DataLab Preparation (Week 3, DataLab II, Wednesday)

2. Overview of the Machine Learning Project Lifecycle

**2a What are the key questions that you should keep on the top of your mind when framing a machine learning problem? (including the questions that you should discuss with stakeholders)**

The key questions are:

1. Why is your customer trying to solve this particular problem?
2. What value will they derive from the solution—how will your model be used, and how will it fit into your customer’s business processes?
3. What kind of data is available, or could be collected?
4. What kind of machine learning task can be mapped to the business problem?
5. What will your input data be?
6. What are you trying to predict?
7. What type of machine learning task are you facing?
8. What do existing solutions look like?
9. Are there any particular constraints you need to keep in mind?

**2b After listing the key questions, write their answers for your creative brief project.**

1. Why is your customer trying to solve this particular problem?

My customer (a restaurant/ bar owner) wants to solve the problem of inventory management to maintain the restaurant running at normal capacity so as to uphold client happiness. This is difficult to do when your employees are not able to complete orders because of low stock on inventory.

1. What value will they derive from the solution—how will your model be used, and how will it fit into your customer’s business processes?

My model (inventory classification) will work best when suppliers deliver the inventory. The model will keep track of the inventory as employees use it, and automatically order new stock when it detects low levels, or when based on previous data, knows to be a busy period for the restaurant (such as during vacations).

1. What kind of data is available, or could be collected?

The available data contains images of alcoholic beverages, non-alcoholic beverages, cocktail fruits, coffee packs, and lemons and limes.

1. What kind of machine learning task can be mapped to the business problem?

Multiclass classification model.

1. What will your input data be?

The input data will be said dataset.

1. What are you trying to predict?

Correctly classify each object to the correct class.

1. What do existing solutions look like?

The employees manually keep track of and order the inventory, which often leads to understocking.

1. Are there any particular constraints you need to keep in mind?

The employees will need to be trained to get acquainted to this programme, as well as keep an eye on the inventory management situation for some time to make sure that the stock levels are at an appropriate level.

**2c Describe three different strategies to create an image dataset.**

1. Pull the dataset from user-uploaded images (for example, the most common image categories).

2. Install cameras in the environment the model will be used and have them collect tens of thousands of images.

3. Collect a database of existing points of interest, in different conditions.

**2d Describe three different strategies to annotate a dataset (data labelling). What are the pros and cons of each strategy?**

1. Manually annotating the data:

- pros: you have full control over the labels; depending on the task at hand, you can train someone to do it for you in-house;

- cons: very time-consuming; a lot more room for mistakes since you are not a professional data labeller;

2. Using a crowdsourcing platform:

- pros: inexpensive; scales well;

- cons: annotations might end up being quite noisy

3. Using a specialized data-labelling company:

- pros: time-saving; professional quality meaning better labelling practices;

- cons: expensive; your dataset might end up becoming a black box since you will not be in any way involved in the actual labelling process, leading to you not being able to perform manual feature engineering, which can be limiting.

**2e What strategy to create and annotate a dataset you are applying in your creative brief project? Explain how you are creating and annotation your dataset.**

I will be manually annotating my dataset since it is an inexpensive method to achieve this process and since the dataset itself is not that large.

This process starts with figuring out the categories that I want to classify my images into. These are: alcoholic beverages, non-alcoholic beverages, cocktail fruits, coffee packs, and lemons and limes. Then I will set out the guidelines for these categories, such as determining ‘alcoholic beverages’ using the bottles where the brand is clearly seen in the photo of the alcoholic drink, colour (i.e dark brown is most likely whiskey), etc. I am confident that most of the labels will not be at risk of confusion, besides alcoholic and non-alcoholic drinks. In such an instance, I would much rather have the model predict alcoholic drinks as non-alcoholic than vice-versa because alcoholic beverages are consumed more often than non-alcoholic, especially in bars.

The next step is the actual annotation, where each image is revised and labelled appropriately based on the categories mentioned.

**2f Describe three different strategies to understand your data.**

For image datasets:

- Take a look at a few samples directly and their labels.

