Machine Learning Assignment

Written Assignment

Name: Amruth Karun M V
RollNo: 2020MCS120004
Date: 05-Nov-2021

Q2. [5 points] **Based on the reading, explain ensemble learning in your own words.**

Ensemble learning is the process by which we can improve the performance of our machine learning models by combining the outputs of several individual models to take the final decision. It just like when we consult others before taking our own decision due to any variability in the past or because of their expertise in that area. We do these kind of consulting when we need to make some investment decisions and our decisions are not accurate all the time and we might end up in a loss. This is because not all decisions are made from the same set of information we have. So, It would be useful to know what we can infer from different sets of data. Similarly, for classifiers we use this concept to form ensemble learning where we create an ensemble of classifiers. We can construct different models by varying the initial seed, taking entirely different data, using different set of parameters, using different techniques. The parameters learned by all these also vary. So, if we combine the results from each of these models, we can improve our machine learning models.

We can use different approaches to arrive at the final decision, use majority voting to pick the class with highest vote and we can add weights to those votes in favour of classifiers that are more likely to be correct. As we know, the models are subjected to bias and variance. If bias is large, our model under-fits the data and if the variance is large, it results in over-fitting. Ensemble learnings is a method used for maintaining a balance between these errors. In our problem 'Titanic Survival' classification we used cross validation and hyper-parameter tuning approaches to train multiple models with different folds of training data and different parameters to find the best model. We doing cross validation we also computed the mean of the cross validated score which is the result of the ensemble of models we trained. We can train these in parallel and sequential fashion to arrive at the final result.

Over-fitting can be reduced by reducing the variance. This can be done by implementing multiple classifiers and training with different set of training data by randomly sampling from the original data set and taking the average of the individual predictions to make the final prediction (eg. Random Forest classifier). This approach is known as **Bagging.** Another technique known as **Boosting** is used to reduce the bias, where we iteratively adjust the weights of an observation based on the previous classification results (eg AdaBoost). Boosting transforms multiple weak learners into strong classifier by correcting them in each iteration. This approach can build strong predictive models but can also result in over-fitting of the training data. The next technique is known as **Stacking** where we combine the output of different learners to produce the final result. This will have an effect on bias or variance depending upon the learners we use. It is key to chose which ensemble technique we use for a problem and this depends upon our knowledge about the training data and problem domain we are dealing with.

Q3. [10 points] **Briefly describe any one use-case (related to your organization/institute) that can harness the benefits of machine learning.**

Organization - Allianz
Use case - Automatic Vehicle Insurance Claim Prediction

I work for a company which is one of the largest insurance and asset management providers in the world. There is a huge amount of insurance and personal data which must be analysed in a confidential manner to build models to automate all manual process which reduces time, cost and improve the overall business of the organization. Here, we consider a vehicle insurance claim scenario and we can leverage machine learning techniques to predict the claim amount. The challenge here is the amount of data we get daily. There are about 1200 accidents in a single day in India and when it comes to the total world count it accounts to millions. So, manual processing of all these documents are very tiring and time consuming. We can use the existing machine learning techniques to automate the whole process.

The task is a regression problem where we can predict the amount based on other features like vehicle model, registration year, city and many other contributing factors. But due the large quantity of these documents it is very difficult to enter the values to create a dataset for the predictor. So, the whole task can be split into multiple stages where we want to use other approaches in computer vision, natural language processing, to first extract this data from the documents and give the recognized values as the training data to the final predictor. So, the photos of the registration and other proof documents might be taken with the mobile camera and might be of lower quality which makes the task even more challenging. This we first need to remove the noise from the original document using some Autoencoders or Random Forest methods to prepare it for the OCR to recognize the fields. And, to make the task simpler we can assign some priority to different fields like vehicle model, registration date, chasis number, whether insurance is taken in previous, licence plate number having the highest priority. So, the necessary fields can be trained recognize using Recurrent Neural Networks or similar convolution based approaches.

Once different fields are recognized, we must make sure that there are no null values and if there are large amount of null value we should drop that attribute. We can fill other missing values using interpolation techniques. After cleaning our data, what remains is the training data for our target predictor. The problem might not be as simple as a linear regression problem because the re are varying factors which might increase or decrease the final output. Then we fit our training data to our training algorithm to generate the final predictor.

In the inference phase, we just need to provide the document to the system, which extracts the data from the document, cleans it and recognizes the text corresponding to each field and this is given as input to the target predicting algorithm and calculates the final claim amount. In this way, we can process multiple documents in a parallel and distributed fashion to improve the speed and incorporate ensemble learning to improve the predictive power of our algorithms.