TARGET ENCODING

- Developed to solve limitations of dummy encoding for high cardinality categorical features.

Goal: Each categorical feature x should be encoded in a single numeric feature $\tilde{x}$.

- Basic definition for regression by Micci-Barreca (2001):

$$\tilde{x} = \frac{\sum_{i:x=l} y^{(i)}}{N_l}, \quad l = 1, \ldots, k,$$

where N_l is the number of observations of the l'th level of feature x.

TARGET ENCODING - EXAMPLE

Foundation	BrkTil	CBlock	PConc	Slab	Stone	Wood
n	311	1244	1310	49	11	5

- Encoding for wooden foundation:

house.id	17	893	986	2898	2899
SalePrice	164000	145500	143000	250000	202000
Foundation	Wood	Wood	Wood	Wood	Wood

$$\frac{164000 + 145500 + 143000 + 250000 + 202000}{5} = 180900$$

TARGET ENCODING - EXAMPLE

- For all foundation types:

Foundation	BrkTil	CBlock	PConc	Slab	Stone	Wood
Foundation(enc)	128107	148284	227069	110458	149787	180900

This mapping is calculated on training data and later applied to test data.

TARGET ENCODING FOR CLASSIFICATION

- Extending encoding to binary classification is straightforward, instead of the average target value the relative frequency of the positive class is used
- Multi-class classification extends this by creating one feature for each target class in the same way as binary classification.

TARGET ENCODING - ISSUES

Problem: Target encoding can assign extreme values to rarely occurring levels.

Solution: Encoding as weighted sum between global average target value and encoding value of level.

$$\tilde{x} = \lambda_l \frac{\sum_{i:x=l} y^{(i)}}{N_l} + (1 - \lambda_l) \frac{\sum_{i=1}^{n} y^{(i)}}{n}, \quad l = 1, \ldots, k.$$

- λ_l can be parameterized and tuned, but optimally, tuning must be done for each feature and level separately (most likely infeasible!).
- Simple solution: Set $\lambda_l = \frac{N_l}{N_l + \epsilon}$ with regularization parameter ϵ.
- This shrinks small levels stronger to the global mean target value than large classes.

TARGET ENCODING - ISSUES

Problem: Label leakage! Information of $y^{(i)}$ is used to calculate $\tilde{x}$. This can cause overfitting issues, especially for rarely occurring classes.

Solution: Use internal cross-validation to calculate $\tilde{x}$.

- It is unclear how serious this problem is in practice.
- But: calculation of $\tilde{x}$ is very cheap, so it doesn't hurt.
- An alternative is to add some noise $\tilde{x}^{(i)} + N(0, \sigma_\epsilon)$ to the encoded samples.