Train In Data

Smoothing
Category encoders

Encoding logic

The encoding values are determined by a mixture of probabilities:

- The posterior → target mean per category
- The prior → target mean for the entire dataset

- In short, when there are few observations in some categories.

Encoding logic

The two probabilities are then "blended" using a weighting factor that is a function of the sample size:

$$\text{Value} = \boldsymbol{\lambda} \times \text{posterior} + (1-\boldsymbol{\lambda}) \times \text{prior}$$

Encoding logic

$$\lambda = \frac{1}{1 + e^{\frac{-(n-k)}{f}}}$$

n: number of observations per category

k: our trust

f: controls the rate of transition from prior to posterior

Encoding logic

Value = $\lambda \times \text{posterior} + (1-\lambda) \times \text{prior}$

$$\lambda = \frac{1}{1 + e^{\frac{-(n-k)}{f}}}$$

n	k	f
10	5	2
5	5	2
1	5	2

(n-k)/f	e	lambda
2.50	0.08	0.92
0.00	1.00	0.50
-2.00	7.39	0.12

Encoding logic

The two probabilities are then "blended" using a weighting factor that is a function of the sample size:

$$\text{Value} = \boldsymbol{\lambda} \times \text{posterior} + (1 - \boldsymbol{\lambda}) \times \text{prior}$$

- The more observations, the more we trust the posterior

Encoding logic

$$\lambda = \frac{1}{1 + e^{\frac{-(n-k)}{f}}}$$

n: number of observations per category

k: our trust

f: controls the rate of transition from prior to posterior

Encoding logic

min_samples_leaf: int k

For regularization the weighted average between category mean and global mean is taken. The weight is an S-shaped curve between 0 and 1 with the number of samples for a category on the x-axis. The curve reaches 0.5 at min_samples_leaf. (parameter k in the original paper)

smoothing: float f

smoothing effect to balance categorical average vs prior. Higher value means stronger regularization. The value must be strictly bigger than 0. Higher values mean a flatter S-curve (see min_samples_leaf).

THANK YOU

www.trainindata.com