MAE и квантильная регрессия

28 сентября 2021 г.

1 MAE

Решим задачку:

$$MAE(C) = \frac{1}{\ell}\sum_{i=1}^{\ell}|C - y_i| \to \min_C$$

Покажем, что минимум MAE достигается при $C = median(y_1, \ldots, y_\ell) = m$. Другими словами, наша задача показать, что при любом C, отличном от m, мы получим $MAE(C) \geqslant MAE(m)$. Рассмотрим $C < m$. Распишем разность $MAE(C) - MAE(m)$:

$$|y_i - C| - |y_i - m| = \begin{cases} C - m, y_i < m \\ -(C + m - 2y_i), C \leqslant y_i \leqslant m \\ -(C - m), y_i > m \end{cases}$$

$$|y_i - C| - |y_i - m| \geqslant -(C - m) + 2(C - m)[y_i \leqslant m]$$

Суммируем по i:

$$\ell MAE(C) - \ell MAE(m) \geqslant -\ell(C - m) + 2(C - m)\sum_{i=1}^{\ell}[y_i \leqslant m]$$

Так как m — медиана, $\sum_{i=1}^{\ell}[y_i \leqslant m] \geqslant \frac{\ell}{2}$. Тогда

$$\ell MAE(C) - \ell MAE(m) \geqslant -\ell(C - m) + 2(C - m)\frac{\ell}{2} = 0.$$

Итак, для $C < m$ выполняется $MAE(C) \geqslant MAE(m)$. Аналогично показывается, что при $C > m$ выполняется $MAE(C) \geqslant MAE(m)$.

2 Функция потерь квантильной регрессии

Задача квантильной регрессии с константным прогнозом записывается так:

$$Q_\tau(C) = \frac{1}{\ell}\sum_{i=1}^{\ell}\rho_\tau(y_i - C) \to \min_C,$$

$$\rho_\tau(x) = \begin{cases} \tau x, & x > 0, \\ (\tau - 1)x, & x \leqslant 0. \end{cases},$$

где $0 < \tau < 1$ фиксировано.

Покажем, что минимум квантильной регрессии достигается при $C = q_\tau$, где q_τ — τ-квантиль вектора $y = (y_1, \ldots, y_\ell)$. Again, рассмотрим случай $C < q_\tau$:

$$\rho_\tau(y_i - C) - \rho_\tau(y_i - q_\tau) = \begin{cases} (\tau - 1)(y_i - C) - (\tau - 1)(y_i - q_\tau), & y_i < C \\ \tau(y_i - C) - (\tau - 1)(y_i - q_\tau), & C \leq y_i \leq q_\tau \\ \tau(y_i - C) - \tau(y_i - q_\tau), & q_\tau < y_i \end{cases}$$

$$\rho_\tau(y_i - C) - \rho_\tau(y_i - q_\tau) = \begin{cases} \tau(q_\tau - C) - (q_\tau - C), & y_i < C \\ \tau(q_\tau - C) - (q_\tau - y_i), & C \leq y_i \leq q_\tau \\ \tau(q_\tau - C), & q_\tau < y_i, \quad q_\tau < y_i \end{cases}$$

$$\rho_\tau(y_i - C) - \rho_\tau(y_i - q_\tau) \geq \tau(q_\tau - C) - (q_\tau - C)\,[y_i \leq q_\tau]$$

Суммируем по i:

$$\ell Q_\tau(C) - \ell Q_\tau(q_\tau) \geq \ell\tau(q_\tau - C) - (q_\tau - C)\sum_{i=1}^{\ell}[y_i \leq q_\tau]$$

Так как q_τ — τ-квантиль, то $\sum_{i=1}^{\ell}[y_i \leqslant q_\tau] \geqslant \ell\tau$. Отсюда

$$\ell Q_\tau(C) - \ell Q_\tau(q_\tau) \geq \ell\tau(q_\tau - C) - (q_\tau - C)\ell\tau = 0$$

Аналогично показывается для случая $C < q_\tau$.

3 Матчасть

На днях Давид Дале в своем канале очень круто описал практические use-кейсы квантильной регресси (https://t.me/matchast/267). Привожу кусочек для потомков :)

Когда использовать квантильную регрессиию?

1. Ваша целевая метрика — MAE, а не $RMSE$ или R^2.

2. В данных есть выбросы, и вы не хотите, чтобы они слишком влияли на результат.

3. Вам важнее правильно предсказать медиану, чем среднее арифметическое.

4. Большие и маленькие ошибки одинаково важны: например, одна ошибка в 300 рублей для вас не более плачевна, чем три ошибки в 100 рублей.

5. Важность ошибок несимметричная, например, ошибка -100 гораздо хуже, чем ошибка +100. Тогда вам может быть полезно предсказывать квантиль, отличную от 50%.

6. Вы хотите доверительный интервал для вашего предсказания, но не хотите завязываться на допущение, что ошибки распределены нормально с одинаковой дисперсией. Тогда вы можете просто предсказать, например, 5% и 95% квантили отдельными формулами.

7. Ваши данные гетероскедастичные, т.е. дисперсия ошибок в разных частях выборки разная. И при этом вы хотите, чтобы модель одинаково усердно старалась предсказывать и в зонах высокой дисперсии, и в зонах низкой.