DataLab Preparation (Week 5, DataLab I, Monday)

3. Introduction to Machine Learning in Production

**3a Describe what is the "Proof of Concept to Production Gap".**

Proof of Concept to Production Gap refers to the challenges and discrepancies that often arise when transitioning a project to the full-scale production deployment stage.

**3b In the speech recognition example presented in video 3 (Case Study: Speech Recognition), what is the problem with some labelers transcribing audio as "Um, today's weather" and others transcribing "Umm…, today's weather"?**

We should not be transcribing "Umm." The correct transcription, which serves the user's needs better, is just "Today's weather".

The first is grammatically incorrect and we should use the second transcription.

The second is grammatically incorrect and we should use the first transcription.

[x]Either transcription is okay, but the inconsistency is problematic.

**3c What are "data definition questions"? Give two examples.**

Data definition questions help to ensure that the data to be used to train the ML algorithm is properly defined and that the baseline is established, as well as to label and organise the data in an efficient and useful way.

Examples:

- Is the data labeled consistently?

- (For speech recognition algorithms) How to perform volume normalization?

**3d How error analysis can be used to improve models?**

Error analysis, which is part of the modeling stage of the machine learning project lifecycle, can tell you how to systematically improve the data, which helps you be more targeted towards the data needed to build an accurate model.

**3e Considering the deployment of the speech recognition system presented in video 3 (Case Study: Speech Recognition) as as example, describe and illustrate how the application that you are building for your creative brief project would work in a production environment.**

The software would be installed on a tablet, such as the ones used to take orders by waiters (edge device). The employees and the distributors would input the quantity amount of the order, which the system would then send the information to the internal storage cloud which would keep evidence of how much stock is left of a certain category of product.

The software would also make sure to send an alert when stock is low on a certain product category.

**3f Which of these are stages of the machine learning project lifecycle?**

[x]Scoping

[x]Deployment

[x]Data

[x]Modeling

Configuration

**3g Which statement is a more accurate description of the full cycle of a machine learning project?**

It is a linear process, in which we move step-by-step from scoping to deployment (that's why we call it a cycle - bicycles are only good at going forward, not backward).

[x]It is an iterative process, where during a later stage we might go back to an earlier stage (that's why we call it a cycle - it's a circular process)

4. Deployment

**4a Describe the key deployment challenges related to your create brief project.**

Software engineering issues (cloud vs edge browser, compute resources).

Additionally, another problem to be faced in my creative brief project is data drift, as new beverages enter the marekt and gain popularity, therefore the system would not keep up with trends over time and would not be able to recognize brands it was not trained on.

5. Selecting and Training a Model

**5a Which of the following sentences is a more accurate description of a data-centric approach to machine learning development?**

Holding the training data fixed, work to improve your neural network's architecture to do well on the problem.

[x]Holding the neural network architecture fixed, work to improve the data to do well on the problem.

**5b What are the main challenges in model development?**

When building a model, there are three main main challenges in the modeling stage of the machine learning project lifecycle. These challenges can also act as milestones if reversed.

1. Model doing poorly on the training set (usually measured by average training error)

2. Model performing poorly on test set/ holdout cross-validation set.

3. Results not adhering to the business metrics/ project goals.

**5c Let's say you have an algorithm that diagnoses illnesses from medical X-rays, and achieves high average test set accuracy. What can you now say with high confidence about this algorithm? Check all options that apply.**

It does well even on rare classes of diseases.

Its diagnoses are roughly equally accurate on all genders and ethnicities, so we are confident it is not biased against any gender or ethnicity.

The system can be safely deployed in a healthcare setting.

[x]None of the options above.

**5d What is Human-Level Performance?**

Human Level Performance (HLP) is the average level of performance achieved by humans on a task given to a ML algorithm. It is used as a baseline to compare the performance of models, being considered a significant milestone in the development of AI systems, as it demonstrates the ability of machines to portray intelligence and capabilities comparable to those of humans.

**5e What is "unstructured data"? What is "structured data"? Give examples of both types of data.**

Unstructured data refers to data that is not organised in a certain matter, such as images, audio or text. For this type of data, HLP is a very useful benchmark for measuring the performance of a model.

Structured data refers to the predefined databases (excel spreadsheets), such as inventory data for a shop.

**5f Which of these statements about establishing a baseline are accurate? Check all that apply.**

Open-source software should not be used to establish a baseline, since the performance of a good open source implementation might be too good and thus too hard to beat.

[x]For unstructured data problems, using human-level performance as the baseline can give an estimate of the irreducible error/Bayes error and what performance is reasonable to achieve.

[x]Human-level performance (HLP) is generally more effective for establishing a baseline on unstructured data problems (such as images and audio) than structured data problems.

[x]It can be established based on an older ML system.

**5g Let's say you run a sanity-check test of trying to overfit a single training example in a speech recognition problem. You pick a clearly articulated clip of someone saying "Today's weather", and the algorithm fails to fit even this single audio clip, and outputs "\_ \_ \_ \_ \_ \_ \_". What should you do?**

[x]Debug the code/algorithm/hyperparameters to make it pass this sanity-check test first, before moving to larger datasets.

Train the algorithm on a larger dataset to help it to fit the data better.

Use data augmentation on this one audio clip to make sure the algorithm hears a variety of examples of "Today's weather" to fit this phrase better.

Create a training set of this example repeated 100 times to force the algorithm to learn to fit this example well.

6. Human-level Performance

In tomorrow's DataLab, you will calculate the Human-Level Performance (HLP) for your project. Therefore, you need to prepare a form where each question is an image, and answers are class labels.

After creating the form, click in the "Collect responses" button, and then copy the generated link and paste it here.

https://forms.office.com/e/Q4NNpKebS2