Онлайн-курс “Быстрый старт в искусственный интеллект”
Модуль 1. Машинное обучение
Дополнительные материалы к уроку 1.3 “Метрики машинного обучения”

Напомним, что в видеолекции мы подробно поговорили о метриках в задаче классификации. В этом материале мы поговорим о метриках в задаче регрессии.

Две самые распространённые метрики в задаче регрессии, о которых мы уже говорили, это MSE (mean squared error) и MAE (mean absolute error). Они вычисляются по следующим формулам:

$$MSE(y_{true}, y_{pred}) = \frac{1}{\ell}\sum_{i=1}^{\ell}(y_{true}^i - y_{pred}^i)^2;$$

$$MAE(y_{true}, y_{pred}) = \frac{1}{\ell}\sum_{i=1}^{\ell}|y_{true}^i - y_{pred}^i|.$$

Также часто используют метрику $RMSE$ (rooted mean squared error), равную корню квадратному из MSE, и $MAPE$ (mean absolute percentage error), которая вычисляется по формуле

$$MAPE(y_{true}, y_{pred}) = \frac{1}{\ell}\sum_{i=1}^{\ell}\left|\frac{y_{true}^i - y_{pred}^i}{y_{true}^i}\right|.$$

Последняя метрика хороша тем, что измеряет среднее значение абсолютной ошибки в процентах от реального значения целевой переменной.

Последняя метрика, о которой мы поговорим, называется R^2 (произносится “R-квадрат”), или коэффициент детерминации. Она вычисляется по формуле

$$R^2(y_{true}, y_{pred}) = 1 - \frac{\sum\limits_{i=1}^{\ell}\left(y_{true}^i - y_{pred}^i\right)^2}{\sum\limits_{i=1}^{\ell}\left(y_{true}^i - \overline{y_{true}}\right)^2},$$

где $\overline{y_{true}}$ — среднее истинное значение целевой переменной. Здесь в числителе стоит MSE, домноженное на ℓ, а в знаменателе — выборочная дисперсия выборки y_{true}, также домноженная на ℓ.

Заметим, что если положить y_{pred} тождественно равным среднему значению целевой переменной: $\overline{y_{true}}$, то R^2 становится равным 0. Из этого следует, что константная модель машинного обучения, которая предсказывает среднее значение целевой переменной независимо от объектов, имеет $R^2 = 0$.

При этом заметим, что R^2 может быть сколь угодно малым отрицательным числом. Такие модели неадекватны, поскольку результат их работы хуже константной модели.

Кроме того, очевидно, R^2 не превосходит 1. Таким образом, значения метрики для “адекватных” моделей находятся в промежутке $[0, 1]$. Для приемлемых моделей предполагается, что коэффициент детерминации должен быть не менее 0.5. Модели с коэффициентом детерминации выше 0.8 считаются весьма хорошими. Значение коэффициента детерминации 1 означает функциональную зависимость между переменными.