Weekly Report 2 - Random Forest

Ganji Varshitha
AI20BTECH11009

Introduction

Random Forest is an ensemble classifier which combines multiple classifiers to achieve better accuracy. It trains several models using bootstrapped dataset and selects the majority vote for classification problems and average for regression problems.

Algorithm

Algorithm 1 Random Forest Algorithm

```
Given a training set S
for i = 1 to k do
   Build subset S_i by sampling with replacement from S
   Learn tree T_i from S_i
   for each node do
      Choose best split from random subset of F features
      Each tree grows to the largest extent, and no pruning
   end for
end for
Make predictions according to majority vote of the set of k trees.
```

The value of F needs to be constant during the algorithm and it should be very less compared to total number of features M.
Possible values of F are $\frac{1}{2}\sqrt{M}, \sqrt{M}, 2\sqrt{M}$.

Why does bagging work?

Decision trees are prone to overfit which results in high variance of the model. Bagging reduces the variance of the model.
Let S be the training dataset.
Let S_k be a sequence of training sets containing a sub-set of S.
Let P be the underlying distribution of S.
Bagging replaces the prediction of the model with the majority of the predictions given by the classifiers S.

$$\phi(x, P) = \mathbb{E}_s(\phi(x, S_k))) \tag{1}$$

- The algorithm works if the dependent variables are linearly related with independent variables.
- It assumes the error terms to have constant variance and no correlation between one another.
- Since we need to compute the gradients, there is need to normalize the training data so that it does not take longer time.
- Learning rate decides how fast the algorithm converges.
- The model tends to overfit if there are many features in the dataset. Besides, there is a high chance correlation between some input variables.

Questions

1. Bias and variance in linear regression model are ________ related.
   A. directly   B. inversely
   Solution:
   B. inversely

2. Explain the geometric and probabilistic interpretation of the model.
   Solution:
   Geometric Interpretation:
   The least-squares regression function is obtained by finding the orthogonal projection of the output vector y onto the subspace spanned by $x_1, x_2, x_3, \cdots, x_d$.
   Probabilistic Interpretation:
   Let us assume the target variable be given by the deterministic function with added gaussian noise. This gives

$$p(y|X, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(y_n|\mathbf{w}^\top x_n, \beta^{-1}) \tag{2}$$

   where β is inverse variance of zero mean Gaussian random variable. We estimate the probability model using maximum likelihood which is same as minimising the least squares error.

$$\arg\max_{\mathbf{w}} L = \arg\min_{\mathbf{w}} E \tag{3}$$

3. Do the missing values in the data affect the model?
   A. Yes   B. No
   Solution:
   B. No

4. Is the model generative or discriminative? Explain why.
   Solution:

Linear regression is a discriminative model. As seen in the probabilistic interpretation of the model above, we learn the parameters that maximises the conditional probability $\mathrm{P}(Y|X)$. This is the basic definition of discriminative model where we assume functional form of $\mathrm{P}(Y|X)$ and estimate its parameters.

5. When to prefer gradient descent approach to ordinary least squares method?
   Solution:
   Time complexity of least squares method is $\mathcal{O}(n^3)$ whereas time complexity of gradient descent is $\mathcal{O}(n)$. Gradient descent is preferred to ordinary least squares method when n i.e number of input features is greater than 10,000.

6. Which error function is sensitive to outliers?
   - A. MSE(Mean squared error)
   - B. MAE(Mean absolute error)

   Solution:
   A. MSE