Mean Target Encoding

$\forall\ D = \{(x_i, y_i)\}_{i=1}^{n}$, $x_i \in \mathbb{R}^m$, $y_i \in \mathbb{R}$

where, for example, x_i – features describing user and ad, y_i – event (click or view)

Assumption

Records in D are ordered by time, i.e

$\forall i,j\ (i<j \Rightarrow y_i$ happend before $y_j)$

Let $k \in \{1,\dots,m\}$ be index of categorical feature.

We would like to replace $x_{i,k}$ with some real value.

MTE

$$\frac{\sum_{j=1}^{i-1} \mathbb{1}\{x_{j,k} = x_{i,k}\} \cdot y_j + \alpha P}{\sum_{j=1}^{i-1} \mathbb{1}\{x_{j,k} = x_{i,k}\} + \alpha} \qquad (1)$$

prevents overfitting

* We consider all records stricly before i with same value of k-th feature
* We do smoothing to avoid overfitting

P – prior knowlege, α – parameter.

For example, P – global CTR, α – initial views

Data Split

lets assume that we are given second dataset $T = \{x'_i\}_{i=1}^{t}$ without labels.

Q: How to replace $x'_{i,k}$ in T without y?

A: Assume that events in T happend after events in D then

$$x'_{i,k} \leftarrow \frac{\sum_{j=1}^{n} \mathbb{1}\{x_{j,k} = x'_{i,k}\}\, y_i + \alpha \cdot p}{\sum_{j=1}^{n} \mathbb{1}\{x_{j,k} = x'_{i,k}\} + \alpha} \qquad (2)$$

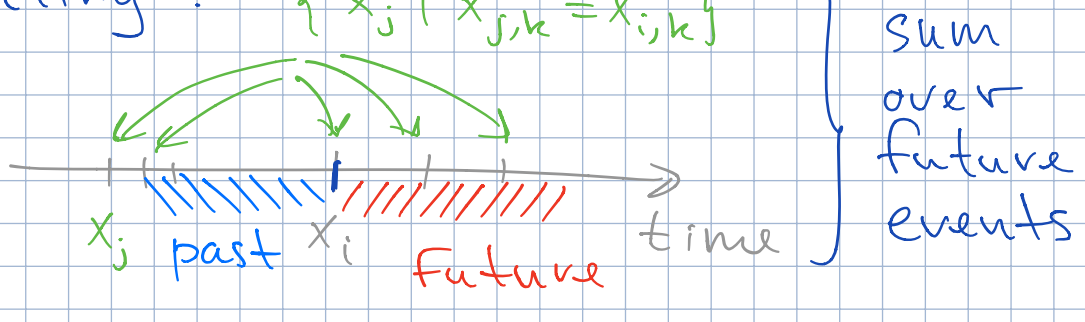
Sum over all records in D because D preceeds T!

Common Mistakes

(1)

Q: What will happen if in Eq. (1) instead of $(i-1)$ upper limit we will use n (sum over all dataset)?

A: Overfitting.



(2)

Q: How to implement and use MTE in Spark Pipeline

A:

1) Split D into Train / Val / Test by timestamp
2) Fit MTE on Train part
3) Transform Val and Test using Eq. (2)

For Train use Eq. (1) (not transform)

This approach will prevent you from looking into the future events.