Assignment_part1 V2 (1)

September 16, 2025

1 Introduction

A very important aspect of supervised and semi-supervised machine learning is the quality of the labels produced by human labelers. Unfortunately, humans are not perfect and in some cases may even maliciously label things incorrectly. In this assignment, you will evaluate the impact of incorrect labels on a number of different classifiers.

We have provided a number of code snippets you can use during this assignment. Feel free to modify them or replace them.

1.1 Dataset

The dataset you will be using is the Adult Income dataset. This dataset was created by Ronny Kohavi and Barry Becker and was used to predict whether a person's income is more/less than 50k USD based on census data.

1.1.1 Data preprocessing

Start by loading and preprocessing the data. Remove NaN values, convert strings to categorical variables and encode the target variable (the string <=50K, >50K in column index 14).

[2]:
```python
import pandas as pd
import numpy as np
```

[4]:
```python
# This can be used to load the dataset
data = pd.read_csv("adult_all.csv", header=None, na_values='?')
data.head()
```

[4]:
```
    0                 1       2          3   4                   5  \
0  39         State-gov   77516  Bachelors  13       Never-married
1  50  Self-emp-not-inc   83311  Bachelors  13  Married-civ-spouse
2  38           Private  215646    HS-grad   9            Divorced
3  53           Private  234721       11th   7  Married-civ-spouse
4  28           Private  338409  Bachelors  13  Married-civ-spouse

                   6              7      8     9    10  11  12  \
0       Adm-clerical  Not-in-family  White  Male  2174   0  40
1    Exec-managerial        Husband  White  Male     0   0  13
2  Handlers-cleaners  Not-in-family  White  Male     0   0  40
```

```
3  Handlers-cleaners        Husband  Black    Male     0   0  40
4     Prof-specialty           Wife  Black  Female     0   0  40

              13     14
0  United-States  <=50K
1  United-States  <=50K
2  United-States  <=50K
3  United-States  <=50K
4           Cuba  <=50K
```

1.1.2 Data classification

Choose at least 4 different classifiers and evaluate their performance in predicting the target variable.

Preprocessing Think about how you are going to encode the categorical variables, normalization, whether you want to use all of the features, feature dimensionality reduction, etc. Justify your choices

A good method to apply preprocessing steps is using a Pipeline. Read more about this here and here.

Evaluation Use a validation technique from the previous lecture to evaluate the performance of the model. Explain and justify which metrics you used to compare the different models.

[10]:

```python
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline

# Define your preprocessing steps here
steps = []

# Combine steps into a ColumnTransformer
ct = ColumnTransformer(steps)

# show the correlation between different features including target variable
def visualize(data, ct):
    pass

# Apply your model to feature array X and labels y
def apply_model(model, X, y):
    # Wrap the model and steps into a Pipeline
    pipeline = Pipeline(steps=[('t', ct), ('m', model)])

    # Evaluate the model and store results
    return evaluate_model(X, y, pipeline)

# Apply your validation techniques and calculate metrics
def evaluate_model(X, y, pipeline):
    pass
```

1.1.3 Label perturbation

To evaluate the impact of faulty labels in a dataset, we will introduce some errors in the labels of our data.

Preparation Start by creating a method which alters a dataset by selecting a percentage of rows randomly and swaps labels from a 0->1 and 1->0.

```
[ ]: """Given a label vector, create a new copy where a random fraction of the␣
      ↪labels have been flipped."""
     def pertubate(y: np.ndarray, fraction: float) -> np.ndarray:
         copy = data.copy()
         # Flip fraction*len(data) of the labels in copy
         return copy
```

Analysis Create a number of new datasets with perturbed labels, for fractions ranging from `0` to `0.5` in increments of `0.1`.

Perform the same experiment you did before, which compared the performances of different models except with the new datasets. Repeat your experiment at least 5x for each model and perturbation level and calculate the mean and variance of the scores. Visualize the change in score for different perturbation levels for all of the models in a single plot.

State your observations. Is there a change in the performance of the models? Are there some classifiers which are impacted more/less than other classifiers and why is this the case?

```
[ ]: # Code
```

Observations + explanations: max. 400 words

Discussion

1) Discuss how you could reduce the impact of wrongly labeled data or correct wrong labels. max. 400 words

   Authors: Youri Arkesteijn, Tim van der Horst and Kevin Chong.

1.2 Machine Learning Workflow

From part 1, you will have gone through the entire machine learning workflow which are they following steps:

1) Data Loading
2) Data Pre-processing
3) Machine Learning Model Training
4) Machine Learning Model Testing

You can see these tasks are very sequential, and need to be done in a serial fashion.

As a small perturbation in the actions performed in each of the steps may have a detrimental knock-on effect in the task that comes afterwards.

In the final part of Part 1, you will have experienced the effects of performing perturbations to the machine learning model training aspect and the reaction of the machine learning model testing section.

1.3 Part 2 Data Discovery

You will be given a set of datasets and you are tasked to perform data discovery on the data sets.

The datasets are provided in the group lockers on brightspace. Let me know if you are having trouble accessing the datasets

The process is to have the goal of finding datasets that are related to each other, finding relationships between the datasets.

The relationships that we are primarily working with are Join and Union relationships.

So please implement two methods for allowing us to find those pesky Join and Union relationships.

Try to do this with the datasets as is and no processing.

[ ]:
```
def discovery_algorithm():
    """Function should be able to perform data discovery to find related datasets
    Possible Input: List of datasets
    Output: List of pairs of related datasets
    """

    pass
```

[ ]:
```

```

You would have noticed that the data has some issues in them. So perhaps those issues have been troublesome to deal with.

Please try to do some cleaning on the data.

After performing cleaning see if the results of the data discovery has changed?

Please try to explain this in your report, and try to match up the error with the observation.

[ ]:
```
## Cleaning data, scrubbing, washing, mopping

def cleaningData(data):
    """Function should be able to clean the data
    Possible Input: List of datasets
    Output: List of cleaned datasets
    """

    pass
```

1.4 Discussions

1) Different aspects of the data can effect the data discovery process. Write a short report on your findings. Such as which data quality issues had the largest effect on data discovery. Which data quality problem was repairable and how you choose to do the repair.

Max 400 words