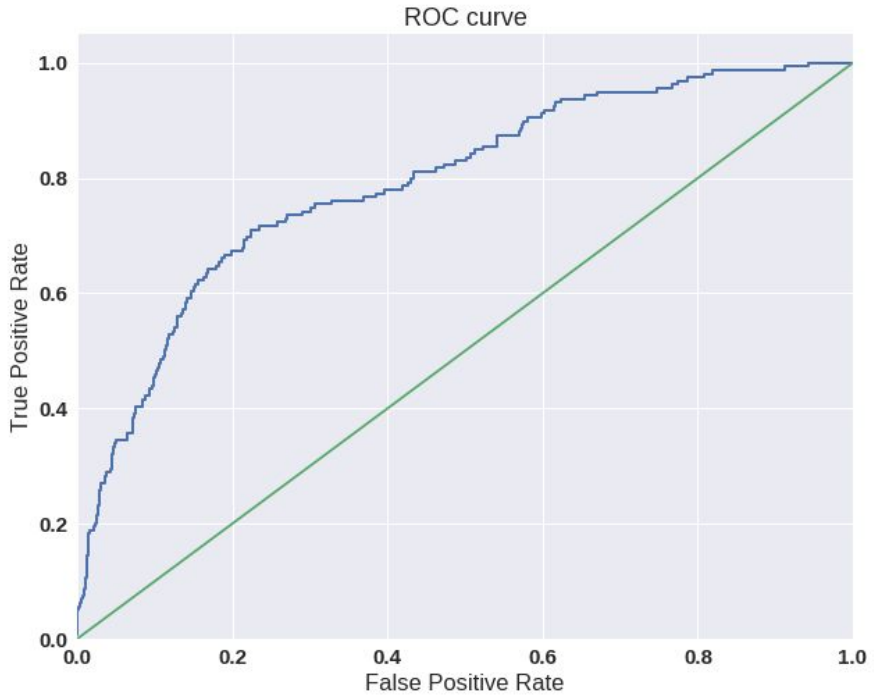
При конвертации вещественного ответа алгоритма в бинарную метку, мы должны выбрать какой-либо порог.

Одним из способов оценить модель в целом, не привязываясь к конкретному порогу, является **AUC-ROC** (или ROC AUC) — площадь (Area Under Curve).

Данная кривая представляет из себя линию от (0,0) до (1,1) в координатах **True Positive Rate** (TPR) и **False Positive Rate** (FPR):

AUC-ROC

AUC-ROC **устойчив** к несбалансированным классам и может быть интерпретирован как вероятность того, что случайно выбранный **positive** объект будет иметь более высокую вероятность быть **positive**, чем случайно выбранный **negative** объект.



LLM метрики

Вот наиболее важные и распространенные метрики для LLM:

- **Релевантность**: Определяет, способен ли вывод LLM информативно и лаконично ответить на заданный ввод.
- **Корректность**: Определяет, является ли вывод LLM фактически правильным, основанным на некоторой базовой истине.
- **Галлюцинация**: Определяет, содержит ли вывод LLM фальшивую или выдуманную информацию.

LLM пользовательская метрика

Вам понадобится как минимум одна пользовательская метрика, специфичная для конкретной задачи.

Например, если ваше приложение LLM предназначено для **суммаризации** страниц новостных статей, вам понадобится пользовательская метрика оценки LLM, которая будет определять:

- Содержит ли саммари достаточно информации из оригинального текста.
- Содержит ли саммари какие-либо противоречия или галлюцинации по сравнению с оригинальным текстом.

Какими свойствами должна обладать метрика

Количественная. Метрики всегда должны вычислять балл при оценке поставленной задачи. Такой подход позволяет установить минимальный порог прохождения, чтобы определить, является ли ваше приложение LLM «достаточно хорошим», и позволяет отслеживать, как эти баллы меняются со временем по мере итераций и улучшения вашей реализации.

Надежность. Как бы непредсказуемы ни были результаты LLM, меньше всего вам хочется, чтобы метрика оценки LLM была такой же нестабильной.

Точность. Отражает реальную эффективность вашей LLM. Секрет того, как сделать хорошую метрику оценки LLM отличной, заключается в том, чтобы она максимально соответствовала ожиданиям человека.

Scorers

BERTScore опирается на предварительно обученные языковые модели типа BERT и вычисляет **косинусоидальное сходство** между контекстуальными эмбеддингами слов в эталонном и сгенерированном текстах. Затем эти сходства суммируются для получения итогового балла.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) используется в основном для оценки текстовых суммаризация моделей NLP и рассчитывает **recall** путем сравнения перекрытия n-грамм между результатами LLM и ожидаемыми результатами. Он определяет долю [0, 1] n-грамм в ссылке, которые присутствуют в выводе LLM.

G-Eval использует LLM (GPT) для оценки результатов LLM (также известных как LLM-Evals), и является одним из лучших способов создания метрик, специфичных для конкретной задачи. (Есть Open source аналог - **Prometheus**)

Scorers сравнение

Metrics	Coherence		Consistency		Fluency		Relevance		AVG	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
ROUGE-1	0.167	0.126	0.160	0.130	0.115	0.094	0.326	0.252	0.192	0.150
ROUGE-2	0.184	0.139	0.187	0.155	0.159	0.128	0.290	0.219	0.205	0.161
ROUGE-L	0.128	0.099	0.115	0.092	0.105	0.084	0.311	0.237	0.165	0.128
BERTScore	0.284	0.211	0.110	0.090	0.193	0.158	0.312	0.243	0.225	0.175
MOVERSscore	0.159	0.118	0.157	0.127	0.129	0.105	0.318	0.244	0.191	0.148
BARTScore	0.448	0.342	0.382	0.315	0.356	0.292	0.356	0.273	0.385	0.305
UniEval	0.575	0.442	0.446	0.371	0.449	0.371	0.426	0.325	0.474	0.377
GPTScore	0.434	–	0.449	–	0.403	–	0.381	–	0.417	–
G-EVAL-3.5	0.440	0.335	0.386	0.318	0.424	0.347	0.385	0.293	0.401	0.320
- Probs	0.359	*0.313*	0.361	*0.344*	0.339	*0.323*	0.327	*0.288*	0.346	*0.317*
G-EVAL-4	**0.582**	**0.457**	**0.507**	**0.425**	**0.455**	**0.378**	**0.547**	**0.433**	**0.514**	**0.418**
- Probs	0.560	*0.472*	0.501	*0.459*	0.438	*0.408*	0.511	*0.444*	0.502	*0.446*
- CoT	0.564	0.454	0.493	0.413	0.403	0.334	0.538	0.427	0.500	0.407

Table 1: Summary-level Spearman (ρ) and Kendall-Tau (τ) correlations of different metrics on SummEval benchmark. G-EVAL without probabilities (*italicized*) should not be considered as a fair comparison to other metrics on τ, as it leads to many ties in the scores. This results in a higher Kendall-Tau correlation, but it does not fairly reflect the true evaluation ability. More details are in Section 4.