Breast Cancer Wisconsin Analysis

Dingchen Sha (dsha2@illinois.edu)
Tengyuan Liang (tl13@illinois.edu)
Shide Pu (shidepu2@illinois.edu)
Jianfei Wang (jwang268@illinois.edu)

Abstract

> Statistical learning methods are applied to the measurements of the breast tumor data in an attempt to accurately distinguish malignant and benign breast cancer types. A variety of machine learning techniques were explored and validated. Simple methods like logistic regression demonstrate the most promising precision, but further data collection and analysis are recommended.

Introduction

Breast cancer is the leading type of cancer in women, accounting for 25% of all cases. Survival rates in the developed world are high, with between 80 and 90% of those in England and the United States alive for at least five years[1]. There are two types of breast cancer tumors: non-cancerous, or “benign”, and cancerous, which are “malignant”[2]. The classification of breast tumors is meaningful in medical treatment, and it can help doctors to make medical decisions. In medical practice, the diagnosis of breast cancer tumors can be classified by regarding various measurements of breast tumors, including the radius of the tumor, area of the tumor, gray-scale values, etc. Accurately detecting breast cancer allows patients to receive appropriate treatment in the early phase of the disease, increasing patients’ survival rate.

In an attempt to construct a tool to diagnosis tumors into benign and malignant, statistical learning techniques have been applied to measures of the breast tumor data from Breast Cancer Wisconsin (Diagnostic) Data Set. The goal of this study is to classify based on characteristics and measurements of breast tumors metrics. The results show potentials for further application of pre-diagnosis breast cancer, but a more extensive data collection and a more comprehensive analysis are recommended.

Methods

Data

The data is obtained from UCI Machine Learning Repository [3]. It is mainly authored by the following three scientists:

1. Dr. William H. Wolberg, General Surgery Dept. University of Wisconsin, Clinical Sciences Center Madison, WI 53792
2. W. Nick Street, Computer Sciences Dept. University of Wisconsin, 1210 West Dayton St., Madison, WI 53706
3. Olvi L. Mangasarian, Computer Sciences Dept. University of Wisconsin, 1210 West Dayton St., Madison, WI 53706

Each observation in the data was classified as either benign or malignant. Patients’ ID (variable `id`), standard deviation of measurements, and `X` were also not included for model training.

Ten related attributes are preserved, among which attribute `diagnosis` is the object of interest.

Exploratory data analysis can be found in the appendix.

Modeling

The presence status of breast tumor is reflected as a binary response.

Four classification methods are established and examined:

- K-Nearest Neighbors;
- Random Forest;
- Stochastic Gradient Boosting;
- Logistic Regression.

Evaluation

Models were tuned using 5-fold cross-validation. All models were tuned to maximize area under the ROC curve. `Specificity` (true negative rate) with `malignant` as the negative class was also essential in evaluating models. Models were ultimately evaluated based on their accuracies in detecting malignant breast cancer.

Results

The results below suggested performances among Random Forest, Stochastic Gradient Boosting, and Logistic Regression vary by a small amount. Also, Random Forest and Logistic Regression demonstrate similar predictive powers. The Random Forest is selected as it classifies benign tumors slightly more precise than Logistic Regression does. Additional intermediate tuning results can be found in the appendix.

Model	Tuning	ROC	Specificity
KNN	k = 9	0.9396	0.7671
Random Forest	mtry = 6	0.9808	0.9126
Stochastic Gradient Boosting	interaction.depth = 2, n.trees = 150	0.9874	0.9010
Logistic Regression	N/A	0.9825	0.9124

Table: **K-Nearest Neighbors**, Cross-Validated Binary Predictions versus Response, Percent

	B	M
Predict: B	60.0000	8.7912
Predict: M	2.1978	29.0110

Table: **Random Forest**, Cross-Validated Binary Predictions versus Response, Percent

	B	M
Predict: B	59.3407	3.2967
Predict: M	2.8571	34.5055

Table: **Stochastic Gradient Boosting**, Cross-Validated Binary Predictions versus Response, Percent

	B	M
Predict: B	59.5604	3.5165
Predict: M	2.6374	34.2857

Table: **Logistic Regression**, Cross-Validated Binary Predictions versus Response, Percent

	B	M
Predict: B	59.1209	3.2967
Predict: M	3.0769	34.5055

Discussion

The table below summarizes the results of the random forest model on a held-out test dataset. Even though results seem statistically

satisfactory, practical use of this classification method is not recommended for reasons that the cross-validated results still indicate a certain amount of failure in malignant tumor detection. As can be seen from the results, there also exists a small chance that a benign tumor is misdiagnosed as cancer, which, although not the primary concern, still can cause unnecessary medical and financial losses.

