Mean encodings

TOTAL POINTS 4

1. What can be an indicator of usefulness of mean encodings? 1 point

- (●) Categorical variables with lots of levels.
- ( ) Learning to rank task.
- ( ) A lot of binary variables.

2. What is the purpose of regularization in case of mean encodings? Select all that apply. 1 point

- [ ] Regularization allows to make feature space more sparse.
- [x] Regularization reduces target variable leakage during the construction of mean encodings.
- [x] Regularization allows us to better utilize mean encodings.

3. What is the correct way of validation when doing mean encodings? 1 point

- ○ Fix cross-validation split, use that split to calculate mean encodings with CV-loop regularization, use the same split to validate the model.
- ○ Calculate mean encodings on all train data, regularize them, then validate your model on random validation split.
- ◉ First split the data into train and validation, then estimate encodings on train, then apply them to validation, then validate the model on that split.

4. Suppose we have a data frame 'df' with categorical variable 'item_id' and target variable 'target'. 1 point

We create 2 different mean encodings:

1. via df['item_id_encoded1'] = df.groupby('item_id')['target'].transform('mean')
2. via OneHotEncoding item_id, fitting Linear Regression on one hot-encoded version of item_id and then calculating 'item_id_encoded2' as a prediction from this linear regression on the same data.

Select the true statement.

- (●) 'item_id_encoded1' and 'item_id_encoded2' will be essentially the same only if linear regression was fitted without a regularization.
- ( ) 'item_id_encoded1' and 'item_id_encoded2' may hugely vary due to rare categories.
- ( ) 'item_id_encoded1' and 'item_id_encoded2' will be essentially the same.