Машинное обучение. Задание 2

Сопильняк Ольга

9 марта 2017 г.

1.1 Ответы в листьях регрессионного дерева

Пусть $X_1, \ldots, X_n$ — таргеты объектов обучающей выборки.

Рассмотрим первый случай. Пусть сначала $\overline{X} = \frac{X_1+\cdots+X_n}{n}$ — средее значение таргета, тогда

$$EMSE(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n}(\overline{X} - X_i)^2 = \frac{n\overline{X}^2 - 2\overline{X}\sum X_i + \sum X_i^2}{n}$$

Теперь пусть Y — таргет случайного объекта. Посчитаем матожидание:

$$EMSE(Y) = \frac{1}{n}E\sum_{i=1}^{n}(Y - X_i)^2 = \frac{nEY^2 - 2\sum X_i EY + \sum X_i^2}{n}$$

Посмотрим на разность этих матожиданий и после некоторых преобразований получим:

$$EMSE(Y) - EMSE(\overline{X}) = \overline{X^2} - (\overline{X})^2 \geqslant 0$$

Поэтому среднее значение таргета лучше.

1.3 Unsupervised Decision Tree

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

— многомерное нормальное распределение.

Его энтропия:

$$H(f) = -\int \cdots \int f(x) \ln f(x) dx$$

Распишем:

$$\int \cdots \int f(x)(\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) + \ln(2\pi)^{n/2}|\Sigma|^{1/2}))dx$$

$$\frac{1}{2}E(\sum(x_i - \mu_i)(\Sigma^{-1})(x_j - \mu_j)) + \frac{1}{2}\ln((2\pi)^n|\Sigma|)$$

Первое слагаемое равно $\frac{n}{2}$, поэтому окончательно получаем:

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