Задача. *1.1 Ответы в листах регрессионного дерева*

Proof. Пусть в листе находятся равновероятные объекты со значениями целевого признака $a_1, \ldots, a_n$. Тогда, обозначив E_1 и E_2 матожидания ошибки в случае ответа средним значением $\overline{a}$ и случайным значением a_r соответственно, имеем:

$$E_1 = E(a-\overline{a})^2 = Ea^2 + \overline{a}^2 - 2\overline{a}Ea = Ea^2 + \overline{a}^2 - 2\overline{a}^2 = Ea^2 - (Ea)^2 \leq$$

$$(Ea^2-(Ea)^2)+(Ea^2-(Ea)^2) = Ea^2+Ea^2-2EaEa_r = Ea^2+Ea_r^2-2Eaa_r = E(a-a_r)^2 = E_2$$

Здесь мы пользуеся тем, что $Ea^2 > (Ea)^2$. Окончательно получаем $E_1 > E_2$, то есть лучше отвечать средним. □

Задача. *1.2 Линейные модели в деревьях*

Возможно проблема в том, что дерево разбивает пространство на маленькие области, так что регрессия будет давать большую погрешность. Идея: на каждом разбиении делать регрессию в половинках и выбрать такое, где ошибка наименьшая

Задача. *1.3 Unsupervised decision tree*

Proof. Рассмотрим многомерное нормальное распределение

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)}$$

$$H(f) = \int f \ln f = \int f(\mathbf{x}) \left(\ln\left((2\pi)^{n/2}|\Sigma|^{1/2}\right) + \frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu) \right) =$$

$$E\left(\ln\left((2\pi)^{n/2}|\Sigma|^{1/2}\right) + \frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu) \right) =$$

$$E\left(\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu) \right) + \ln\left((2\pi)^{n/2}|\Sigma|^{1/2}\right) =$$

$$\frac{1}{2}\left(E\left((\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right) + \ln\left((2\pi)^n|\Sigma|\right) \right) =$$

$$\frac{1}{2}\left(n + \ln\left((2\pi)^n|\Sigma|\right) \right) =$$

$$\frac{1}{2}\ln\left((2\pi e)^n|\Sigma|\right) =$$

□