МегаФон

МЕТРИКИ КАЧЕСТВА В ML

ЧЕРНЫШЕВ ИЛЬЯ, 2021

Задачи машинного обучения

Регрессии

Цена квартиры, следующий платеж

Классификации

Имеет конечное количество ответов (как правило, в формате «да» или «нет»)

Кластеризации

Распределение данных на группы

Уменьшения размерности

Сведение большого числа признаков к меньшему

Выявления аномалий

Отделение аномалий от стандартных случаев.

Группы метрик

Группа метрики зависит от типа задачи, от модели ML, от приложения/инструмента

- Метрики классификации (точность, отзыв, F1-оценка, ROC, AUC,…)
- Метрики регрессии (MSE, MAE)
- Рейтинговые метрики (MRR, DCG, NDCG)
- Статистические метрики (корреляция)
- Метрики компьютерного зрения (PSNR, SSIM, IoU)
- Метрики НЛП (недоумение, оценка BLEU)
- Метрики, связанные с глубоким обучением (начальная оценка, начальная дистанция Фреше)

Что такое метрика качества?

- Метрика или метрика качества это характеристика качества работы вашей модели на выборке.
- Чаще всего под качеством алгоритма подразумевают значение метрики на тестовом множестве

Зачем нужны метрики качества?

- Для оценки качества работы модели
- Для сравнения моделей
- Для интерпретации результатов

РЕГРЕССИЯ

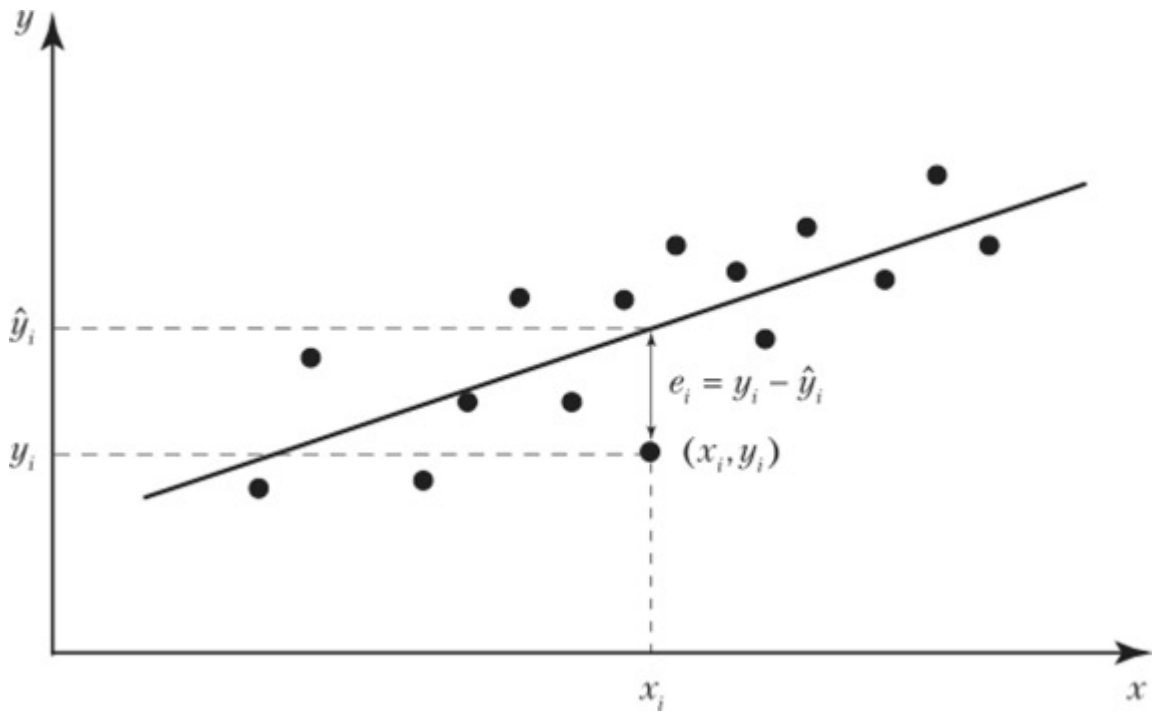
(MSE, MAE, R^2 и др)

