Осипов Николай 492

Теоретические задачи.

① Покажем, что выбирать среднее значение таргета на объектах обучающей выборки лучше для MSE.

Посчитаем ошибку в обоих случаях:

$$E\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 = \frac{1}{n}E\sum_{i=1}^{n}y_i^2 - \frac{2}{n}Ey_i\bar{y} + \bar{y}^2 =$$

$$= Ey_1^2 - \bar{y}^2$$

$$E\frac{1}{n}\sum_{i=1}^{n}(y_i-y^*)^2 = Ey_1^2 - \frac{2}{n}\sum_{i=1}^{n}Ey_iy^* + \frac{2}{n}Ey^{*2} +$$

$$+ Ey^{*2} = Ey_1^2 - 2\frac{n-1}{n}\bar{y}^2 + \frac{n-2}{n}y^{*2} =$$

$$= Ey_1^2 + \frac{n-2}{n}Dy^{*2} - \bar{y}^2 \Rightarrow$$

Выбирая таргет для случайного объекта, получаем, что ошибка будет увеличиваться на $Dy^{*2}\cdot\frac{n-2}{n}$

Задача

(2)

$$G(j,t) = \frac{|L|}{|Q|} H(L) + \frac{|R|}{|Q|} H(R) \to \min_{j,t}$$

$$H(x) = \sum_{i=1} (x_i - \hat{x}_i)^2$$

Предикат листа имеет вид feature<threshold

=> алгоритм ищет разбиение, которые хорошо описываются кривой y = const

Для улучшения работы надо в качестве ф-ии H взять ошибку линейной регрессии и вместо порога подбирать коэффициенты a, b нашей линейной модели

$y = ax + b$.

(3) Ищем энтропию многомерного нормального распределения.

$$p(x) = e^{-\frac{\frac{1}{2}(\bar{x}-\bar{\mu})}{\Sigma(\bar{x}-\bar{\mu})}} \cdot \frac{1}{\sqrt{(2\pi)|\Sigma|}}$$

$$H(x) = -\int p(x) \ln p(x)\, dx =$$

$$= -\int p(x) \left[-\frac{1}{2} \ln (2\pi)^n |\Sigma| + \left(-\frac{1}{2}\right)(x-\mu)^T \Sigma^{-1} (x-\mu) \right] dx =$$

$$= \frac{1}{2} \ln (2\pi)^n |\Sigma| + \frac{1}{2} \int p(x) \left((\bar{x}-\bar{\mu})^T \Sigma^{-1} (x-\mu) \right) dx \ominus$$

$$tr\, (\bar{x}-\bar{\mu})^T \Sigma^{-1} (\bar{x}-\bar{\mu}) =$$

$$= tr\, \Sigma^{-1} (\bar{x}-\bar{\mu})^T (\bar{x}-\bar{\mu}) = tr\, E = n =$$

$$= (\bar{x}-\bar{\mu})^T \Sigma^{-1} (\bar{x}-\bar{\mu}) \Rightarrow$$

$$\ominus\ \frac{1}{2} \ln (2\pi)^n |\Sigma| + n/2$$

$$\blacksquare$$