Group 9

STAT 451 Group Project Proposal

Our project aims to apply machine learning techniques to the Wisconsin breast cancer research dataset in order to accurately classify cases and pinpoint the most significant factors. Analyzing relationships between features and diagnoses aims to advance diagnostic precision and deepen our understanding of malignant and benign cell nuclei.

Dataset

We will use the Wisconsin Breast Cancer dataset [1], developed by two computer scientists and a surgeon here at the University of Wisconsin-Madison. It comprises 569 samples, each with a diagnosis label (M for malignant, B for benign) and 30 real-valued attributes, which have been computed from the analysis of digitized images obtained from fine needle aspiration (FNA) of breast masses, detailing the characteristics of cell nuclei within the images. We can download and read the dataset into a pandas data frame as follows:

```
pip install ucimlrepo
from ucimlrepo import fetch_ucirepo

breast_cancer_wisconsin_diagnostic = fetch_ucirepo(id=17)
X = breast_cancer_wisconsin_diagnostic.data.features
y = breast_cancer_wisconsin_diagnostic.data.targets
```

Methodology

The primary goal in this project is to train models that accurately classify new samples as benign or malignant. For this goal, we will use and try different classification algorithms like logistic regression, SVMs, kNN, and decision trees. We can also try other methods to improve the results, such as normalization of the dataset, balancing of classes, or even combinations of models.

In addition to high classification accuracy, our goal is to determine whether all 30 attributes are equally important for diagnosing breast cancer. We seek to gain deeper insight into the causes of breast cancer, determining which attributes hold more significance than others.

Lastly, given the criticality of the application, we will also want to do a more fine-grained assessment of our classifiers rather than just checking the overall classification error. We will thus study the confusion matrices of our classifiers to ensure that they prioritize minimizing false negatives over false positives.

References

[1] William H. Wolberg, W. Nick Street, Olvi L. Mangasarian: Breast Cancer Wisconsin (Prognostic). UCI Machine Learning Repository, 1995. https://doi.org/10.24432/C5DW2B.