WEEK-3 REPORT P-4,P-5 ; WEEK-4 REPORT P-1,P-2

DASARI SRINITH (CS21BTECH11015)

GBM and XGBoost and LightGBM and CatBoost:

Theory:

- GBM , XGBoost , LightGBM and CatBoost are 4 Boosting Algorithms.
- Boosting is one of the techniques that uses the concept of ensemble learning.It is a technique that attempts to build a strong classifier from the number of weak classifiers.
- 1. **GBM** :
    - A Gradient Boosting Machine or GBM combines the predictions from multiple decision trees to generate the final predictions.
    - Different DT's capture different information from the data beacause the nodes in every decision tree take a different subset of features for selecting the best split.
    - And also, each new tree takes into account the errors or mistakes made by the previous trees.
    - The Pseudo-code for GBM can be seen as,
        (a) Initialize the outcome.
        (b) Then we have to iterate the follwing from 1 to total number of trees
            i. Updating the weights based from the pervious iteration.
            ii. Fitting the model on subset of data
            iii. Making predictions for full set of data
            iv. Updating the output with current results taking into account the learning rate.
        (c) Get the final output.
2. **XGBoost** :
    - XgBoost stands for Extreme Gradient Boosting. XGBoost is an implementation of Gradient Boosted decision trees.
    - The working procedure of XGBoost is the same as GBM. The trees in XGBoost are built sequentially, trying to correct the errors of the previous trees.
    - XGBoost has an option to penalize complex models through both L1 and L2 regularization. Regularization helps in preventing overfitting.
    - The following steps are involved in gradient boosting:

(a) $F_0(x)$ – with which we initialize the boosting algorithm – is to be defined:

$$F_0(x) = argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma) \tag{1}$$

(b) The gradient of the loss function is computed iteratively:

$$r_{im} = -\alpha \left[\frac{d(y_i, F(x_i)}{d(F(x_i))} \right]_{F(x)=F_{m-1}(x)} \tag{2}$$

where α is the learning rate.

(c) Each $h_m(x)$ is fit on the gradient obtained at each step.

(d) The multiplicative factor γ_m for each terminal node is derived and the boosted model $F_m(x)$ is defined:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{3}$$

3. **LightGBM** :
   - LightGBM is used to handle large amount of data , but it does not work well with small amount of data .
   - This is because ,the trees in LightGBM have a leaf-wise growth, rather than a level-wise growth.
   - After the first split, the next split is done only on the leaf node that has a higher delta loss. The leaf-wise split of the LightGBM algorithm enables it to work with large datasets.
   - Light GBM use histogram based algorithm i.e it buckets continuous feature values into discrete bins which fasten the training procedure.
4. **CatBoost** :
   - CatBoost is a boosting algorithm that can handle categorical variables in the data
   - Target Encoding is an efficient way to deal with categorical variables i,e, to substitute them with some numerical values (usually some target statistics).
   - We usually apply some smoothing in calculation (in target encoding) with a 'prior' term.

$$avg_{value} = \frac{count_{inclass} + prior}{total_{count} + 1} \tag{4}$$

where $total_{count}$ is the total number of objects with catogorical feature value

   - In ordered target encoding we find the values and number of times up to the current one ,i.e we dont use global statistics.This helps us prevent overfitting problem.
   - Numerically missing values are set to min value for that feature

- Both random forest and GBDT build a model consisting of multiple decision trees. The difference is in how the trees are built and combined.
- XGBoost is a scalable and highly accurate implementation of gradient boosting

Questions-GBM

1. What is difference one other difference between Random forest and GBDT ?

   Solution:
   Random forest "bagging" minimizes the variance and overfitting, while GBDT "boosting" minimizes the bias and underfitting.

2. How do I stop gradient boost overfitting??

   Solution:
   Regularization techniques are used to reduce overfitting effects, eliminating the degradation by ensuring the fitting procedure is constrained.

3. How do you optimize a gradient boost?

   Solution:
   Choose a relatively high learning rate then determine the optimum number of trees for this learning rate and tune tree-specific parameters for decided learning rate and number of trees finally lower the learning rate and increase the estimators proportionally to get more robust models

4. Do you need to normalize data for gradient boosting?

   Solution:
   No.

5. Is gradient boosting sensitive to outliers?

   Solution:
   Yes , as each tree is built on the residuals/errors of the previous tree , so more the outliners , greater the sensitivity.

Questions-XGBoost

1. Why is XGBoost faster than GBM?

   Solution:
   XGBoost uses advanced regularization (L1 & L2), which improves model generalization capabilities.

2. Does XGBoost handle missing values?

   Solution:
   Yes , XGBoost has an in-built capability to handle missing values.

3. How do I reduce variance in XGBoost?

   Solution:

   (a) increase depth of each tree (max_ depth),

   (b) decrease min_ child _weight parameter,

   (c) decrease gamma parameter,

   (d) decrease lambda and alpha regularization parameters.

4. Can XGBoost handle correlated variables?

   Solution:
   XGBoost automatically removes perfectly correlated variables before starting the calculation.

5. When to use XGBoost ?

   Solution:

   (a) When you have large number of observations in training data.

   (b) Number features ¡ number of observations in training data.

   (c) When the data has numerical and categorical features or only numerical values.

Questions-LightGBM

1. Is LightGBM faster than Random Forest ?

   Solution:
   A properly-tuned LightGBM will most likely win in terms of performance and speed compared with random forest.

2. One reason why LightGBM is faster

   Solution:
   Histogram based splitting

3. For what can tasks LGBM be used?

   Solution:
   LightGBM can be used for regression, classification, ranking

4. How do I increase the accuracy of LightGBM?

   Solution:

   (a) Use large max_ bin (may be slower)

   (b) Use small learning_ rate with large num_ iterations.

(c) Use large num_ leaves (may cause over-fitting)

5. How do I reduce overfitting in LightGBM?

   Solution:
   The min_data_in_leaf parameter is a way to reduce overfitting

Questions-CatBoost

1. When would you use a CatBoost classifier? ?

   Solution:
   We can use CatBoost without any explicit pre-processing to convert categories into numbers. CatBoost converts categorical values into numbers using various statistics on combinations of categorical features and combinations of categorical and numerical features.

2. Is scaling required for CatBoost?

   Solution:
   We need to apply feature scaling only for the Decision Tree Classification and not for XGBoost and CatBoost.

3. When to use XGBoost or CatBoost?

   Solution:
   If you want accurate model but it takes more time , then XGBoost should be used ; conversely , if you are ok with less accurate model but less time cousuming , then CatBoost should be used.

4. Is Target encoding data leakage?

   Solution:
   Target encoding does not cause target leakage because we "learn the target encodings from the training dataset only".

5. What is Target mean encoding?

   Solution:
   Target encoding is the process of replacing a categorical value with the mean of the target variable.