Project 3 Proposal - Classification and Interactive Dashboards
Breast Cancer Tumor Diagnosis - Malignant or Benign?
Flora Xinru Cheng
October 15, 2019

Domain

For this project, I would like to implement and compare different classification algorithms, to see which one gives the highest accuracy when predicting whether a breast cancer tumor is malignant or benign. Even though this is not directly classifying image data from hospitals, the insights provided could still be valuable to the medical field. In suggesting which model gives the best accuracy for this type of problem, we could potentially help speed up the diagnosis process for breast cancer and other similar types of cancers, and reduce patient waiting times.

Learning from previous projects, I deliberately cut down on project design and data collection time for this project, and narrowed down the scope, in order to perform a more thorough analysis on the results. I'm also using the cookiecutter template for the first time in this project, in order to improve file organization and increase efficiency. I also plan to extend the findings with a Flask website and/or Tableau visualization during the second week of the project.

Data

The main dataset used for this analysis is the Breast Cancer Wisconsin (Diagnostic) Data Set, from UCI Machine Learning, downloaded from Kaggle. The features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

The dataset has 33 features, 569 observations. We are mainly interested in the Diagnosis (M = malignant, B = benign) which would be the target for classification (this is also the only column with non-numeric values), and the ten features computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area

e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

There is also the option of introducing a related but smaller dataset: Breast Cancer Wisconsin (Prognostic) Data Set, which has additional information -- Outcome (R = recur, N = nonrecur), Time (recurrence time if field 2 = R, disease-free time if field 2 = N).

Minimum Viable Product

An MVP for this project would classify observations as benign or malignant using at least 3 cross-validated classification models. This would likely be done with a smaller number of features. A later version would involve the entire dataset, a larger set of classification algorithms to choose from, and benefit from feature engineering and model tuning.

Tools

Python, Pandas, NumPy
SciKit-Learn, Matplotlib, Seaborn
Flask, Tableau
GitHub
Cookiecutter (datasciencemvp template)
Jupyter Notebook