Регрессия

- Используется для прогнозирования непрерывных целевых значений
- Метрики, используемые для оценки этих моделей, должны иметь возможность работать с набором непрерывных значений

Задача предсказания стоимости тарифа

Фактическое значение	Предсказание модели	Ошибка модели
120	100	-20
300	310	+10
210	200	-10
250	250	0
100	120	+20
400	450	+50
390	380	-10
310	330	+20



Метрики регрессии

- Используется для прогнозирования непрерывных целевых значений
- Метрики, используемые для оценки этих моделей, должны иметь возможность работать с набором непрерывных значений

Задача предсказания стоимости тарифа

Фактическое значение	Предсказание модели	Ошибка модели
120	100	-20
300	310	+10
210	200	-10
250	250	0
100	120	+20
400	450	+50
390	380	-10
310	330	+20

MSE (Mean Square Error)

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

RMSE (Root Mean Square Error)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

Метрики регрессии

- Используется для прогнозирования непрерывных целевых значений
- Метрики, используемые для оценки этих моделей, должны иметь возможность работать с набором непрерывных значений

Задача предсказания стоимости тарифа

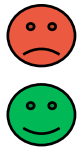
Фактическое значение	Предсказание модели	Ошибка модели
120	100	-20
300	310	+10
210	200	-10
250	250	0
100	120	+20
400	450	+50
390	380	-10
310	330	+20

MAE (Mean Absolute Error)

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Метрики регрессии

- MSE сильнее штрафует за большие отклонения по сравнению с MAE, поэтому более чувствителен к выбросам
- MSE подходит для сравнения двух моделей или для контроля качества во время обучения, но не позволяет сделать выводов о том, насколько хорошо модель решает задачу



0,12	0,93	0,01	0,25	0,78	0,89	1	0,76	0,51	0,69
1120	1930	1010	1250	1780	1890	1000	1760	1510	1690

MSE = 10

- Коэффициент детерминации R^2 $$R^2 = 1 - \frac{\sum_{i=1}^{\ell}(a(x_i) - y_i)^2}{\sum_{i=1}^{\ell}(y_i - \bar{y})^2}$$
- Нормированная среднеквадратичная ошибка. Если она близка к единице, то модель хорошо объясняет данные, если же она близка к нулю, то прогнозы сопоставимы по качеству к константным предсказанием.

Другие метрики регрессии

MAPE (Mean absolute percentage error)

$$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

RMSLE (Root mean square logarithmic error)

$$\mathrm{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1) - \log(a_i+1))^2}$$

SMAPE (Symmetric absolute percentage error)

RMSLE is used when y has long tail distribution, or we are interested in the ratio of true value and predicted value.

Когда что использовать

- Хотим учитывать выбросы – MSE, RMSE
  - Хотим интерпретировать результат – R^2
- Не хотим учитывать выбросы – MAE
  - Хотим интерпретировать результат – MAPE, SMAPE

