DataLab Preparation (Week 5, DataLab II, Wednesday)

2. Error Analysis and Performance Auditing

**2a Explain why error analysis is considered the heart of the machine learning development process.**

Error analysis is considered the heart of the machine learning development process because it helps us identify and select the problems which impediment the machine learning algorithm functioning at its highest potential. Without this step in the ML development process, we would not be able to improve the performance of the model in such an efficient way, wasting time and resources.

**2b Explain the importance of creating tags (like "car noise", "people noise", etc.) during error analysis.**

Tags are used during the error analysis process to help identify in which class most misclassifications happen, how those misclassifications look, what percentage of all the data has the specific tag, etc. Therefore, tags help understand more about certain errors, and they can be used to decide which problem to prioritise and guide us in coming up with an appropriate solution.

**2c Provide examples of different types of tags that could be used for visual inspections in the dataset of your creative brief project. How do these tags contribute to the error analysis process?**

'Low-lighting', 'Multiple bottles' , 'Blurry', 'Reflection', 'Labeled bottle(s)', 'Non-labeled bottle(s)', 'Symmetrical', 'Smooth', 'Rough', 'Colour inconsistency', etc.

Tags aid in the error analysis process by helping me identify error patterns and the root cause, leading to an improved decision-making process and faster improvement of the model performance.

**2d How analyzing the percentage of errors associated with a particular tag (like 'car noise') allows you to make informed decisions on what aspects of a machine learning model to improve?**

It helps determining the overall accuracy of the model.

[x]It indicates the relevance of a particular issue.

It is useful to reduce the dataset size.

It helps identifying redundant tags.

**2e Considering the example presented in video 15 (Prioritizing What to Work On), why might it be more beneficial to focus on "people noise" instead of "car noise"?**

[x]People noise has a larger data percentage.

Car noise is less important for overall performance.

People noise is easier to improve.

None of the options above.

**2f Explain what is a skewed dataset.**

A skewed dataset occurs when the distribution of data points across different categories is not evenly balanced. Instead, there is a significant imbalance, with one or more categories having a disproportionately large or small number of observations compared to others.

**2g Consider that you are working on a binary classification ML algorithm that detects whether a patient has a specific disease. In your dataset, 98% of the training examples (patients) don't have the disease, so the dataset is very skewed. Accuracy on both positive and negative classes is important. You read a research paper claiming to have developed a system that achieves 95% on \_\_ metric. What metric would give you the most confidence they have built a useful and non-trivial system?**

Accuracy

Precision

[x]F1 Score

Recall

**2h Considering the problem described in the previous exercise, if your algorithm simply print('1') (i.e., it says everyone has the disease). Which of these statements is true?**

The algorithm achieves 0% recall

The algorithm achieves 0% precision

[x]The algorithm achieves 100% recall

The algorithm achieves 100% precision

**2i Describe the process of conducting a performance audit on a machine learning model before deploying it to production. What are the key steps and why are they important?**

The process of conducting a performance audit goes as follows:

1. Check for accuracy, fairness/ bias, and other problems by:

- Brainstorming the ways the system might go wrong: performance on subsets of data (e.g., ethnicity, gender), how common certain error are (e.g., FP, FN), performance on rare classes.; TensorFlow has a package for model analysis (TFMA) that computes detailed metrics on new ML models.

2. Establish metrics to asses performance against these issues on appropriate slices of data, rather than the whole dataset.

3. Get business/ product buy-in.

3. Data Iteration

**3a Explain what is the data-centric view in the development of machine learning models.**

A data-centric approach means holding the quality of the data as paramount instead of focusing on the model itself. This view includes using tools to improve the data quality such as data augmentation and error analysis allowing a more general good performance across multiple models.

I.e., holding the code fixed and improving the data.

**3b Let's say that you are considering applying data augmentation to a phone visual inspection problem. Which of the following statements are true about data augmentation?**

Data augmentation should distort the input sufficiently to make sure they are hard to classify by humans as well.

Data augmentation should try to generate more examples in the parts of the input space where the algorithm is already doing well and there's no need for improvement.

[x]Data augmentation should try to generate more examples in the parts of the input space where you would like to see improvement in the algorithm's performance.

Generative Adversarial Network (GANs) can be used for data augmentation.

**3c Explain why adding more data to a dataset rarely hurts the accuracy of a model.**

For unstructured data problems with a large enough model that has a low bias, and where the labelling is properly made (i.e., humans can make accurate predictions on only the input), increasing the dataset with accurately labeled data rarely hurts the model accuracy because the model can learn from a diverse set of cases and it will do a good job even on the portion of data that is new, therefore generalising better on test sets and reducing variance in the model's predictions.

**3d Give an example where adding more data can negatively affect the accuracy of a model.**

Adding more data can negatively affect the accuracy of a model in instances where the added data is too ambiguous, even for baseline metrics such as HLP, leading to decreased accuracy. As an example, adding more data from a different neighbourhood to predict whether students will pass an exam may introduce biases. If the new data represents a wealthier area with better resources, the model may become skewed, leading to less accurate predictions.

**3e Why experiment tracking is important in a machine learning project?**

Having a system that tracks the ML experiments can help improve efficiency in making decisions about the model, hyperparameters, data, etc. to systematically improve the model's performance.

What to track?

- Algorithm/ code version

- Dataset used

- Hyperparameters

- Results (at least the high-level metrics)

Tracking tools

- Text files - good for smaller experiments; does not scale well

- Spreadsheets - good for teams, more organised

- Experiment tracking systems

Desirable features:

- Does it give me all the information needed to replicate the results? (Replicability)

- Tools that help you quickly experiment results, ideally with the summary metrics/ statistics

- Resource modeling, visualisation tools, model error analysis tools

**3f What can be considered "good data" in a machine learning project?**

Good data:

- Covers important cases (good coverage of inputs x); data augmentation is not enough data/ data is not diverse

- Is defined consistently (definition of labels y is unambiguous)

- Has timely feedback from production data (distribution covers data and concept drift)

- Is sized appropriately