Another issue is that the data set only contains 569 observations. This insufficiency of data might not be a good reflection of reality. Moreover, the random forest classification does not have an overwhelming advantage over stochastic gradient boosting and logistic regression in terms of their classification accuracies. It is possible that the random forest classification would be outcompeted by the other three methods based on a different dataset.

In addition to the above, the fact that all observations were collected from the state of Wisconsin has dramatically diminished the data's representativeness. Furthermore, the information was donated to the UCI Machine Learning Repository in 1995. Given that the medical knowledge and technologies have been improved dramatically over past twenty years, the practicability and usefulness of this study in the present days are worth reconsidering.

Table: Test Results, **Random Forest**, Percent

	B	M
Predict: B	57.8947	2.6316
Predict: M	7.0175	32.4561

Appendix

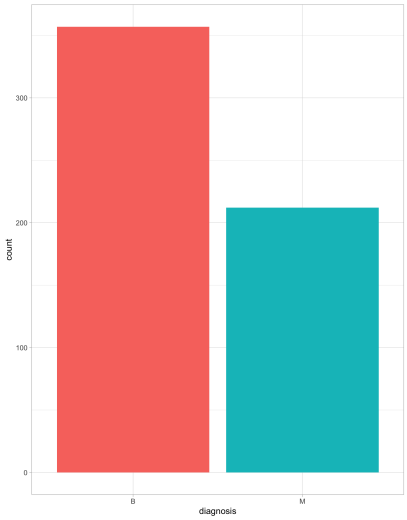
Data Dictionary

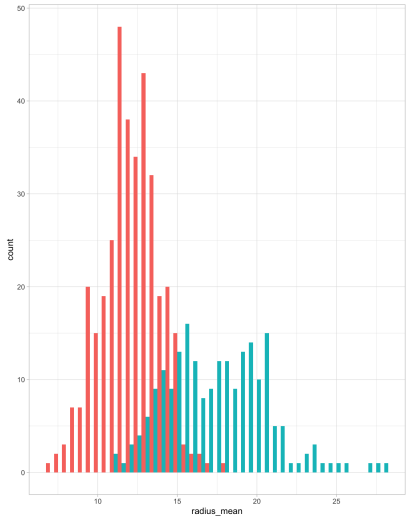
- `diagnosis` - The diagnosis of breast tissues (M = malignant, B = benign)
- `radius_mean` - mean of distances from center to points on the perimeter
- `texture_mean` - standard deviation of gray-scale values
- `perimeter_mean` - mean size of the core tumor
- `area_mean`
- `smoothness_mean` - mean of local variation in radius lengths
- `compactness_mean` - mean of perimeter ^ 2 / area - 1.0
- `concavity_mean` - mean of severity of concave portions of the contour
- `concave points_mean` - mean for number of concave portions of the contour
- `symmetry_mean`
- `fractal_dimension_mean` - mean for "coastline approximation" - 1

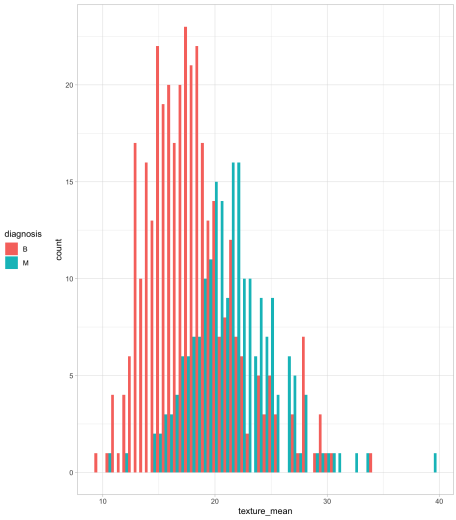
See the UCI website for additional documentation.

