1 Теоретические задачи

1.1 Ответы в листьях регрессионного дерева

Обозначения:

$$MSE(X, Y) = \sum_{y_i \in Y} (\xi_i - y_i)^2$$

Где таргеты в листе:

$$X = \{x_i\}_i^n$$

тестовые таргеты

$$Y = \{y_i\}_i^m$$

$\xi_i \sim U(X)$ случайная величина - таргет случайного объекта из листа (из условия задачи).

$$p_{ij} = Pr(\xi_i = x_j) = \frac{1}{n}$$

Тогда матожидание ошибки по MSE можно посчитать как:

$$\mathbb{E}MSE = \sum_{k=1}^{n} \mathbb{E}(MSE(X, Y)|X, Y) \cdot I(X, Y)$$

То есть если мы посчтаем эти суммы для обоих случаев и сравним слагаемые (что я и сделаю дальше), то мы получим искомое неравенство. В случае сэмплирования ответов из таргета имеем:

$$\mathbb{E}(\sum_{i=1}^{m} (\xi_i - y_i)^2 | X, Y) =$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} (\xi_i - y_i)^2 =$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} \xi^2 - 2 p_{ij} \xi_i y_i + p_{ij} y_i^2 =$$

$$= \textstyle\sum_{i=1}^{m} (\sum_{j=1}^{n} \frac{x_j^2}{n} - \sum_{j=1}^{n} \frac{2}{n} x_j y_i + y_i^2) \ (*)$$

В случае выдачи констного среднего ответа по таргетам в листе:

$$\textstyle\sum_{i=1}^{m} (\displaystyle\frac{\sum_{j=1}^{n} x_j}{n} - y_i)^2 =$$

$$\textstyle\sum_{i=1}^{m} ((\sum_{j=1}^{n} \frac{x_j}{n})^2 - \sum_{j=1}^{n} \frac{2}{n} x_j y_i + y_i^2) \ (**)$$

Получаем (*) $>$ (**), поскольку:

$$\sum_{j=1}^{n} \frac{x_j^2}{n} > (\sum_{j=1}^{n} \frac{x_j}{n})^2$$

1.2 Линейные модели в деревьях Критерий построения разбиений в регрессионном дереве никак не учитывает то, насколько в каждой из дочерних ветвей зависимость близка к линейной.(можно взять точки на прямой с большим расстоянием друг от друга). Тогда в качестве меры "хорошести"(было неоднородности") множества можно использовать, например, MSE модели $a(x)$ обученной на данном множестве, те.:

$$H(R) = \frac{1}{|R|} \sum_{x_j \in R} (y_i - a(x_i))^2$$

и подставить этот $H(R)$ в критерий разбиения(в обозначениях лекции):

$$\frac{|L|}{|Q|} H(L) + \frac{|R|}{|Q|} H(R)$$

1.3 unsupervised dicision tree

По определению

$$H(f) = -\int_{x \in \mathbb{R}^n} f(x) \ln(f(x)) dx =$$

$$= -\mathbb{E} \ln(f(x))$$

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

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

$$\mathbb{E} \ln(f(x)) = -\ln\left((2\pi)^{\frac{1}{2}} |\Sigma|^{\frac{1}{2}} \right) - \frac{1}{2} \mathbb{E}(x-\mu)^T |\Sigma|^{-1} (x-\mu) =$$

Где

$$\frac{1}{2} \mathbb{E}(x-\mu)^T |\Sigma|^{-1} (x-\mu) = -\frac{1}{2} tr(\Sigma \Sigma^{-1}) = -\frac{1}{2} n$$

Итого

$$H(f) = -\mathbb{E} \ln f(x) = \frac{n}{2} + \ln\left((2\pi)^{\frac{n}{2}} \right) |\Sigma|^{\frac{1}{2}} =$$

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

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