Теоретические задачи

1.1 Ответы в листьях регрессионного дерева

Рассмотрим некоторый лист дерева. Пусть в него попали объекты $y_1, ...y_n$. Посчитаем мат-ожидание MSE для некоторого объекта y в первом и во втором случае:

$$MSE = E(y - \hat{y})^2 = Ey^2 + E\hat{y}^2 - 2Ey \cdot \hat{y} = Ey^2 + E\hat{y}^2 - 2Ey \cdot E\hat{y}$$

Значит, разность ошибок:

$err_1 - err_2 = E\hat{y_1}^2 - E\hat{y_2}^2 - 2Ey \cdot (E\hat{y_1} - E\hat{y_2})$

1) $\hat{y_1} = \frac{\sum_{i=1}^n y_i}{n}$

$E\hat{y_1} = \frac{\sum_{i=1}^n Ey_i}{n} = \frac{\sum_{i=1}^n y_i}{n}$

$E\hat{y_1}^2 = E(\frac{\sum_{i=1}^n y_i}{n})^2$

2)$E\hat{y_2} = \frac{\sum_{i=1}^n y_i}{n}$

$E\hat{y_2}^2 = \frac{\sum_{i=1}^n y_i^2}{n}$

Разность ошибок:

$err_1 - err_2 = E(\frac{\sum_{i=1}^n y_i}{n})^2 - \frac{\sum_{i=1}^n y_i^2}{n} - 2Ey \cdot (\frac{\sum_{i=1}^n y_i}{n} - \frac{\sum_{i=1}^n y_i}{n}) \leq 0$

Использовано неравенство Коши-Буняковского$((\sum_{i=1}^n y_i)^2 \leq (\sum_{i=1}^n y_i^2) \cdot (\sum_{i=1}^n 1^2) = n \cdot \sum_{i=1}^n y_i^2)$

1.3 Unsupervised decision tree

Энтропия непрерывного распределения выражается через его плотность по формуле: $-\int ... \int f(x) \cdot ln(f(x)$, где $f(x)$ - плотность распределения. Плотность многомерного нормального распределения: $f(x) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}(x-a)^T \Sigma^{-1}(x-a)}$.

Заметим, что $\frac{1}{2}(x-a)^T \Sigma^{-1}(x-a) = \sum_i \sum_j (x_i - a_i) \cdot \Sigma_{i,j} \cdot (x_j - a_j)$(в силу произведения матриц).

Найдем энтропию нормального распределения:

$H = -\int ... \int f(x) \cdot (-\frac{1}{2}(x-a)^T \Sigma^{-1}(x-a) - \frac{1}{2} \cdot ln((2\pi)^n |\Sigma|) = \frac{1}{2} E(\sum_i \sum_j (x_i - a_i) \cdot \Sigma_{i,j}^{-1} \cdot (x_j - a_j)) +$
$\frac{1}{2} \cdot ln((2\pi)^n)|\Sigma| = \frac{1}{2} \sum_i \sum_j \Sigma_{i,j}^{-1} \cdot E(x_i - Ex_i)(x_j - Ex_j) + \frac{1}{2} \cdot ln((2\pi)^n)|\Sigma| = \frac{1}{2} \sum_i \sum_j \Sigma_{i,j}^{-1} \cdot \Sigma i, j + \frac{1}{2} \cdot$
$ln((2\pi)^n)|\Sigma| = \frac{1}{2} \sum_i \sum_j \Sigma_{i,j}^{-1} \cdot \Sigma j, i + \frac{1}{2} \cdot ln((2\pi)^n)|\Sigma| = \frac{1}{2} \sum_i E_{i,i} + \frac{1}{2} \cdot ln((2\pi)^n)|\Sigma| = \frac{1}{2} \cdot ln((2\pi \cdot e)^n)|\Sigma|$