Reading 9 : AI ML Challenges

Name: Pushpak Vijay Katkhede **Date: 03/13/2023**

---

1. *Give (a) one reason you can think of that continual learning (re-training a classifier after the initial deployment) would be beneficial (good) and (b) one reason it could be detrimental (bad).*

Ans: Continual Learning →

1. Beneficial → Apart from cost of retraining over a bigger load of data, the second most important benefit I think of continual learning is that it allows us to understand our data better. By incrementally adding new data to our model we can observe how new set of examples improves accuracy of our model and hence patterns that were previously not visible due to stateless training can be evidently spotted in our data. This pattern can be helpful to us to simulate our data whenever we plan on maybe changing the model behind completely or exploring different options.
2. Detrimental → I assume that from the reading every increment leads to a better accuracy for our model pipeline which will eventually tend to maximum accuracy . This may cause the problem of overfitting the model. As discussed in previous point about the data patterns the downside of these patterns could lead highly bound model to the training data and leads very low variation in data. As a result, there is a risk that the model may perform bad on the unseen data.

2. (a) Imagine you want to have the best spam classifier possible for your personal email inbox. How often would you want this classifier to re-train?

Ans: As per my understanding here it will not be beneficial to retrain the model based on some specific time criterion. Moreover this could be more relevant if we retrain based on number of new emails or count of newly identified spam emails which were actually not spam. Meaning the models accuracy has been degraded and model is prone to decay. So, we can specify certain threshold say on encountering 100 or 500 such False positive observation we should retrain the model.

(b) What new labeled data would your spam classifier use each time its re-trains ("data iteration")?

Ans: There can be new emails that can be spam for example from a new sender which actually have a different email structure than usual. Or, for an example previous spam emails didn't have images but now the new spam emails can be with images or suspicious links. These emails should be labeled by manual efforts as our model might not be able to identify these emails as spam as it is blind towards them. Overall, our goal is to continually update the model with new data to ensure that it remains accurate and up-to-date with the latest trends and techniques used by spammers.

3. Random forests consist of multiple decision trees each trained on a different subsample of the training data. In an online setting in which new data arrives over time, how could you update an existing (trained) random forest to include the new data without re-training the random forest from scratch?

Ans: One of the approaches might be to add a new tree to the forest that consists of the new data. Later, the weights / hyperparameters for the trees in the forests should be updated accordingly. This can be achieved through a process called bootstrapping in which the decision tree is trained on a subsample of newer data. This approach can be feasible if the amount of new datapoints is significant to justify creating a new tree. Else we it is better option to just feed that to the existing trees in the forest.

4. *Reflect on your work for the course over the past week. What did you do that was effective and increased your knowledge? What could you do or change to increase what you gain from this course? Is there anything about this course you are anxious / worried about? (There are no wrong answers here; this is your chance to maximize what you get out of the course and to let me know about any concerns.)*

Ans: Last week was much dedicated to learning more about skewness of the data as it was the last challenge remained in my project. I tried exploring different trends in my data by applying techniques like normalization and binning with different parameters. Now, I believe that after this practice I know my data way better than before. Apart from this, I am really excited to give the presentation and convey my work to our whole class. There is still not much of anxiety or stress about anything and really cheerful for the last week of this wonderful learning journey!!