- Check for target leaking.

- Printing the number of instances in each class to get a better understanding of the distribution of classes to account for an imbalance if found.

**2g As stated in the book Deep Learning with Python, "it's pretty bad practice to treat a dataset as a black box". Considering this statement, describe the strategies that you are applying in your creative brief project to understand the dataset.**

I am applying all the strategies mentioned in the previous exercise.

**2h Why data needs to be preprocessed in order to be effectively used by a machine learning model? Why not using raw data? List and explain three data preprocessed strategies used to preprocess datasets.**

Raw data is impractical to use for ML models because the output will not coincide with what we are trying to achieve. Preprocessing the raw data at hand makes it more amenable to NNs. These can be domain-specific.

Three basic data preprocessed strategies used are:

1. Vectorisation

Whatever data is used to solve the problem at hand, it must first be transformed into tensors. This is because all inputs and targets in a neural network must typically be tensors of floating-point data so the NN can understand it.

2. Value Normalisation

Typically, data input into the networks should have two characteristics:

- Take small values (most values should be in the 0-1 range)

- Be homogenous (all features should take values in roughly the same range)

Normalisation can be done in multiple ways. The most popular one are dividing by the highest value in the range so obtain values between 0-1 (such as image data encoded as integers in the 0–255 range which is then divided by 255), or to normalize each feature independently to have a mean of 0 and to have a standard deviation of 1 (this is a bit stricter).

3. Handling missing values

One method of handling missing values in a dataset is to completely remove a feature that has too many such values, and therefore is considered obsolete in solving the business problem at hand. This could result in important data loss, however.

If the feature is categorical, it’s safe to create a new category that means “the value is missing.” The model will automatically learn what this implies with respect to the targets.

If the feature is numerical, avoid inputting an arbitrary value like "0", because it may create a discontinuity in the latent space formed by your features, making it harder for a model trained on it to generalize. Instead, consider replacing the missing value with the average or median value for the feature in the dataset. You could also train a model to predict the feature value given the values of other features.

If missing categorial features are expected in the test data, but the network was trained on data without any missing values, the network won’t have learned to ignore missing values! In this situation, training samples with missing entries should be artificially generated : copy some training samples several times, and drop some of the categorical features that you expect are likely to be missing in the test data.

**2i Explain why it is important to define a baseline in a machine learning project.**

Defining a baseline to beat in a machine learning project is important because it helps perform feature selection and develop new, more useful features. Additionally, a baseline aids in selecting a good-enough training configuration, such as the right loss function, batch size and learning rate. Trying multiple architectures is key to beating the baseline.

**2j Discuss how you would select an appropriate machine learning model for a given problem. What factors would influence your decision?**

The most important factor in choosing an appropriate model is taking into account the business problem at hand and what the model will be used for, as well as considering the dataset’s variable type (images, numbers, sound, etc.). Then, deciding what type of problem we are dealing with (classification, regression, etc.), and what its complexity level is.

Additionally, I would then assess the performance of the chosen model and decide if it was a good decision or not.

**2k Describe the process and objectives of performing error analysis in a machine learning project. Why is this phase critical in the lifecycle?**

Error analysis involves examining and analysing the errors made by the model during training or inference. The primary objective is to identify patters, understand the root of the errors and improve the model itself or, if necessary, the underlying data.

This phase is critical to the lifecycle for several reasons. Error analysis is the base of achieving good model generalisation because it helps identify areas where the model might make systematic errors (overfitting, sources of bias, underfitting, etc.), leading to prioritizing development efforts where they are most needed. It also helps uncover issues in the training data itself. Additionally, error analysis contributes to higher interpretability by providing explanations for the predictions.

**2l Explain the challenges and considerations involved in deploying a machine learning model in a production environment.**

Deploying a ML model in a production environment might pose challenges such as programming languages that do not match. Therefore, one might deploy the model in the language used in the environment where it will be used. This required research beforehand. Secondly, since the production model will only be used to output predictions, rather than for training, there is room to perform various optimizations that can make the model faster and reduce memory footprint.

