Листья регрессионного дерева

1) отвечаем средним:

$$\frac{1}{n}\sum_{i=1}^{n}(a(x_i)-y)^2 \to min$$

$$a(x_i) = \sum_{i=1}^{n}\frac{y_i}{n} = E(y)$$

$$E\left(\frac{1}{n}\sum_{i=1}^{n}(a(x_i)-y)^2\right) = D(y)$$

дисперсия y как по случайной величине равномерно распределенной по ответам в листе

2) отвечаем случайно

$a(x_i) = y_j$ j - произвольный $j \in \overline{1,n}$

$$E\left(\sum_{i=1}^{n}\frac{1}{n}(a(x_i)-y_i)^2\right) = \sum_{i=1}^{n}\frac{1}{n}E(a^2(x_i)) - 2y_i E(a(x_i)) + y_i^2) = \sum_{i=1}^{n}\frac{1}{n}\left[y_i^2 - 2y_i\sum_{j=1}^{n}\frac{y_j}{n} + E(a(x_i))^2\right] =$$

$$D(a(x_i)) = E(a(x_i))^2 - (E(a(x_i)))^2$$

$$E(a(x_i))^2 = D(a(x_i)) + (E(a(x_i)))^2$$

$$= \sum_{i=1}^{n}\frac{1}{n}\left[\left(y_i - \sum_{j=1}^{n}\frac{y_j}{n}\right)^2 + D(a(x_i))\right] =$$

$$= \sum_{i=1}^{n} D(a) + D(y) + D(a(x))$$

NOTES

Т. е. предпочтительней выбирать среднее, чем случайное (мат. ожидание ошибки ниже)