Yandex

Yandex

Dealing with categorical features and overfitting

Nikita Kazeev[1 2] on behalf the CatBoost team

[1] National Research University Higher School of Economics (HSE) [2] Yandex

Categorical features

Categorical features

Image: www.petsworld.in

One-hot encoding

$$\text{[proton, pion, kaon]} \rightarrow [[1, 0, 0], [0, 1, 0], [0, 0, 1]]$$

One-hot encoding

[proton, pion, kaon] $\rightarrow [[1, 0, 0], [0, 1, 0], [0, 0, 1]]$

› Doesn’t scale well with the number of categories

CTR (aka click-through ratio)

For each pair (target_class, categorical_feature_value):

$$\text{ctr}_i = \frac{\text{countInClass} + \text{prior}}{\text{totalCount} + 1}$$

- countInClass – number of objects in the i-th class with the current categorical feature value
- prior – algorithm parameter
- totalCount – total number of objects with the current categorical feature value

CTR example

fruit	target	ctr
apple	0	0.625
orange	0	0.25
apple	1	0.625
apple	1	0.625

prior $= 0.5$

Classes counter

For each pair (target_class, categorical_feature_value):

$$\text{count}_i = \frac{\text{curCount} + \text{prior}}{\text{totalCount} + 1}$$

- curCount – number of objects with the current categorical feature value
- prior – algorithm parameter
- totalCount – total number of objects

Counters example

fruit	target	ctr	counter
apple	0	0.625	0.7
orange	0	0.25	0.3
apple	1	0.625	0.7
apple	1	0.625	0.7

prior $= 0.5$

Overfitting

Gradients bias in gradient boosting

› Each subsequent tree is fit to the gradient between the current predictions on train and the true labels

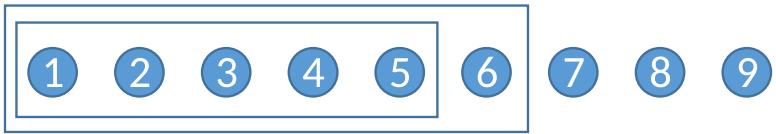
Gradients bias in gradient boosting

- Each subsequent tree is fit to the gradient between the current predictions on train and the true labels
- The gradient is estimated using the model fitted on the very dataset used for training

Gradients bias in gradient boosting

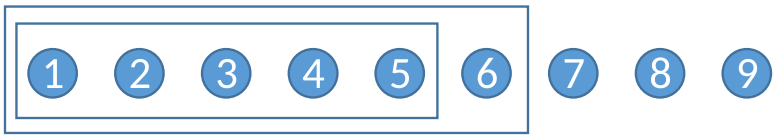
- Each subsequent tree is fit to the gradient between the current predictions on train and the true labels
- The gradient is estimated using the model fitted on the very dataset used for training
- The gradients are likely to be overfitted

Dynamic boosting



› Order data randomly

Dynamic boosting



› Order data randomly

› For each element maintain prediction based on the previous model elements

GB matters

> Gradient boosting is the workhorse of Yandex Internet business

Meet CatBoost

- Gradient boosting on decision trees
- Categorical features handling (even more advanced than discussed!)
- A novel boosting scheme (submitted to NIPS)
- Released into open source by Yandex on Tuesday
- Used in the LHCb PID

Contacts

Nikita Kazeev
Researcher

 kazeevn@yandex-team.ru

 nikita.kazeev.9