Задание 2. Теория

Боймель Александр

8 марта 2017 г.

I. Ответы в листьях регрессионного дерева.
Пусть в листе, в который попал новый объект, находятся k объектов со значениями целевой переменной $\{y_i\}_{i=1}^k$. Ответ на попавшем объекте y.

Тогда, посчитаем матожидание ошибки, предсказывая среднее значение по листу.

$$\begin{aligned} E(MSE_{\overline{y}}) =& E\left(y - \frac{\sum_{i=1}^k y_i}{k}\right)^2 = \\ =& Ey^2 + k \cdot E(\frac{y_1}{k})^2 - 2 \cdot k \cdot Ey \cdot E\frac{y_1}{k} + k \cdot (k-1) \cdot (E\frac{y_1}{k})^2 = \\ =& Ey^2 - 2 \cdot Ey \cdot Ey_1 + \frac{Ey_1^2}{k} + (1 - \frac{1}{k}) \cdot (Ey_1)^2 \end{aligned}$$

Теперь посмотрим на матожидание ошибки отвечая значением на случайном объекте из листа(считая все объекты равновероятными).

$$\begin{aligned} E(MSE_{rand}) =& E(y - y_i)^2 = \\ =& Ey^2 + Ey_1^2 - 2 \cdot Ey \cdot Ey_1 \end{aligned}$$

где $i \in \{1, \cdots k\}$, но так как все равновероятны, то $i \to 1$ Сравним их:

$$E(MSE_{rand} - MSE_{\overline{y}}) = (1 - \frac{1}{k}) \cdot (Ey_1^2 - (Ey_1)^2) = (1 - \frac{1}{k}) \cdot Dy_1 > 0$$

Таким образом, получается, что во втором случае в среднем ошибка больше, значит лучше предсказывать среднее по объектам листа.

II. Линейный модели в деревьях

III. Unsupervised decision tree Покажем, что $H(S) = \frac{1}{2} \cdot \ln((2 \cdot \pi \cdot e)^n \cdot |\Sigma|)$ - энтропия многомерного нормального распределения.

Плотность многомерного нормального распределения имеет вид $p(x) = \frac{1}{(2\cdot\pi)^{\frac{n}{2}} \cdot \sqrt[2]{|\Sigma|}} \cdot e^{-\frac{1}{2}\cdot(x-\mu)^T\cdot\Sigma^{-1}\cdot(x-\mu)}$

$$
\begin{aligned}
H(p) = &- \int \cdots \int_{R^n} p(x) \cdot \ln p(x)\, dx = \\
&= \int \cdots \int_{R^n} p(x) \cdot \left(\frac{1}{2} \cdot (x-\mu)^T \cdot \Sigma^{-1} \cdot (x-\mu) + \ln((2\cdot\pi)^{\frac{n}{2}} \cdot \sqrt[2]{|\Sigma|}) \right) dx = \\
&= \frac{1}{2} \cdot E\left(\sum_{i,j} (x_i - \mu_i) \cdot (\Sigma^{-1})_{i,j} \cdot (x_j - \mu_j) \right) + \frac{1}{2} \cdot \ln((2\cdot\pi)^n \cdot |\Sigma|) = \\
&= \frac{1}{2} \cdot \sum_{i,j} \left(E((x_i - \mu_i) \cdot (\Sigma^{-1})_{i,j} \cdot (x_j - \mu_j)) \right) + \frac{1}{2} \cdot \ln((2\cdot\pi)^n \cdot |\Sigma|) = \\
&= \frac{1}{2} \cdot \sum_i \sum_j (\Sigma)_{i,j} \cdot (\Sigma^{-1})_{i,j} + \frac{1}{2} \cdot \ln((2\cdot\pi)^n \cdot |\Sigma|) = \\
&= \frac{1}{2} \cdot \sum_i (\Sigma \cdot \Sigma^{-1})_{i,i} + \frac{1}{2} \cdot \ln((2\cdot\pi)^n \cdot |\Sigma|) = \\
&= \frac{1}{2} \cdot \sum_i (E)_{i,i} + \frac{1}{2} \cdot \ln((2\cdot\pi)^n \cdot |\Sigma|) = \\
&= \frac{n}{2} + \frac{1}{2} \cdot \ln((2\cdot\pi)^n \cdot |\Sigma|) = \\
&= \frac{1}{2} \cdot \ln((2\cdot\pi\cdot e)^n \cdot |\Sigma|) =
\end{aligned}
$$

Ч.Т.Д.