Programming assignment #6

Course: CHE1147H - Data Mining in Engineering

1 Model performance

A very common question in every machine learning problem is: how many data samples do we need to model the system behaviour adequately. Unfortunately, just like many other topics in machine learning, there is no straight answer. In many toy problems presented in textbooks, a classification problem is solved with only 50-100 data points. In real world problems, a classification problem may be very difficult even with millions of data points.

Generally, the model performance depends on the following factors:

1. Are the classes easily separated or they are pretty mixed? Are they separated linearly or non-linearly? Is a linear or non-linear model used?
2. The features quality. Do they carry information with respect to the output/class? More features does not necessarily mean better performance. The famous quote "Garbage in, garbage out" is used to describe uninformative features.
3. The number of data points. Intuitively, more data points lead to better performance. But after some point, it is expected that the increase in model performance diminishes.

The last point is the subject of this section. From a business perspective, you want to know how many samples you need to model the clients behaviour adequately. This information is crucial when the conditions change and you may want to re-fit your model.

For example, with Covid-19 the clients behaviour changed dramatically. Let's assume that you are at the beginning of Covid-19 in March 2020 and your manager is asking you to re-fit the retail response problem you solved in Assignment #5 (apologies for putting you mentally back at the beginning of Covid-19, we are almost out of it). The question that comes with this request is: **how many data points** do you need to re-fit the model with adequate performance?

You know that generally more data points means better performance, but you cannot wait for too long to collect new data post-March 2020 because your business will not have a reliable model for as long as you collect data. A similar situation may appear in an industrial setting, let's say after the annual maintenance of a machine or a reactor. How many data points do you need to model the machine or reactor behaviour after the maintenance?

1.1 Dataset size vs model performance

Here, you will quantify the relationship between the dataset size and the model performance. Essentially, you will answer the question: how much data is enough to model client behaviour? In order to do this, you will pick the best single tree model you created in Assignment #5 and evaluate it with datasets of different sizes using the monthly features you created in Assignment #3.

Perform the evaluation with the following steps:

1. Split the train/test sets with 9:1 ratio This split should give you approximately 291k/32k samples in train/test set, respectively.

2. Initialize and create a for loop in which you take N samples (e.g. 50), build a tree model with the N samples and evaluate the test set AUC. Repeat the sampling process 10 times and append the test set AUC. The following table shows the desired output:

N = 50 samples	
sample #	Test AUC
1	0.545
2	0.561
⋮	⋮
10	0.551

   From this table, you can calculate the mean and standard deviation of the test AUC for N samples.

3. Repeat the procedure you performed in the previous step for different sample size N (e.g. 100, 500, 1000, 2000, 5000, 10000) [1].

4. Build a table that contains the values of:

   - Sample size N
   - Test AUC mean
   - Test AUC standard deviation

5. Using the matplotlib function errorbar, plot the model performance captured in the test AUC mean and standard deviation as a function of the sample size. From this plot, can you estimate what is the minimum number of samples needed to model the behaviour adequately?

[1] The N values here are just my educated guesses. You should try values that will give you a meaningful result as described in the next steps.