

HW2

Линейное регрессии

(N) Какое спрашивается изложение в листочке регрессионного дерева  
приводит к наименшему критерию ошибки по MSE:

- среднее значение target-а на листах обучающей выборки, попавших в лист
- target имеющийся обозначаючиый лист (среднее листа обучающих равновероятно)

$$\text{tk} \quad E\left(\sum_{i=1}^k (y_i - \bar{y})^2\right) \leftarrow \text{где среднее значение}$$

$$\cdot E\left(\sum_{i=1}^k (y_i - \tilde{y})^2\right) \leftarrow \text{где ожидаемое значение листа}$$

$$E\left(\sum_{i=1}^k (y_i - \bar{y})^2\right) = E\left(\sum_{i=1}^k (y_i - \tilde{y})^2\right) \Leftrightarrow \begin{array}{l} \text{сумма квадратов} \\ \text{ошибок} \end{array}$$

$$1) E\left(\sum_{i=1}^k (y_i - \bar{y})^2\right) = k E(y_1)^2 - 2E(y_1) + E(\bar{y})^2 = k E(y_1)^2 - k E(\bar{y})^2,$$

$$\text{т.к. } 2E(\bar{y} y_1) = 2E(\bar{y})^2 \quad \forall i \quad E y_i = \frac{\sum_{i=1}^k y_i}{k} = \bar{y} \quad \begin{array}{l} \text{исходящий} \\ \text{равноделенный} \\ \text{распределение} \end{array}$$

$$2) E\left(\sum_{i=1}^k (y_i - \tilde{y})^2\right) = k E(y_1)^2 - 2E\sum_{i \neq j} y_i y_j - 2E y_k^2 + k E \bar{y}^2 = \\ = k E(y_1)^2 - 2(k-1) E y_1^2 + (k-2) E(\bar{y})^2 = \\ = / \quad \bar{y} y_1 = E(\bar{y})^2 - (E y_1)^2 \quad \Rightarrow \quad k E(y_1)^2 + (k-2) \bar{y} y_1 + E(\bar{y})^2 (k-2-2k+2) = \\ = k E(y_1)^2 + (k-2) \bar{y} y_1 + E(\bar{y})^2 (k)$$

$$\Leftrightarrow k E(y_1)^2 - k E(\bar{y})^2 - k E(y_1)^2 - (k-2) \bar{y} y_1 + k E(\bar{y})^2 = (2-k) \bar{y} y_1$$

видно, что выбор среднего лучше, чем выбор ожидаемого.  
Сумма равняется  $(k-2) \bar{y}$ .

(N3) из лекции мы знаем, что линейная модель в регрессионном анализе

$$G(\hat{y}_i | t) = \min_{L, R} \left\{ \frac{L}{Q} K(L) + \frac{R}{Q} H(R) \right\}$$

и в таких задачах обычно  $H(\cdot)$  минимизирует среднеквадратичное отклонение (смисл N3)  $KSE \approx H(\Sigma) = \sum_i (x_i - \hat{x}_i)^2$

Статистика Волдса является константным ответом в линейной модели линейной регрессии, единственной на отрезке от  $t$  до  $t+1$ , не давая оптимального приближения, т.к. алгоритм регрессии так раз подбирает так, чтобы подогнать к функции  $y = \text{const}$ , что по сути, и есть константный ответ.

Если же это значение функции  $H(\cdot)$  является среднеквадратичного отклонения на отрезок при линейной модели линейной задачи, тогда надо пытаться найти ее и если, то некоторую линейную функцию  $y = ax + b$ . Если такого  $x^{(i)}$  подбирают коэффициенты  $a$  и  $b$  для линейной модели и брали ошибку  $KSE$ , то можно увидеть зависимость

(N3) линейная модель  $S$ :  $K(S) = \frac{1}{2} \ln((2\pi e)^n |\Sigma|)$   
 (скомбинировавшись с предыдущими) матрица ковариаций остатков распределения нормально)  
 линейных многочленов нормального распределения

$$K(x) = \int f(x) \ln(f(x)) dx \quad (1)$$

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

$$(2) - \int_{\mathbb{R}^n} f(x) \left( -\frac{1}{2} \ln((2\pi)^n |\Sigma|) \right) dx - \int_{\mathbb{R}^n} f(x) \left( -\frac{1}{2} (x-\bar{x})^\top \Sigma^{-1} (x-\bar{x}) \right) dx =$$

$$= \frac{1}{2} \ln((2\pi)^n |\Sigma|) + \frac{1}{2} \int_{\mathbb{R}^n} f(x) ((x-\bar{x})^\top \Sigma^{-1} (x-\bar{x})) dx$$

$$\Sigma = \text{cov}(x, x) = E((x - Ex)(x - Ex)) \Rightarrow \Sigma^{-1} \frac{1}{2} \Sigma = \frac{1}{2}$$

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