

Breast Cancer Prediction Dataset

(<https://www.kaggle.com/merishnasuwal/breast-cancer-prediction-dataset>)

Team 8










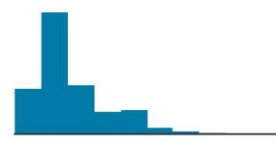


Luke Maccariello, James Cooper, Geoffrey Bonnanzio and Mathias Helder

What is the problem?

- The task is to use machine learning to correctly diagnose breast cancer
- Diagnosis for breast cancer begins when an abnormal lump is found or a tiny speck of calcium is detected.
- Our dataset was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg.
- The dataset contains numeric values for the mean of radius, texture, perimeter, area, and smoothness. It also provides the diagnosis from the values.



Our Dataset

# mean_radius 	# mean_texture 	# mean_perimeter 	# mean_area 	# mean_smoothness 	# diagnosis 
 6.98 28.1	 9.71 39.3	 43.8 189	 144 2.5k	 0.05 0.16	 0 1
17.99	10.38	122.8	1001.0	0.1184	0
20.57	17.77	132.9	1326.0	0.08474	0
19.69	21.25	130.0	1203.0	0.1096	0
11.42	20.38	77.58	386.1	0.1425	0
20.29	14.34	135.1	1297.0	0.1003	0



Why is it challenging?

- The first step to cancer diagnosis is detecting a tumor (mammograms or MRIs).
- If a tumor is detected, it needs to be analyzed to determine if it is benign or malignant. This step is done through biopsy, taking a sample of the tumor which is then analyzed in a lab. Doctors will use that sample to determine whether it is malignant or not and what kind of treatments need to be applied.
- The most challenging part of ML approach is how to increase the effectiveness of determining whether a tumor is benign or malignant.

(Credit: <https://umm-csci.github.io/senior-seminar/seminars/fall2020/northwood.pdf>)



Why is it challenging?

Which model is the BEST fit for our project?

- Logistic Regression
 - Very powerful modeling tool
 - Assess the likelihood of a diseases or health condition as a function of a risk factor
 - Used primarily for predicting binary or multiclass dependent variables.
- Support Vector Machine
 - A classifier which divides the datasets into class to find a maximum hyper plane via the nearest data points
- Decision Tree
 - A predictive modeling tool that can be applied across many areas
 - Can split dataset in different ways based on the different conditions
- Random Forest
 - An ensemble of decision tree algorithms
 - The most popular and widely used ML algorithms given its good or excellent performance across a wide range of classification and regression predictive modeling problems

What is our approach?

Breast cancer prediction is an example of a binary classification problem

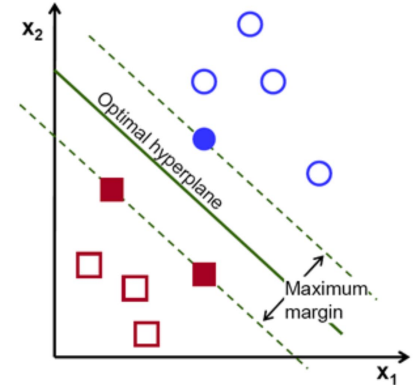
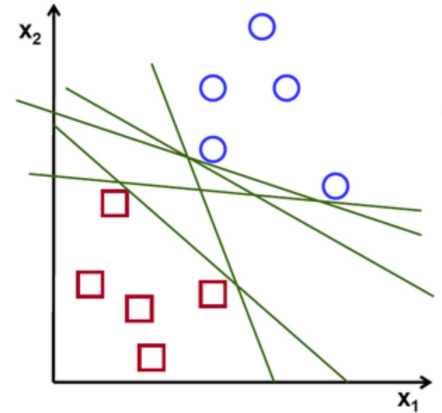
- Outcomes: benign or malignant
- Supervised machine learning

We will employ a support vector machine because we have a fair number of features (30)

with a relatively small number of training examples (569 total patients)

- Effective in high dimensional space
- Uses a subset of training points to create support vectors (memory efficient)

Will use a subset of data to test logistic regression to compare models



How to train and test our model?

- There is no data preprocessing needed for this dataset.
- To train the model we will use gradient descent.
- We will split our data set into 2 parts one for training and the other for testing.
 - This split will be 66% and 33% respectively
 - Further testing will be done for other breakdowns
- The evaluation metric we will use will be accuracy rate as it is suited for a binary classification problem.
- We can compare our model to other breast cancer prediction models after completion.



Questions?

