Classification and Prediction of 6-Month Mortality in Stroke Patients

Aidan Horvath & Henry Luong

Overview:

Stroke is a common, serious medical condition with a variety of causes, treatments, and overlap in symptoms with other medical conditions. As such, a tool for rapid assessment of patients presenting with stroke symptoms could be useful for prioritizing specific diagnostic paths or early treatments. Additionally, stroke occurs with a wide degree of severity, both in the initial event and with the possibility of recurrent episodes. Machine learning could also be useful in predicting whether the Bamford classification of a patient's stroke and if they are likely to experience another in the days following their first occurrence.

Our aim for the Milestone II project is to accurately label the different types of strokes while predicting the 6-month mortality rate in each patient. We will be performing a semi-supervised learning approach on individual patient data from the International Stroke Trial (IST) database. The trial was conducted back in 1991-1996 with over 19,000+ deidentified patient records. By incorporating unsupervised learning methods to identify the underlying structure or distribution of the data, we hope to select and extract certain features that will allow us to segue into a supervised framework for model prediction and multi-classification.

Supervised learning

Learning approaches (subject to change):

- Dimensionality reduction for Naive-Bayes classifiers will be used as a benchmark against other classification models.
- MinMaxScaler or Standard scaler for Kernelized Support Vector Machines to help classify the types of strokes for non-linear decision boundaries.
- MinMaxScaler or Standard scaler for Gradient-boosted Decision Tree for an ensemble approach that will allow us to identify both stroke types and 6-month mortality.
- ROC Curve and Precision-Recall to identify the binary error rates as well as the True Positives, False Positives, True Negatives and False Negatives.
- Classifier Decision/Prediction functions and a Multi-Class confusion matrix to visualize and evaluate the performance of the classification models.

Unsupervised learning

Formulating questions:

- Are there various features that tend to outperform others when classifying either stroke types or predicting the 6-month mortality of each patient?
- Which inflictions/indicators within the dataset tend to co-occur?

Data Manipulation:

- One hot/Label encoding and Data Imputation

Unsupervised Learning Approaches:

- K-means clustering through MinMax Scaling or Standard Scaling as a feature representation.

- Dimensionality Reduction such as t-SNE and Principal Component Analysis.

Evaluation/Visualization

- Heatmap and scree plot to identify the best principal components.

Challenges:

- One of the challenges that we might face is how to create a classifier that can be generalizable with a high-dimensional dataset.
- Another challenge will be preventing data leakage, as several of the features could be proxies for outcome labels, e.g. “Infarc visible on CT” would denote an Ischaemic stroke.
- We must also be aware of any class imbalance and missing data that may occur within the dataset.

Timeline and Contribution:

	Wk2	Wk3	Wk4	Wk5	Wk6	Wk7
Both contributing equally	Data preparation and cleaning	Unsupervised learning	Supervised learning	Data training	Parameter tuning and evaluation	Analyzing and summarizing report