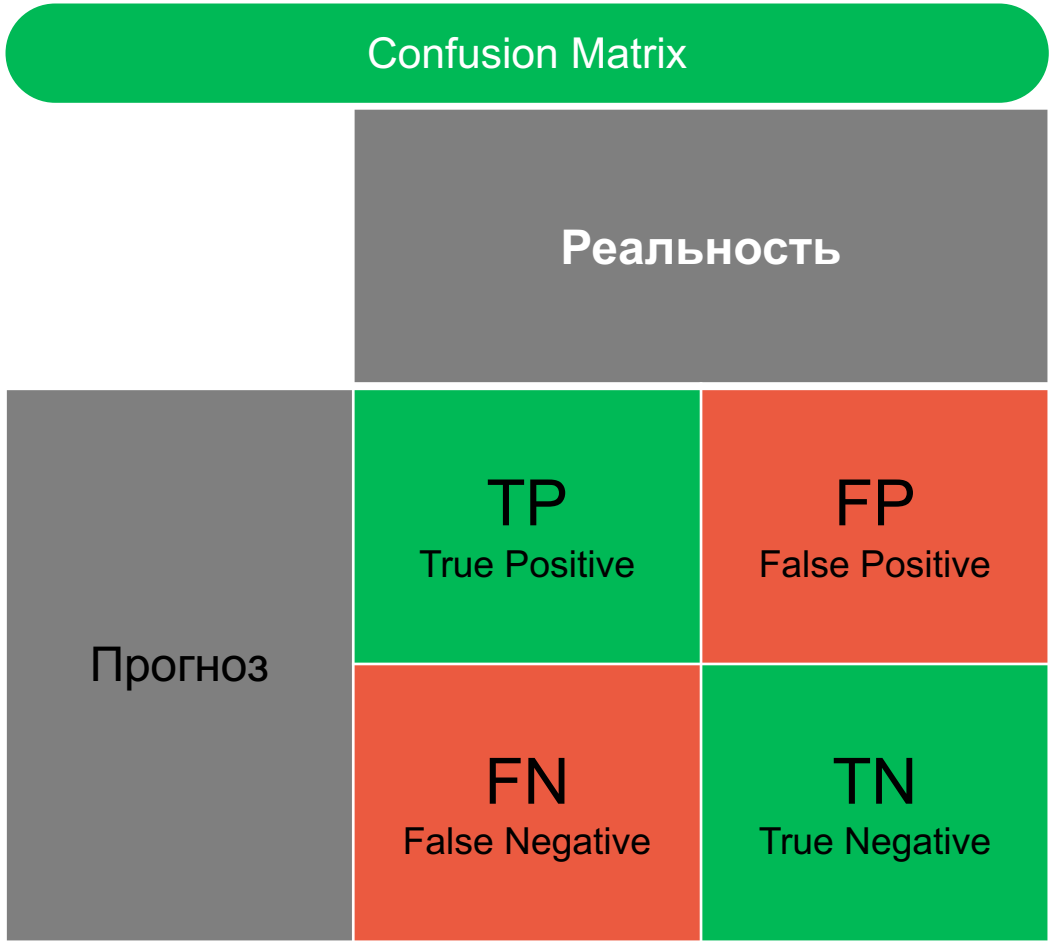
КЛАССИФИКАЦИЯ

(CONFUSION MATRIX, ACCURACY, ROC AUC и др)

