DataLab Preparation (Week 5, DataLab II, Thursday)

2. Define Data and Establish Baseline

**2a You are building a system to detect cats. You ask labelers to "use bounding boxes to indicate the position of cats". Different labelers label as seen below. What is the most likely cause of this?**

Lazy labelers

Labelers have not had enough coffee

[x]Ambiguous labeling instructions

That this should have been posed as a segmentation rather than a detection task

**2b Which of these statements do you agree with regarding structured vs. unstructured data problems?**

It is generally easier for humans to label data on unstructured data, and easier to apply data augmentation on structured data.

[x]It is generally easier for humans to label data and to apply data augmentation on unstructured data than structured data.

It is generally easier for humans to label data on structured data, and easier to apply data augmentation on unstructured data.

It is generally easier for humans to label data and to apply data augmentation on structured data than unstructured data.

**2c Describe the challenges of labeling large datasets and how they differ from small datasets.**

For large datasets, the labeling emphasis has to be on the data process. It is more difficult to have a consistent communication channel between labellers, as you can do with smaller datasets. Therefore, usually a smaller team establishes consistent labelling conventions which are then passed to the rest of the labelers. Additionally, with small datasets, labelers can be more thorough by manually looking through the dataset and fixing the inconsistent labels, unlike with big datasets, where more resources would have to be used in order to achieve this level of thoroughness.

**2d Discuss the implications of using a large, poorly labeled dataset versus a small, meticulously labeled dataset in developing a machine learning model for financial fraud detection.**

When using a large dataset that has noisy labels, the algorithm can still fit a function over it pretty confidently. Therefore, it would still be able to detect fraudulent financial cases. However, if the dataset is poorly labeled, the model's performance may be compromised. In contrast, a small dataset with meticulous labeling may yield a higher-quality model with better accuracy, despite its limited size.

Labeling a large dataset requires substantial human and financial resources. If the dataset is poorly labeled, these resources may be wasted on annotating inaccurate or inconsistent data.

Additionally, even with big datasets, there can still be small dataset problems. For example, in the case of financial fraud detection, even though a big poorly labeled dataset may still be able to yield decent results, in rare events such as when there is a sudden shift in fraud patterns (that is not new, but is rare), a large, poorly labeled dataset may fail to capture these nuances effectively. This can lead to blind spots in the model's understanding. This is where label consistency and reliance makes a big difference.

**2e What is essential when dealing with small datasets in machine learning?**

Large number of features

[x]Clean and consistent labels

Complex algorithms

High computational power

**2f What can help improve the performance of a machine learning model when dealing with rare events in large datasets?**

Increasing the dataset size

Ignoring the rare events

[x]Ensuring label consistency

Using a simpler model

**2g Take a phone visual inspection problem. Suppose that even a human inspector looking at an image cannot tell if there is a scratch. If, however, the same inspector were to look at the phone directly (rather than an image of the phone), then they can clearly tell if there is a scratch. Your goal is to build a system that gives accurate inspection decisions for the factory (not publish a paper). What would you do?**

Carefully measure HLP on this problem (which will be low) to make sure the algorithm can match HLP.

Get a big dataset of many training examples, since this is a challenging problem that will require a big dataset to do well on.

[x]Try to improve their imaging (camera/lighting) system to improve the quality or clarity of the input images, x.

Try to improve the consistency of the labels, y.

**2h In what scenarios might creating a new class for ambiguous data not be a viable solution? Discuss the potential drawbacks.**

In some situations, a new class meant to simplify the labeling process might end up creating a new set of problems itself.

In the situation where labelers might decide to merge two classes to eliminate all potential subjective disagreements, they might end up also losing important data for the task at hand. For example, for a phone manufacturing factory, two classes labeled 'small scratch' and 'big scratch' are useful when determining the root cause of the defect. Merging the two classes might alleviate labeling problems, but it also complicates the process of fixing the bigger problem at hand.

Models with multiple classes, including a class for ambiguous data, may be less interpretable than simpler models with fewer classes. This could also lead to loss of important data as sometimes even if the labelers are not 'correct' in their labeling decisions, defining correctness in a machine learning context is entirely left to the engineering team.

**2i Considering the speech recognition problem. Some labelers transcribe with "…" (as in, "Um… today's weather") whereas others do so with commas "," ("Um, today's weather"). Human-level performance (HLP) is measured according to how well one transcriber agrees with another. If you make your team of labelers to consistently use commas ",". What effect will this have on HLP?**

HLP will stay the same.

[x]HLP will increase.

HLP will decrease.

3. Label and Organize Data

**3a Discuss why it might not be beneficial to spend a long time collecting data before training the initial model.**

Spending a long time collecting data before training the initial model would lock the project out of the iterative loop, delaying the machine leaning development process. Instead, it is best to spend a fixed amount of time collecting data.

**3b Do you agree with the following statement: "to implement the data iteration loop effectively, the key is to take all the time that's needed to construct the right dataset first, so that all development can be done on that dataset without needing to spend time to update the data"? Why?**

I do not agree with the statement because of multiple reasons:

- spending a long time constructing the 'perfect' dataset would delay the project too much

- even if the dataset is right, the concept of 'data drift' or 'concept drift' might occur, which would make the dataset redundant and the time spent on it wasted.

