# Машинное обучение: задание 2

Болотин Пётр

Март, 2017

## 0.1 Задача 1

Лучше брать среднее.

*Доказательство.* Вследствие линейности мат.ожидания, достаточно доказать, что мат.ожидание ошибки будет меньше на каком-то произвольном листе, в котором находится $k$ объектов. Далее, предполагая, что все величины одинаково распределены:

$$E\sum_{i=1}^{k}(y_i-\overline{y})^2 = k(Ey_1^2 - 2E(\overline{y}y_1) + E\overline{y}^2) \tag{1}$$

$$E\sum_{i=1}^{k}(y_i-y^*)^2 = k(Ey_1^2 - 2E(y^*y_1) + E(y^*)^2) \tag{2}$$

$y^*$ обозначает случайно взятый элемент листа. Теперь нужно показать, что (1) $\leq$ (2) Первые слагаемые очевидно совпадают. Рассмотрим вторые, учтя, что выбираем равновероятно. $E(y^*y_1) = E(y_1\frac{1}{k}\sum_{i=1}^{k} y_i) = E(\overline{y}y_1)$ Осталось равнить третьи слагаемые.

$$kE\overline{y}^2 = \frac{1}{k}E(\sum_{i=1}^{k}y_i^2+\sum_{i\neq j}y_iy_j) = Ey_1^2+(k-1)Ey_1^2 = Ey_1^2-(Ey_1)^2+k(Ey_1)^2 \tag{3}$$

$$kE(y^*)^2 = E\sum_{i=1}^{k}y_i^2 = kEy_1^2 \tag{4}$$

$(3)-(4) = \sigma - k\sigma = \sigma(1-k)$ Где $\sigma$ это дисперсия случайной величины, то есть при $k > 1$ получаем, что ошибка при выборе среднего строго меньше. $\square$

## 0.2 Задача 2

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

Как правило, в задачах регрессии среднеквадратичное отклонение от среднго используется в качестве функции $H()$, поэтому и нет результата, ведь по сути алгоритм учится находить те разбиения, которые хорошо описываются прямой вида $y = const$, чтобы результат был, этот функцию $H()$ нужно заменить на ошибку при оценке линейной регрессией и подбирать не порог, а параметры $a, b$ уравнения $y = ax + b$

## 0.3 Задача 3

$\int_{R^n} f(x)lnf(x) = \frac{1}{2}E((x-\mu)^T\Sigma^{-1}(x-\mu)) + ln((2\pi)^{n/2}|\Sigma|^{1/2})$ Рассмотрим мат.ожиданание. $E(x-\mu)^T\Sigma^{-1}(x-\mu) = E\sum_{j,k}(x-\mu)_j(x-\mu)_k\Sigma_{j,k}^{-1} = n$ Потому что $(x-\mu)_j(x-\mu)_k = \Sigma_{j,k}$ Далее $ln((2\pi)^{n/2}|\Sigma|^{1/2}) + \frac{n}{2} = H(S)$