Yu Long Wang
Homework 2 Report

Part 1: Exploratory Data Analysis (EDA) & Data Transformation (DT)

Exploratory Data Analysis (EDA)

Checking what the dataframe looks like through head, describe, and isnull.
This helps me understand the dataset like the data types, names of the columns, and missing values.

Plot target distribution to see how many malignant and benign cases there are and to check if the dataset is balanced.

Plotting the heatmap of the correlation matrix to see which features are highly correlated.

Data Transformation (DT)

Converted the target to -1 (malignant) and 1 (benign).

Dropping some features that are highly correlated (ones that are bright red in the heatmap) such as mean radius, mean perimeter, mean area (brighter than like 0.7-1).

I also notice that there are more benign cases than malignant cases, so I tried to balance the dataset by randomly getting a subset of benign points. But this doesn't help with the performance of the model, so a random subset of benign points to match the number of malignant points, does not perform well.

For now I left the dataset unbalanced.

During the experiment, I also tried to standardize the data to have a mean of 0 and a standard deviation of 1. This puts all the features in the same scale, making it easier to compare and interpret the results.

Part 2: Training and Validation

Experiment:
loaded the breast cancer dataset, converted the target to -1 (malignant) and 1 (benign), and then split the data into training and validation sets. Also removed features that are highly correlated.

Ran the experiment for different features, one without the data transformation of removing features that are highly correlated, and one with the data transformation. For each of these

experiments, I also ran the experiment with and without standardizing the data. Using pocket algorithm.

- No data transformation, no standardization
- No data transformation, standardization
- Data transformation, no standardization
- Data transformation, standardization

I also ran the experiment for different sample sizes, from 50 to 500 in increments of 50 (10 different sample sizes).

I split the data into training and validation sets for each sample size.
80% for training, 20% for validation.

Since there is a total of 569 data points:
- 455 points for training
- 114 points for validation
Going higher than 500 for the sample size doesn't change the results much.

Ein and Eout evaluation:

My Ein and Eout for the dataset with data transformation does not perform as well as the one without data transformation.

My Ein and Eout gap narrows a bit which shows some improvement, but not much.
It also does not reach a full convergence.
The error rate is higher than when it is not data transformed. (0.4)

Compared to the Ein and Eout without data transformation it has a lower error rate (0.05), more stable Ein and Eout, and reaches a full convergence.

So I am removing too many features and or features that are highly correlated that may be important for the model. It was not good to remove the features that I had selected.

Discussion

I observed that the performance of the non transformed dataset plot showed better results than the transformed dataset plot.

The transformed dataset plot had a higher Ein and Eout error rate and a higher Ein and Eout gap.

All this was unexpected since I thought that removing features that are highly correlated would help the performance of the model but it did not but could be due to the features that I had

selected. I learned that it is important to select the right features to remove or performance would decrease. As I did not expect the error rate to increase that high.

In this case, having more complex features still performs better than removing features that are highly correlated. So making it less complex by removing features may not always perform better, then again it depends on the features that you are removing.