CatBoost

Mannem Charan AI21BTECH11019

Abstract

This report consists of my basic understanding of one of the popular Ml methods ”Catboost”.

1 Catboost

Catboost algorithm is an algorithm for gradient boosting on decision trees.It is a newly mechanised ml technique which introduces two advances in the classic GBM algorithm - the implementation of ordered boosting, a permutation-driven alternative to the classic algorithm, and an innovative algorithm for processing categorical features.

2 Handling Categorical features

The key features of CatBoost is one of the significant reasons why it was selected by many boosting algorithms such as LightGBM, XGBoost algorithm ..etc

With other machine learning algorithms. After preprocessing and cleaning your data, the data has to be converted into numerical features so that the machine can understand and make predictions.

This is same like, for any text related models we convert the text data into to numerical data it is know as word embedding techniques.

This process of encoding or conversion is time-consuming. CatBoost supports working with non-numeric factors, and this saves some time plus improves your training results.

3 Ordered Boosting

CatBoost has two modes for choosing the tree structure, Ordered and Plain. Plain mode corresponds to a combination of the standard GBDT algorithm with an ordered Target Statistic. In Ordered mode boosting we perform a random permutation of the training examples - $\sigma 2$, and maintain n different supporting models - M1, . . . , Mn such that the model Mi is trained using only the first i samples in the permutation. At each step, in order to obtain the residual for j-th sample, we use the model Mj-1. Unfortunately, this algorithm is not feasible in most practical tasks due to the need of maintaining n different models, which increase the complexity and memory requirements by n times. Catboost implements a modification of this algorithm, on the basis of the gradient boosting algorithm, using one tree structure shared by all the models to be built. In order to avoid prediction shift, Catboost uses permutations such that $\sigma 1 = \sigma 2$. This guarantees that the target-yi is not used for training Mi neither for the Target Statistic calculation nor for the gradient estimation.

4 Advantages of Catboost

1) No one-hot-encodings/sparse dataframe
2) Keeps original format of dataframe, making collaboration easier as well
3) Training is faster
4) Categorical features are more important
5) Model is more accurate
6) You can now work with features that you could not before like ID's, or categorical features with high unique counts