Machine Learning Course Project (25 points)

Predicting Text Writing Dates Using Named Entities

Project Description

In this project, you will develop a machine learning model to predict the year in which a historical text was written. The task involves working with real-world data that contains features extracted from the text and metadata about its possible time of authorship. Your model will utilize person and place names mentioned in the text to predict the date of writing.

Dataset Details

1. Features Dataset:

- A one-hot encoded matrix of features, with 11599 person names (nam_id_XXXX) and 4784 place names (geo_id_XXXX) as columns.
- The first column contains unique text_id corresponding to individual texts.

https://www.dropbox.com/scl/fi/y478ga33bf5ep8aurhxe1/encoded_df_blanks_as_na.csv?rlkey=2pn5fj6istuhzeagr3nzdfyuf&st=hb9pemkn&dl=0

2. Labels Dataset:

- A separate CSV file containing for each text:

  - text_id: same as the text id in the features dataset
  - y1: The earliest possible date the text was written.
  - y2: The latest possible date the text was written.
- If y1 and y2 are equal, the date of writing is known with certainty.

https://www.dropbox.com/scl/fi/vwl5623hydnmbrgrt01zi/20240103_texts_with_dates.csv?rlkey=3llwbnca32hbw99dvm4fmsstx&dl=0

Objective

Your goal is to build a classifier that predicts the year of writing for each text. You should determine whether to predict a single year or provide a range of years based on the provided y1 and y2.

Tasks

1. Data Exploration and Preprocessing (5 points):

- Explore the distribution of the person and place features and their relationship to the target labels (y1 and y2).
- Consider strategies for handling uncertain dates ($y1 \neq y2$).

2. Feature Engineering (5 points):

- Evaluate the importance of different features and explore feature reduction techniques if necessary.
- Consider aggregating or transforming the one-hot encoded features to better capture patterns.

3. Model Development (5 points):

- Apply appropriate machine learning algorithms covered in class.

- Experiment with models to optimize accuracy and interpretability.

4. Evaluation (5 points):

- Use appropriate metrics to evaluate your model's performance, such as Mean Absolute Error (MAE) for year prediction or accuracy for year classification.
- Perform cross-validation to ensure robust results.

5. Presentation (5 points):

- Document your process, including data preparation, model development, and performance evaluation.
- Summarize your findings and provide insights into the features most predictive of the text's writing year.

Deliverables

- A well-documented Python notebook or script demonstrating your analysis, model development, and evaluation.
- A concise report summarizing your methodology, key findings, and recommendations for future work.
- A presentation slide deck to communicate your results effectively to an audience.