Wisconsin Breast Cancer

Martin Skarzynski

Johns Hopkins University

April 24, 2018

Introduction

- Presenting the *Wisconsin Breast Cancer* dataset

Data Description

- I obtained the Cervical Cancer dataset from the UCI Machine Learning Repository
- Number of observations: 569
- Number of features: 32
- Number of target variables: 1
- All of the variables are already numeric, but there are some missing values.

Methods

- I first read in the data using the `pandas` library from the dataset url.
- I then selected the target variables, `Diagnosis`, and made sure the other targets were not include among my predictor variables.
- Next, I compared the number of positive (M=Malignant) to negative (B=Benign) cases in the `Diagnosis`.
- After confirming that the data are ready for modeling, I separated the data into test and train subsets using the `train_test_split()` function from the `scikit-learn` library using a `test_size` of 0.2 (or 20% of the data).
- Algorithms tested were:
  - Logistic Regression with Principal Component Analysis (PCA),
  - Linear and RBF Support Vector Machine (SVM),
  - Decision Tree

Results

- K-fold cross-validation was used to assess the Logistic Regression models
- The modeling was described using
  - Learning and Validation Curves,
  - a Confusion Matrix, and
  - a Receiver Operating Characteristic curve
- In the first step of the analysis, the logistic regression had a high prediction accuracy (roughly 96%)
- The SVM is the best performing method in the final analysis

Conclusions

- The best performing model has a high level of prediction accuracy (roughly 99%)
- With further refinement on a larger dataset can be a very useful tool to inform business decisions related to credit.