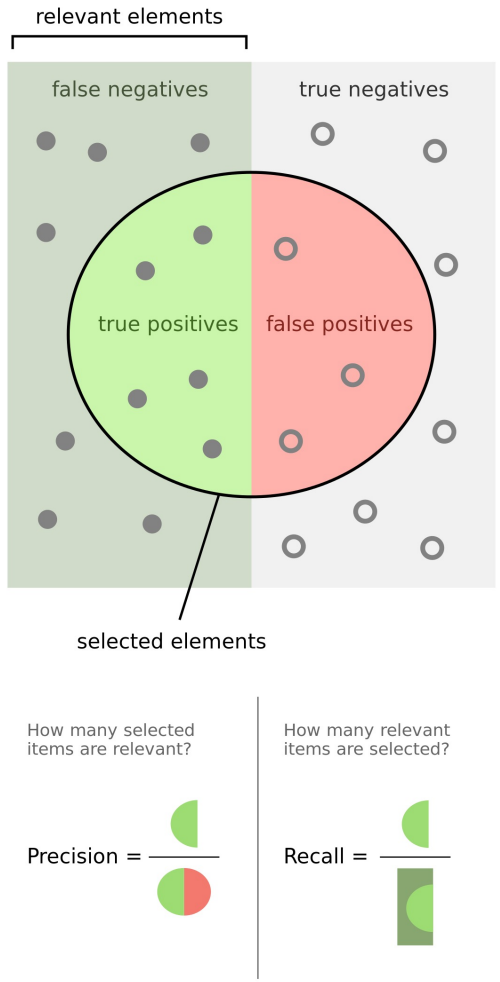
Классификация

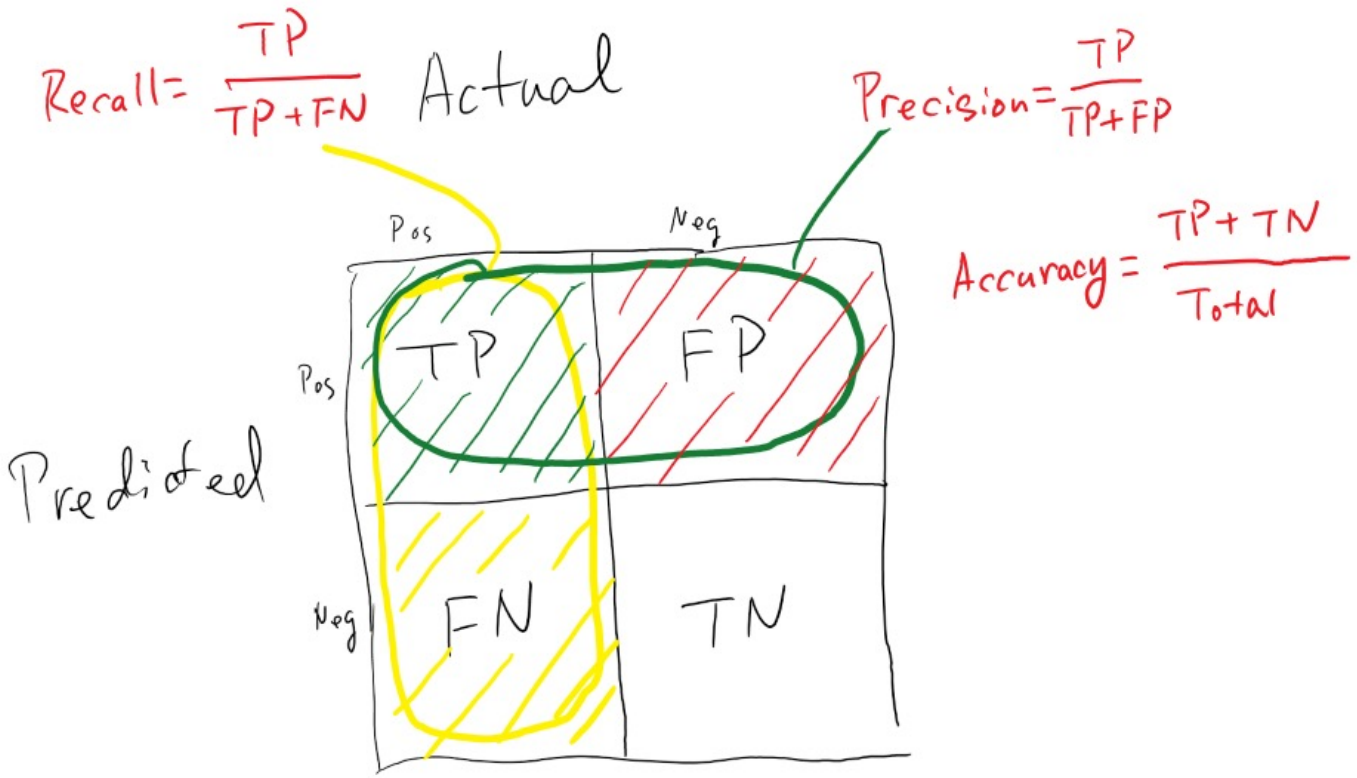
- Задача классификации – получение категориального ответа на основе набора признаков. Имеет конечное количество ответов (как правило, в формате «да» или «нет»): оттечет ли абонент завтра, готов ли абонент подключить услугу, нужен ли абоненту новый девайс.

Фактическое значение	Предсказание модели	Ошибка модели
1	1	TP
0	1	FP
1	0	FN
0	0	TN
0	1	FP
0	0	TN
0	0	TN
1	1	TP



Метрики классификации. Accuracy, Precision, Recall





Задача

Допустим, мы хотим оценить работу спам-фильтра почты. У нас есть 100 не-спам писем, 90 из которых наш классификатор определил верно и 10 спам-писем, 5 из которых классификатор также определил верно.