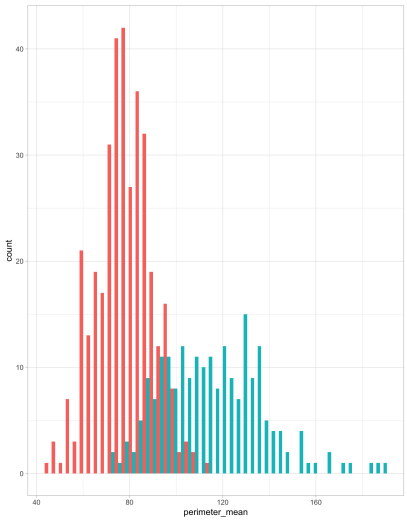
EDA

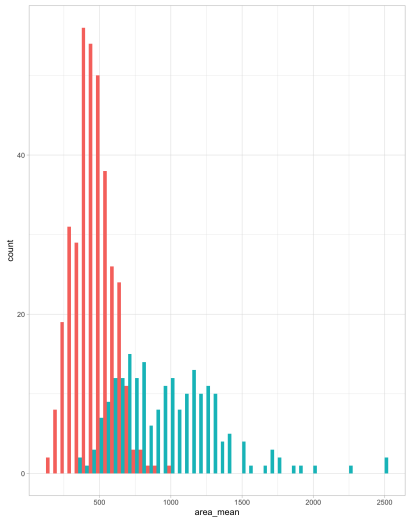
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

