Задание 2. Теория

Александр Захаров, 494

8 марта 2017 г.

1 Задача 1.1 Ответы в листьях регрессионного дерева

У элементов, пришедших в лист, значения целевой переменной: $\{y_1...y_n\}$

Найдем матожидание MSE в обоих случаях

1.1 Ответ $\overline{y}$:

$$\mathbb{E}MSE_0 = \frac{\sum\limits_{i=1}^{n}(y_i-\overline{y})^2}{n} = \sum_{i=1}^{n}\frac{y_i^2}{n} - \sum_{i=1}^{n}\frac{2y_i\sum\limits_{j=1}^{n}y_j}{n^2} + \left(\frac{\sum\limits_{i=1}^{n}y_i}{n}\right)^2 = \sum_{i=1}^{n}\frac{y_i^2}{n} - \left(\frac{\sum\limits_{i=1}^{n}y_i}{n}\right)^2 = \sum_{i=1}^{n}\frac{y_i^2}{n} - \sum_{i=1}^{n}\frac{y_i^2}{n^2} - 2\frac{\sum\limits_{i\neq j}y_iy_j}{n^2}$$

1.2 Отвечаем равномерно из пришедших ответов

То есть мы отвечаем y_i с вероятностью $\frac{1}{n}$

$$\mathbb{E}MSE_1 = \frac{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}(y_i-y_j)^2}{n^2} = 2\left(\frac{\sum\limits_{i=1}^{n}y_i^2}{n} - \frac{\sum\limits_{i\neq j}y_iy_j}{n^2}\right)$$

Таким образом $\mathbb{E}MSE_0 \leq \mathbb{E}MSE_1$. Значит, в среднем отвечать $\overline{y}$ лучше.

2 Задача 1.2 Линейные модели в деревьях

Построение линейных моделей в листах решаещего дерева не дает какого-либо значительного прироста качества, так как при обучении этого дерева мы минизиравали в листе квадрат отклонение от среднего и не предполагали никакой линейной зависимоти в ответах. Таким образом скорее все все коэффициенты кроме коэффициента перед косնстантой в линей модели будут близки к 0.

Для того, чтобы обучение линейны моделей в листах имело смысл, необходимо минимизировать не MSE при выборе разбиения, а обучать линейные модели на части разбитых данных и минимизировать MSE относительно предсказаний линейных моделей. Но такое построение решающего дерева будет происходить гораздо медленнее, чем стандартный способ построения без использования линейных моделей.

3 Задача 1.3 Unsupervised decision tree