- in the modeling phase of the Machine Learning Project Lifecycle, sometimes discoveries relating the dataset needed are made, therefore leading to backtracking on the iterative process anyways, and data being changed.

- no matter the task at hand, data will need to be changed anyways after a set amount of time.

**3c What does the term "data pipeline" refer to in a machine learning project?**

[x]A sequence of data processing steps from raw data to final output.

A tool used for data visualization.

The process of collecting raw data.

A type of algorithm used for data analysis.

**3d What is the primary goal of building a PoC (proof of concept) system?**

[x]To check feasibility and help decide if an application is workable and worth deploying.

To build a robust deployment system.

To collect sufficient data to build a robust system for deployment.

To select the most appropriate machine learning architecture for a task.

**3e MLOps tools can store meta-data to keep track of data provenance and lineage. What do the terms data provenance and lineage mean?**

Data provenance refers the input x, and data lineage refers to the output y.

Data provenance refers data pipeline, and data lineage refers to the age of the data (i.e., how recently was it collected).

[x]Data provenance refers to where the data comes from, and data lineage the sequence of processing steps applied to it.

Data provenance refers to the sequence of processing steps applied to a dataset, and data lineage refers to where the data comes from.

**3f You are working on phone visual inspection, where the task is to use an input image, x, to classify defects, y. You have stored meta-data for your entire ML system, such as which factory each image came from. Which of the following are reasonable uses of meta-data?**

As another input provided to human labelers (in addition to the image x) to boost HLP.

As an alternative to having to comment your code.

[x]To suggest tags or to generate insights during error analysis.

[x]Keeping track of data provenance and lineage.

4. Scoping

**4a What is scoping in the context of machine learning?**

In the context of machine learning, "scoping" refers to the process of defining the boundaries, objectives, and constraints of a machine learning project or task, i.e., clarifying the problem the team will be working on.

**4b Considering that you are working for an e-commerce company and you need to consider several machine learning projects to develop. Which factor is most important when deciding which project to undertake?**

The complexity of the machine learning algorithms.

[x]The potential impact on sales and customer satisfaction.

The personal preferences of the data science team.

The current trends in machine learning research.

**4c What is the emphasize of the initial step of the scoping process?**

Identifying existing AI solutions.

Developing machine learning algorithms.

[x]Assessing the technical feasibility of ideas.

Understanding business problems.

**4d Explain why it is important to separate the identification of business problems from the generation of AI solutions in the scoping process.**

The separation of the identification of business problems helps form a clear understanding of the business problem helps engineers come up with better solutions. As not all problems are solved by AI, knowing what the client wants and needs helps speed up the process of implementing a solution and bridges the gap between what AI engineers can do and what clients expect.

**4e Which of the following criteria are typically used for assessing the technical feasibility of a new project with unstructured data?**

[x]Human-level performance (HLP).

External benchmarks.

Availability of predictive features.

History of the project.

**4f How does using human-level performance (HLP) as a benchmark help in assessing the feasibility of machine learning projects?**

Human-level performance is very useful as a benchmark in assessing the feasibility of a machine learning projects, mostly when talking about unstructured data, because the human eye and mind is meticulously trained since a young age to detect patterns and associate objects, i.e., to make sense of the world around us at all times. This is something that ML models are trained to replace or even outperform. Therefore, it is logical to use HLP as a benchmark help, especially when it comes to new tasks that a model is supposed to solve.

**4g Explain why the history of a project is considered a good predictor for future progress in machine learning applications**

The history of a project can be a surprisingly good predictor for the future progress in ML applications because analysing past stages of the model, we can get a good idea if it is possible to improve it, by how much, over what period of time, etc. It constitutes then a starting point for the task at hand.

**4h Imagine you are evaluating a project to implement a machine learning model for predicting traffic patterns in a city. Describe how you would use the criteria of HLP, predictive features, and project history to assess the feasibility of this project.**

HLP- I would use the criteria of Human-Level Performance as a starting point to understand what type of traffic patterns are most common, how easy they are recognised, and what the corresponding following step is.

Predictive Features- I would identify relevant features and data sources that could be used to predict traffic patterns effectively (like traffic volume, road infrastructure, weather conditions, time of day, day of week, special events, etc.).

Project history - I would use project history to assess the feasibility of the project itself.

**4i What is a common challenge when estimating the value of a machine learning project?**

Finding qualified data scientists.

[x]Aligning machine learning metrics with business metrics.

Deciding on the programming language to use.

Choosing the right machine learning algorithm.

**4j In the context of voice search, why is query level accuracy considered more important than word level accuracy?**

It is easier to measure.

It directly impacts the efficiency of the algorithm.

It is the only metric that matters to business owners.

[x]It aligns more closely with user experience.

**4k Discuss the importance of ethical considerations in deciding whether to pursue a machine learning project.**

Ethical considerations are a detrimental factor in the machine learning project lifecycle because they represent the 'unseen' consequences of deploying a model into the real world. When starting out as a business problem, a ML project has to be considered ethically sound due to their impact on society, fairness, transparency, privacy, potential harm, and long-term sustainability. Machine learning projects can significantly affect individuals, communities, and society at large, thus ethical considerations ensure these impacts align with societal values.