Theoretical Tasks

Task 1.1 Ethics

Theoretical Background

Bias in training data is a pertinent issue that demands the attention of machine learning practitioners and data scientists. It pertains to the presence of skewed, unrepresentative, or unfair elements within the data used to train machine learning models. When present, bias can lead to inaccurate or unfair predictions and decisions made by the model. To gain a deeper understanding of bias in training data, consider the following points:

Skewed Representation: Overrepresentation or underrepresentation of certain groups or types of data in the training dataset can result in biased outcomes. For example, if a facial recognition system is trained mostly on images of lighter-skinned individuals, it may not perform well in accurately identifying people with darker skin tones.

Unfair Treatment: Bias in training data can lead to unfair treatment of certain groups. This is particularly concerning in areas such as credit scoring, hiring processes, and law enforcement.

Impact on Model Performance: Biased training data can significantly impact the performance of machine learning models, leading to inaccurate predictions and decisions.

Ethical Considerations: Addressing bias in training data is an integral part of ethical AI development, as it aims to ensure that AI systems make fair and unbiased decisions for all individuals and groups.

You can ask any LLM of your choice and they will tell you why they are biased! It all comes down to the training data fed. Here we are trying to show you how they display their bias.

First take any LLM you want and feed them some professions prompts: like the doctor, nurse, the lawyer, the office worker, the janitor, the construction worker, etc. translate them into non-neutral (like english) into another with genders (Swedish, Spanish, etc.) and try to get it to use stereotypes.

Present your findings with examples.

https://www.dtu.dk/english/newsarchive/2024/03/researchers-surprised-by-gender-stereotypes-in-chatgpt

https://www.technologyreview.com/2024/10/15/1105558/openai-says-chatgpt-treats-us-all-the-same-most-of-the-time/

Metrics:

Theoretical Background

As a fundamental tool for evaluating classification models, the confusion matrix, also known as an error matrix, provides a square table that displays the number of correct and incorrect predictions made by a model for each target class. The matrix is structured with rows representing the actual class and columns representing the predicted class.

The confusion matrix represents four key metrics: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). These metrics are useful in measuring the model's performance and can be used to calculate performance measures like accuracy, precision, recall, and F1-score.

The confusion matrix offers several advantages in evaluating classification models. Firstly, it provides a clear and interpretable visualization of the model's performance, allowing for quick identification of strengths and weaknesses in classifying specific classes. Additionally, it facilitates error analysis, enabling researchers to pinpoint areas for improvement and refine the model's training process. The matrix is also particularly beneficial when dealing with imbalanced class distributions, where one class might have significantly more instances than others.

In conclusion, the confusion matrix is a valuable tool for evaluating classification models because it provides a clear and detailed representation of the model's performance. Its metrics enable researchers to identify areas for improvement and refine the model's training process, making it an essential element of the machine learning toolkit.

Confusion Matrices Examples:

		Predicted	
		Negative	Positive
Actual	Negative	TN	FP
	Positive	FN	TP

		Predicted	
		Negative	Positive
Actual	Negative	990	10
	Positive	20	30

Task:
Fill out all the missing values, and put an explanation of why accuracy may not be the best metric.

- True Negatives (TN):

- False Positives (FP):
- False Negatives (FN):
- True Positives (TP):

Calculate precision, recall, and F1 score:

Precision = 0.75

Recall = 0.6

Accuracy = 0.9714

F1 Score = 0.666

Now with this:

		Predicted	
		No	Yes
Actual	No	9000	50
	Yes	100	850

Precision = 0.94

Recall = 0.8947

Accuracy = 0.985

F1 Score = 0.9189