(True Negative = 90, False Positive = 10)
(True Positive = 5, False Negative = 5)

Accuracy = 86,4

А что если мы просто будем предсказывать все письма как не-спам?

(True Negative = 100, False Positive = 0)
(True Positive = 0, False Negative = 10)

Accuracy = 90,9

Метрики классификации. F-мера

Задача определения оттока клиентов

Мы хотим находить **всех** уходящих в отток клиентов и только их.

Определив стратегию и ресурс для удержания клиентов, мы можем подобрать нужные пороги по **precision** и **recall**. Например, можно сосредоточиться на удержании только высокодоходных клиентов или тех, кто уйдет с большей вероятностью, так как мы ограничены в ресурсах колл-центра.

F-мера — среднее гармоническое precision и recall

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

β принимает значения в диапазоне 0<β<1 если вы хотите отдать приоритет **точности**,
а при β>1 приоритет отдается **полноте**.
При β=1 формула сводится к предыдущей и вы получаете сбалансированную F-меру (также ее называют F1)

Метрики классификации

Результат работы модели

	top 1	top 5	top 10	top 30	top 50	top 70	top 100
lift	2.78	2.52	2.49	2.13	1.67	1.33	1.0
precision	88.10	79.70	78.80	67.30	52.90	42.00	31.7
recall	2.80	12.50	24.80	63.70	83.50	92.70	100.0
count	59.00	295.00	593.00	1784.00	2975.00	4158.00	5950.0
q	0.99	0.95	0.90	0.70	0.50	0.30	0.0

Метрики классификации. ROC-AUC

Одним из способов оценить модель в целом, не привязываясь к конкретному порогу, является AUC-ROC (или ROC AUC) — площадь (Area Under Curve) под кривой ошибок (Receiver Operating Characteristic curve).

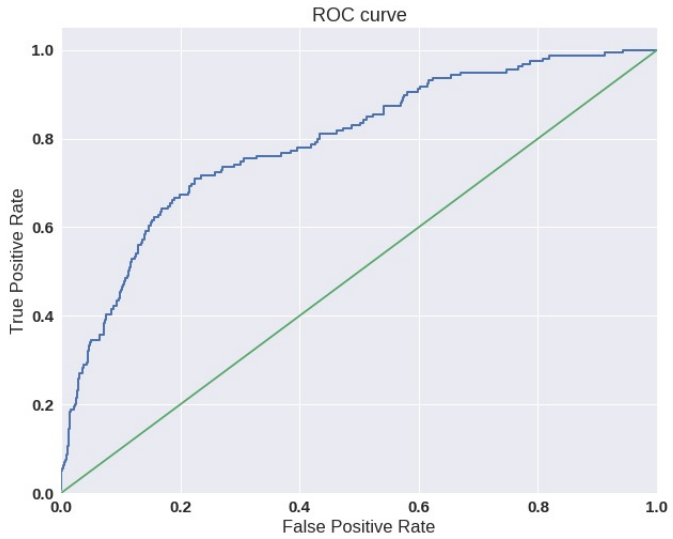
Данная кривая представляет из себя линию от (0,0) до (1,1) в координатах True Positive Rate (TPR) и False Positive Rate (FPR).

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

$$FPR = \frac{FP}{FP + TN}$$

FPR показывает, какую долю из объектов negative класса алгоритм предсказал неверно

Скор модели	Сортировка	Предсказание модели	Фактическое значение
0,1	0,1	0	1
0,9	0,2	1	0
0,2	0,4	1	1
0,5	0,5	1	0
0,4	0,7	1	0
0,8	0,8	1	0
0,7	0,9	1	0
1	1	1	1



Метрики классификации. Индекс Джини

$$GINI = 2 * ROC\text{-}AUC - 1$$

По сути это площадь между ROC-кривой и диагональю соединяющей точки (0,0) и (1, 1)