ДЗ №1

Даниил Дмитриев, 494

9 марта 2017 г.

1 Задание 1

Формулы для матожидания ошибок в обоих случаях:

$$E_1 = (y - \overline{y})^2 = (y - \frac{1}{n}\sum_{i=1}^{n} y_i)^2, E_2 = \frac{1}{n}\sum_{i=1}^{n}(y - y_i)^2$$

$$E_1 = y^2 + \frac{1}{n^2}\left(\sum_{i=1}^{n} y_i\right)^2 - \frac{2}{n}y\sum_{i=1}^{n} y_i, E_2 = \frac{1}{n}\sum_{i=1}^{n}(y^2 + y_i^2 - 2yy_i)$$

После сокращения получаем:

$$E_1 : \left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2$$

$$E_2 : \frac{\sum_{i=1}^{n} y_i^2}{n}$$

Так как квадрат суммы не больше суммы квадратов, получаем, что

$$E_1 \leq \frac{E_2}{n}$$

. То есть отвечать средним выгоднее

2 Задание 2

По тому, как мы разделяем выборку на две, мы минимизируем ошибку при условии, что будем отвечать константой - средним значением по выборке. То есть регрессия в среднем не должна давать прирост, так как в данных не обязательно есть линейная зависимость (можем случайно переобучиться к тому же). Чтобы она была, можно при построении разбиения оценивать функцию $H(R)$ не как MSE для случая, когда мы отвечаем средним, а строить на каждом из разбиений линейную регрессию, и получать функцию оттуда. То есть мы будем строить много линейный регрессий при постройке дерева. В данном случае линейная регрессия на таргетах должна работать хорошо, но постройка дерева будет занимать больше времени.