Assignment 3

CMPT 419 -- Nick Vincent

Nima Seifi
301 398 993

Part 1

Methods

LLM Dataset

The LLM dataset I was interested in happened to be quite small. It was the math fine-tuning data set from Nvidia: nvidia/OpenMathInstruct-1. I was able to download this onto my computer as is.

Once downloaded, I loaded it into a dataframe, iteratively took random samples and stopped taking samples once the token count had exceeded 300k.

See code in github if you like (part1.ipynb)

Breast Cancer Dataset

Since my project is going to be a data analysis tool, there isn't any 1 given dataset that is of high importance to it. Instead, for this assignment I will choose a commonly used breast cancer prediction dataset from kaggle.

Part 2

LLM Dataset

1. **What do the instances that comprise the dataset represent**

Each instance in the data is a pair of problems and solutions generated by a Mixtral-8x7B model. The problem is from 2 pre-existing datasets: GSM8K and MATH, the solutions are generated synthetically by the model.

2. **How many instances are there in total**

The data is pre-split into train and validate, there are roughly 5 million train examples and 1 million validate examples.

3. **Does it contain all possible instances or a sample from the larger set**

Since this is a synthetic dataset, this is just a sample from an infinite set.

4. **What does each instance consist of**

Each instance is a row of features, the primary ones being the problem + generated solution along with a correctness feature. There is also additional data such as dataset source, error message, and generation type. All of which should be intuitively interpretable from their names, except generation type. That feature just identifies if the solution was generated without a reference solution or with a masked reference solution (ie: did we show the model a partial solution or did it generate from scratch)

5. **Is there a label associated with each instance**

The correctness feature can be seen as a label.

6. **Is any information missing from individual instances**

There are no missing values in this dataset

7. **Are relationships between individual instances made explicit**

The notion of “relationship” in this matter is quite abstract but I’d say no. There are similar types of math problems throughout the problem set, but there is no feature mentioning which problems are similar.

8. **Are there recommended data splits**

Yes the data is already pre-split

9. **Are there any errors, sources of noise, redundancies in the dataset**

Since the solutions are synthetically generated, there is an argument for the existence of stochastic noise.

10. **Is the dataset self-contained, or does it link to or otherwise rely on external resources**

It relies on external resources. It relied on 2 external datasets for its problem space and then relied on a separate mixture of experts models to generate the solutions. But now that the data has been aggregated, it no longer needs the prior sources.

11. **Does the dataset contain data that might be considered confidential**

No.

12. **Does the dataset contain data that if viewed directly might be offensive, insulting threatening or might otherwise cause anxiety**

No.

13. **Does the dataset identify any subpopulations**

No.

14. **Is it possible to identify individuals either directly or indirectly**

No.

15. **Does the dataset contain data that might be considered sensitive in any way**

No.

16. **Any other comments**

I should mention that this is not a pre-training dataset but a fine-tuning dataset.

Breast Cancer Dataset

1. **What do the instances that comprise the dataset represent**

Patients who were tested for breast cancer alongside some relevant information about the individual at the time of taking the tests.

2. **How many instances are there in total**

1000

3. **Does it contain all possible instances or a sample from the larger set**

This is a tiny portion of all the possible instances.

4. **What does each instance consist of**

Each instance is a row of features. They are the features that are relevant to breast cancer. I won't bother enumerating them here, they are written on kaggle already: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

5. **Is there a label associated with each instance**

The label is the state of the tumor: malignant vs benign.

6. **Is any information missing from individual instances**

None are missing

7. **Are relationships between individual instances made explicit**

No.

8. **Are there recommended data splits**

No

9. **Are there any errors, sources of noise, redundancies in the dataset**

Since many of the measurements need to be measured with instruments, there will be slight noise from instrument error margins. There will also be strong correlations between features since they often describe the physical features of the nucleus. This may make some variables redundant

10. **Is the dataset self-contained, or does it link to or otherwise rely on external resources**

The data is self contained, but is hosted on Kaggle and UCI so it may rely on those for access.

11. **Does the dataset contain data that might be considered confidential**

No, the data does not include anything about people, purely information about the tumors.

12. **Does the dataset contain data that if viewed directly might be offensive, insulting threatening or might otherwise cause anxiety**

No.

13. **Does the dataset identify any subpopulations**

No.

14. **Is it possible to identify individuals either directly or indirectly**

No.

15. **Does the dataset contain data that might be considered sensitive in any way**

No.

16. **Any other comments**

No.

Part 3

Check part3.ipynb to see the full tables:

LLM Dataset

To evaluate these rows, I considered a few factors.

1. If the problem is interesting (applicability + difficulty)
2. If the solution is correct
3. If the solution can be used to teach someone how to solve the problem

I don't believe point 3 was a goal when generating this dataset, but my personal use of LLMs would have greatly benefited if it was.

Breast Cancer Dataset

For this section, it was not possible for me to determine data quality without finding a domain expert so I had to rely on purely data/statistical metrics. To make the process manual, I looked at plots which LLMs said were likely to be important. I saw clusters forming between benign and malignant cells which was a good sign, and I got a few points that exhibited abnormal behaviour and I noted those in the table.