Project 2

Nikhil Sharma

Introduction

The goal of Project 2 was to make the class more familiar with exploratory data analysis and pre-processing, as well as utilizing various classifiers and fitting data to compare and contrast the results. The data was uploaded to a dataframe in pandas which allowed for visual and statistical analysis. All relationships between variables were visualized using matplotlib and seaborn. The dataset was then fitted to the different classification models using sklearn, specifically the K-nearest neighbor classifier, logistic regression, and Naive-Bayes. Lastly, the precision, recall, f1-score, and support values of the training and test data were outputted to determine the goodness of fit for each model.

Data Preparation

As mentioned previously, the data was uploaded to a pandas dataframe, where it could be parsed into columns and analyzed. The data contained 286 rows, or patients, and 10 columns—class of the cancer, the age of the patient, the menopause status, the size of the tumor, the number of invasive nodes, the nodal capsulation, the degree of malignancy, which breast the tumor is in (left or right), which quadrant of the breast the tumor is in, and the irradiation status (yes or no). Initially, the data contained nine columns of type object and one column of type int64.

After checking for duplicate entries, there were 14 found. After removing these entries, there were only 272 rows, or unique patients.

For each variable, the unique values were explored to determine if there were any null or missing values in each column. When seeing ? as a unique value for node-caps and breast-quad, it denoted that values were missing and needed to be replaced. For numerical values, they can simply be replaced by the mean or median, but due to the lack of order in the categorical variables, the missing values had to be replaced by the mode, or the highest occurring value.

The dataset was also transformed using one-hot encoding on each categorical column. This allowed for categorical variables to be represented as booleans of value true or false. In the models, they are represented as 0 or 1. Some examples of the new variables created from the one-hot encoding are age_30-39, menopause_lt40, and breast-quad_right_low.

Statistical analysis was performed on each variable in the dataset to view univariate behavior. A histogram of each variable was created to see the distribution of values. The variable class only has two unique values: no-recurrence-events and recurrence-events. There is not an even distribution, as no-recurrence-events is the clear mode. Variables like age, tumor-size, and inv-nodes had numerical ranges for their values, theoretically allowing for a more traditional histogram. However, because the datatype was object, the bins were unordered. Even so, there were unimodal peaks present for each plot.

Training the Model

Once the exploratory data analysis and pre-processing were complete, the breast cancer data was split into training and test datasets using an sklearn method. 30% of the data was reclassified to test data, while the other 70% remained as data to train the model. For each different classifier, whether it was K-nearest neighbor, logistic regression, or Naive-Bayes, a model was then created, and the feature and label training data were fit to the model. If the features from new data, such as the test data, were inputted into this model equation, a predicted breast cancer recurrence output would be calculated.

For the K-nearest neighbor model, a parameter grid was created to determine the most accurate value of k for the model. The range of k was set from 1 to 100 and the optimal value that was returned via the best estimator feature was 1. Additionally, a grid search cross validation was created to test each of the values in the parameter grid. A 5-fold cross validation was used with the primary scoring metric being the recall metric, whose importance will be explained in the performance of the model.

Model Performance

In order to quantify the performance of the model for each classifier, the precision, recall, f1-score, and support values were outputted from the classification report. However the most important statistic out of these four is the recall metric as we are looking to minimize the number of false negatives. When putting the numbers into the context of breast cancer recurrence, we want to make sure that patients with recurrent breast cancer are being informed that they have it rather than not being informed, as a false negative in this situation can lead to health risks and potential death.

Here are the results from the classification report for each model tested:

		Precision	Recall	F1-Score
K-Nearest Neighbor	Test	0.39	0.39	0.39
	Train	0.95	1	0.97
Logistic Regression	Test	0.36	0.17	0.24
	Train	0.63	0.33	0.43
Naive-Bayes	Test	0.32	0.96	0.48
	Train	0.36	1.00	0.52

The K-nearest neighbor classifier model, the training data displayed significantly better metrics than the test data. The recall metric for the training dataset was 1.00, showing a strong ability to detect false negatives. However, the recall metric for the test data was only 0.39, which is the biggest disparity between test and training recall metrics out of the three different models.

The recall values for both the training and test data for the logistic regression model are both low, at 0.33 and 0.17, respectively. This shows that the number of true positives identified are low. For the test data, the precision and f1-score are both low, but are relatively higher for the initial training data.

For the Gaussian Naive-Bayes model, the recall is extremely high, suggesting that the model was very accurate in determining both true positives and minimizing false negatives. The precision scores for training and test data were both lower than those of the logistic regression model, and the f1-score for both datasets were around 0.5.

In conclusion, the most accurate of the three models tested was the Gaussian Naive-Bayes model due to its extremely high recall metric. Of the four possible results—true positive, false positive, true negative, and false negative—the most important to minimize is the false negative in the context of breast cancer. The recall metrics of 0.96 and 1.00 show that this model can keep doctors accurate in determining breast cancer recurrence status for their patients. While the K-nearest neighbor recall metric for training data was also 1.00, it showed inconsistent results with the test data. As a result, I would recommend the Gaussian Naive-Bayes model to predict breast cancer recurrence in patients.

References

Class repository: https://coe-379l-sp24.readthedocs.io/en/latest/index.html
Dataset: https://raw.githubusercontent.com/joestubbs/coe379L-sp24/master/datasets/unit02/project2.data