**2m Discuss the importance of monitoring and maintaining a deployed machine learning model. What are some common issues that might arise in this phase, and how can they be addressed?**

Monitoring a model is trivial to improving its performance and achieving even greater generalisation ability. Monitoring its behaviour, its performance on new data, its interaction with the rest of the application and, eventually, its impact on business metrics is crucial to properly implementing a machine learning model.

Common issues and their solutions include:

* Lower user engagement for ads -> randomized A/B testing to isolate the impact of the model itself from other changes (subset of cases go through the new model, while another control subset sticks to the old process);
* Making sure the model’s predictions on producing data are proper output -> a part of the production data is to be manually annotated, then comparing the model’s predictions to the new annotations OR using user surveys

Maintaining a model is important to prevent concept drifting (the gradual degrading of its performance and relevance).

Common issues and their solutions include:

* Model irrelevance -> be aware of changes in production data and new features becoming available in correlation to the label set; expand and keep annotating data.

3. Exploring the Phases of the Machine Learning Project Lifecycle

**Consider the following scenario: "you received the task of developing a machine learning model to classify images of skin lesions as either benign or malignant, which will be used to assist dermatologists in preliminary diagnostics".**

**Based on your understanding of the Machine Learning Project Lifecycle, define the following aspects related to each phase of the project lifecycle:**

**NOTE: Be creative in defining additional aspects related to the app! Try to describe as many elements of the project lifecycle as possible, but don't worry if you don't know the answer to some of the elements yet. You will study more about the machine learning project lifecycle in week 5 of this block.**

* **Project Definition:**
* **Goal and Scope: Define the primary objective and scope of the machine learning model.**

The primary objective of this model is to classify images of skin lesions as either benign or malignant, to aid dermatologists in preliminary diagnostics.

The scope is focused on developing a binary classification model. The model will analyse images of skin lesions and provide predictions of the likelihood of malignancy, which will serve as an initial assessment to be used as a base for dermatologists’ further research into the diagnostic process.

* **Stakeholder Analysis: Identify key stakeholders and their relations to the project.**

Key stakeholders include the hospital’s board of directors, the staff, the patients, and the organisation that will provide funding for the project.

* **Challenges and Benefits: Outline potential challenges and benefits of the project.**

Potential challenges: availability of training data, class imbalance, ethical considerations (such as patient privacy laws), properly integrating the system to be used in collaboration with dermatologists.

Potential benefits: better diagnostic accuracy, efficiency, time-saving, more standardized diagnostics.

* **Data Definition and Baseline Establishment:**
  + **Data Requirements: Specify image data types, formats, and potential sources.**

Types of images: skin lesions captured through various industry-capturing methods (such as clinical photography).

Formats: JPEG, PNG, DICOM (for medical images).

Potential sources: dermatology databases and repositories, hospital archives, clinical trials, crowdsourcing platforms, in-house data collection.

* + **Ethical Considerations: Address ethical aspects of using medical images.**

The main concern for the use of such images is patient privacy, as they might not agree to their photos being used for training such a model. Another concern is the issue of skin colour, as some lesions would be harder to see on dark-coloured individuals. Additionally, issues of transparency might arise.

* + **Baseline Model Concept: Propose a simple model or technique for establishing a baseline.**

For establishing a baseline in the skin lesion classification task, a simple technique is to use a random classifier. This approach sets the minimum performance expectation for any subsequent machine learning model.

The random classifier assigns a class label (benign or malignant) to each skin lesion image randomly, without considering any features or patterns in the data.

This approach represents the simplest non-machine learning technique and serves as a trivial baseline for comparison.

* + **Evaluation Metrics: Suggest metrics for baseline performance evaluation.**

Accuracy, precision, recall, F1-score, and AUC ROC.

* **Data Preprocessing and Labeling:**
  + **Preprocessing Steps: Plan steps for cleaning and preparing image data (be creative and imagine at least two types of problems to be addressed in the preprocessing phase).**

Firstly, get acquainted with the type of image data present in the dataset by sampling a couple pictures to use and seeing if there are class imbalances.

