Toronto Business College

MaCHINe Learning 2

module 11 Case study

|  |  |
| --- | --- |
| **Course** | 2021F-T2 AISC2000 – Machine learning |
| **Instructor** | Mahnaz Malekzadeh |
| **Students** |  |
|  | Charitha Priya Dongari – 500191110 |
|  | Katterapalli Venkata Satya Narayana Reddy – 500190693 |
|  | Rumana Banu Iliyas Ahmed – 500186725 |
|  | Rohan Bhatt – 500187633 |
|  | Shivam Jolly – 500196452 |

1. **Consider the K-Nearest Neighbors (KNN) classifier. Think about *k* parameter and explain if it can result in overfitting.**

KNN classifier is used as a supervised learning algorithm that can be applied for regression and classification tasks. The approach is simple, for regression tasks, the value of data points is calculated as the average of the k closest data points target variable value, and for classification, it is the mode of the classes of the k closest data points. Here k is the number of data points we must consider before deciding. K is the most important part of the algorithm, and we must choose it carefully. Depending on the value of K our model performance can increase or decrease. If the value of K is very less, the model will not be able to generalize much. The model created will be noise sensitive, will give high accuracy on the training set, whereas the model will perform poorly on the test set as the values will vary even on slight changes in data points, and hence, it will lead the model to overfit.

1. **How is Bagging technique different from Random Subsampling used in Random Forest?**

Bootstrap aggregation, which is also commonly known as bagging is an ensemble learning method. In bagging, samples of random data are generated with replacement, from the training set meaning the same data points can be selected more than once, these samples are trained independently on weak models, and then a strong model is created which helps in regression or classification tasks by choosing the mean of values or by majority vote. We use random subsampling in Random Forest which is basically selecting features randomly and using a bagging technique with it. In this, at each split point, the algorithm only looks at some specified number of features and selects the most optimal points. This helps in making the resulting predictions less correlated and hence leads to better performance.

1. **Research about postpruning technique. Explain**
2. **Why it is used,**
3. **How it works,**
4. **And if it outperforms prepruning, why?**

Decision trees are tree-like structures that are generated for classification or regression tasks from the dataset. Most of the time, we need to find the optimum size to have better accuracy for the problem statement. When a tree has too many branches and leaves, it is prone to overfitting. Hence, we use a technique like a post pruning to reduce the redundant subtree. Post pruning means pruning when the tree is built. We prune the tree in a bottom-up fashion using the measures like Gini impurity or information gain. For example, when we try to prune a subtree based on the least information gain, we replace it with the leaf node. And then we observe, what leaf node the algorithm would have created if this subtree was not there. Post pruning tends to be more effective than the pre – pruning as the latter one is greedy in nature i.e., it makes the algorithm avoids the partition and it is possible that the next partition created could be extremely valuable.

**References:**

* <https://towardsdatascience.com/how-important-is-the-k-in-knn-algorithm-3b6fce726110>
* <https://www.ibm.com/cloud/learn/bagging>
* <https://www.r-bloggers.com/2021/08/feature-subsampling-for-random-forest-regression/>
* <https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/>
* https://www.educative.io/edpresso/what-is-decision-tree-pruning-and-how-is-it-done
* <https://towardsdatascience.com/pre-pruning-or-post-pruning-1dbc8be5cb14>