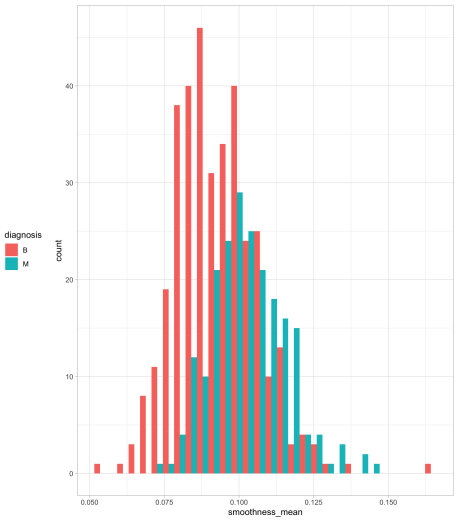

count
300
200
100
0
B
M
diagnosis

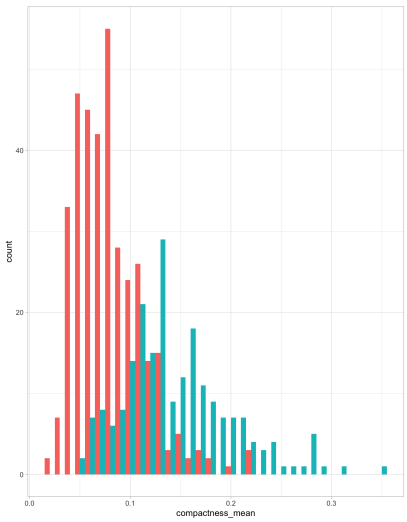

count
50
40
30
20
10
0
10
15
20
25
radius_mean
diagnosis
B
M

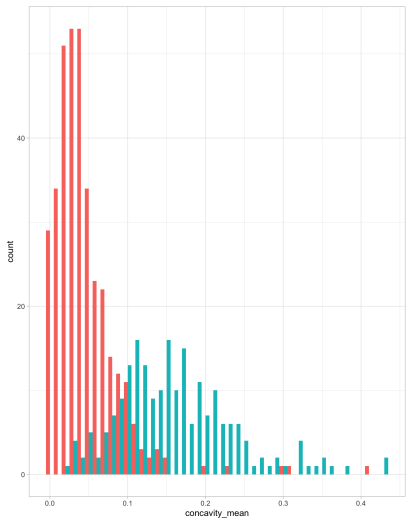

count
20
15
10
5
0
10
20
30
40
texture_mean
diagnosis
B
M

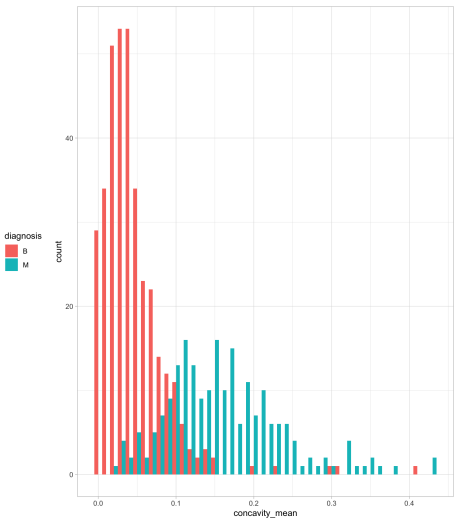

count
40
30
20
10
0
40
80
120
160
perimeter_mean
diagnosis
B
M

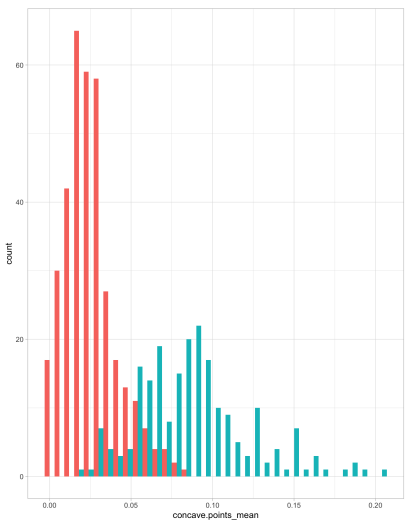

count
40
20
0
500
1000
1500
2000
2500
area_mean
diagnosis
B
M

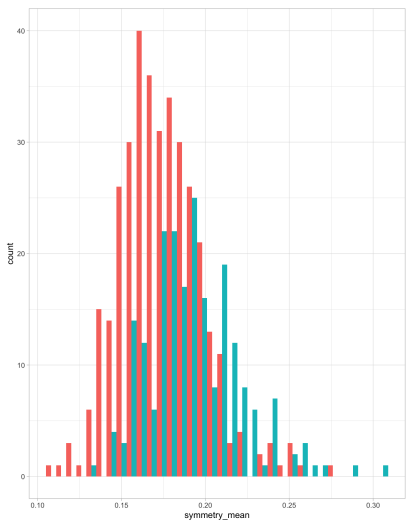

count
40
30
20
10
0
0.050
0.075
0.100
0.125
0.150
smoothness_mean
diagnosis
B
M


count
40
20
0
0.0
0.1
0.2
0.3
compactness_mean
diagnosis
B
M


count
40
20
0
0.0
0.1
0.2
0.3
0.4
concavity_mean
diagnosis
B
M


count
40
20
0
0.0
0.1
0.2
0.3
0.4
concavity_mean
diagnosis
B
M


count
60
40
20
0
0.00
0.05
0.10
0.15
0.20
concave.points_mean
diagnosis
B
M


count
40
30
20
10
0
0.10
0.15
0.20
0.25
0.30
symmetry_mean
diagnosis
B
M

Additional Tuning

Table: Stochastic Gradient Boosting Binary Classification

	shrinkage	interaction.depth	n.minobsinnode	n.trees	ROC	Sens	Spec	ROCSD	SensSD	SpecSD
1	0.1	1	10	50	0.982	0.968	0.889	0.015	0.034	0.061
4	0.1	2	10	50	0.984	0.958	0.918	0.014	0.040	0.064
7	0.1	3	10	50	0.983	0.961	0.918	0.011	0.040	0.039
2	0.1	1	10	100	0.985	0.972	0.907	0.013	0.027	0.039
5	0.1	2	10	100	0.985	0.961	0.918	0.014	0.040	0.044
8	0.1	3	10	100	0.985	0.968	0.913	0.011	0.036	0.036
3	0.1	1	10	150	0.987	0.968	0.907	0.011	0.034	0.039
6	0.1	2	10	150	0.987	0.958	0.907	0.013	0.040	0.057
9	0.1	3	10	150	0.987	0.968	0.907	0.011	0.036	0.039

Table: KNN Binary Classification

k	ROC	Sens	Spec	ROCSD	SensSD	SpecSD
5	0.938	0.947	0.790	0.027	0.025	0.054
7	0.937	0.968	0.767	0.022	0.015	0.060
9	0.940	0.965	0.767	0.021	0.033	0.060

Table: Random Forest Binary Classification

mtry	ROC	Sens	Spec	ROCSD	SensSD	SpecSD
2	0.980	0.961	0.907	0.017	0.040	0.038
6	0.981	0.954	0.913	0.014	0.042	0.036
10	0.978	0.961	0.907	0.019	0.047	0.038

Table: Logistic Regression Binary Classification

parameter	ROC	Sens	Spec	ROCSD	SensSD	SpecSD
none	0.983	0.951	0.912	0.014	0.034	0.047

1. Wikipedia: Breast cancer↩
2. Wikipedia: Neoplasm↩
3. UCI: Breast Cancer Wisconsin (Diagnostic) Data Set↩