Next, to enhance the images, I would increase image contrast to improve the visibility of important details in the skin and normalise the colour distribution to reduce variability cause by different lighting conditions.

For class imbalances, I would augment the minority class by generating synthetic samples and assigning higher weights to samples.

* + **Labeling Strategy: Describe the process for labeling images, ensuring accuracy and consistency.**

Since the use of the model is in the medical field, where it is trivial to get good evaluation metrics scores to ensure a good generalisation of the mode and where experts are crucial, I would send the photos to a specialized data-labelling company.

* + **Data Augmentation: Recommend data augmentation techniques. You will learn about data augmentation next week. For now, you can read this article to have a general idea about what is data augmentation and then suggest a data augmentation technique to be used in this case.**

Increasing the brightness, increasing the image contrast, horizontally flipping the images, using a colour jitter, introducing noise (salt and pepper), and deletion.

* **Model Selection and Training:**
* **Model Selection: Propose a type of model suitable for image classification and justify your choice.**

My proposal of a model is a binary classification model, as the task at hand involves classifying images of skin lesions into one of two classes: malignant or benign.

* **Training Strategy: Outline the approach for training the model, including data splitting and validation methods.**

For data splitting, I would do the typical split ratios: 70 for training, 20 for validation (using holdout validation as I assume the dataset is sufficiently large), and 10 for testing. I would keep class imbalance in mind for this step.

I would then train the model and evaluate the performance, tuning the hyperparameters if necessary.

* **Error Analysis and Model Evaluation:**
* **Error Analysis Plan: Describe how to identify and analyze model errors for this case.**

I would use a confusion matrix to visualise the distribution of TPs, TNs, FPs, and FNs. I would then take into consideration the chosen evaluation metrics.

After inevitably uncovering problems such as certain types of lesions that the model consistently struggles to classify and start assessing them.

* **Performance Improvement: Suggest strategies for improving model performance based on error analysis.**

Increasing model capacity (implementing multiple layers in the NN, bigger layers, more epochs), hyperparameter tuning, and if all else fails, reconsidering my approach to the problem at hand.

* **Deployment Strategy:**
* **Deployment Plan: Outline how the model will be integrated into the existing healthcare systems.**

To integrate the model into existing healthcare systems, firstly I would contact the IT staff of the hospital(s) and develop an API or web service. Next, the staff would receive training about how the model works, what it can do and most importantly, what it cannot do, to make sure the expectations remain close to reality. Dermatologists would take photos of the lesions of their patients (with the patient’s written consent) that they deem suspicious, and use to model to help them make an informed decision on how to proceed. The model would then be deployed.

* **Performance Monitoring: Describe methods for monitoring the model's performance post-deployment.**

The model would be mainly monitored for reliability, as its main purpose dictates that this is what I want to increase. If patients’ lesions are correctly and consistently classified (after further investigations reveal the nature of the lesion), there would be no need for further intervention. However, as time goes on, I would gather more images to put into the dataset itself.

Naturally, I would also monitor anomalies and security.

* **Maintenance and Updating:**
* **Maintenance Plan: Propose a schedule and procedures for regular model maintenance.**

Once every couple of months, software updates, data quality assurance (very important), and model retraining would be performed. Designated personnel for reporting this information as it is happening would be assigned. Continuous user training and support would be provided to healthcare professionals because it can be difficult to fully incorporate such technology from the start.

* **Model Updating: Discuss how and when the model should be retrained or updated.**

As mentioned, every couple of months new data would be used to retrain the model and a performance analysis and comparison to the old model would be done. New technology could also offer more opportunities for improvement.

I would argue that concept drifting is not a particular problem for this model because relative stability of medical imaging techniques, consistency in lesion characteristics and standardization in diagnosis.

Finally, collaboration among healthcare professionals, data scientists, and regulatory experts is crucial in determining the optimal timing and approach for model retraining or updating to ensure its effectiveness in assisting clinicians with skin lesion diagnosis.