Lab 08: Classification
CS3300 Data Science

Learning Outcomes

1. Select, apply, and interpret appropriate visual and statistical methods to analyze distributions of individual variables and relationships between pairs of variables.

2. Communicate findings through generated data visualizations and reports.

3. Determine and apply appropriate experimental setup, evaluation metrics, and models for supervised learning problems.

4. Perform and interpret feature selection to identify relationships between features and predicted variables.

5. Apply methods to real world data sets.

Overview

You are going to build a Logistic Regression model to predict whether breast tissue samples are malignant or benign. You will perform an exploratory data analysis (EDA). You will then set up an experiment, perform feature selection, and then train and evaluate a final model.

The data set is available here:
https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

Use the wdbc.data and wdbc.names files.

All of the features are numerical, and the data set contains no missing data. The Diagnosis field is categorical and is the output variable you are predicting. Ignore the ID column -- this contains unique sample IDs and will be useless as a predictor. The 30 numerical features are engineered from 10 attributes by calculating the mean, standard error, and worst or largest of each attribute for a sample.

Instructions

I want you to create your own notebook analyzing the data. In particular, you will try to identify variables that will be good predictors of the fuel economy (mpg) of the car. The basic steps will involve:

1. Load, transform, and clean the data. You will need to set the column names using the descriptions in the `wdbc.names` file. Describe the final DataFrame using `head()` and `info()`.

2. Characterize each variable. Plot the distributions of each variable and describe the range of values.

3. Explore the relationships between each variable and the tumor type. Choose the appropriate plots based on the variable types (e.g., categorical, numerical, ordered.) Indicate which variables are strongly correlated, somewhat correlated, or not correlated.

4. Set up your experiment and divide the data into training and testing sets.

5. Scaling the data.

6. Create three logistic regression models:

a. A baseline or null model

b. A model built using a "greedy" approach. Build a model for each variable individually. Sort the variables from least error to most error vs the baseline. Starting with the baseline model, add each variable one at a time to the model. If variable improves the model over the last model, keep that variable in the model. If the variable does not improve the accuracy, skip that variable. At the end, you should have a single model with multiple variables.

c. Lastly, build a model using all of the variables.

7. Evaluate the models

a. For three models, calculate accuracy and plot ROC curves. Which model has the lowest RMSE? Which model has the highest RMSE?

b. Calculate the precision, recall, and a confusion matrix for each model. What kinds of errors (e.g., false positives or false negatives) are the models prone to making?

c. Assume you were creating a model to be used for diagnosing patients in a clinical setting. What types of errors (false positives or false negatives) would be better? Justify your reasoning.

I will be will looking for following:

- Write an introduction (including your own statement of the problem) and a written summary of your results at the top of the notebook in Markdown. Make sure to put your name at the top of the notebook.
- That you successfully imported the data and verified that the correct shape of the data
- Used the appropriate plots to investigate the distributions of each variable.
- Used the appropriate plots to investigate relationships between the other variables and mpg.
- Thoughtfully evaluated the models.
- Overall, I want to see a finished, relatively polished product. Use Markdown cells in appropriate places to explain what you are doing, interpret your results, and describe your conclusions. One way to verify that your notebook is in a good state is to restart the kernel and re-run everything. This ensures that all of the necessary code is there.

Submission Instructions

To submit, please save the notebook as a HTML file and upload a zip file of the HTML file to Blackboard.

Rubric

Followed submission instructions	10%
Data Loading	10%
Exploratory Data Analysis (EDA)	20%
Experimental Setup and Data Transformation	20%
Logistic Regression Models	20%
Model Evaluation and Reflection	10%