Train In Data

Mean encoding Smoothing

Original article:

Micci-Barreca D. “**A Preprocessing Scheme for High-Cardinality Categorical Attributes in Classification and Prediction Problems**”.

ACM SIGKDD Explorations Newsletter, 2001.

https://dl.acm.org/citation.cfm?id=507538

The encoding logic

Map individual values (categories) of a high-cardinality categorical variable to an estimate of the **probability** or the **expected value** of the dependent attribute (target)

The encoding logic: mean encoding?

Map individual values (categories) of a high-cardinality categorical variable to an estimate of the probability or the expected value of the dependent attribute (target)

- **Is the target mean value per category not a good estimate?**
  - **Yes, but…**

When is the target mean suitable?

If the number of observations per category is sufficiently large,

so that we can get a good, reliable measure of the target mean.

When is the target mean NOT suitable?

- When variables have high cardinality, or rare labels.

- In short, when there are few observations in some categories.
  - We can’t trust the target estimates in these cases.

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.

Implementations

Implementations of the blended probabilities

- Category encoders
- Feature-engine

Train In Data

THANK YOU

www.trainindata.com