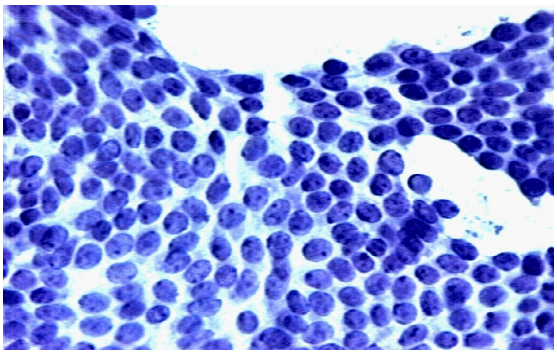
Prediction Accuracy and Sensitivity Analysis on Different Models for Breast Cancer Wisconsin Data Set

Shiyu Liu
shiyu_liu@brown.edu
Data 1030 | Fall 2019



Problem Narrative

Forecasting breast cancer can significantly increase the survival rate of patients, and classifying the sample tumor cells as malignant or benign is one of the best ways to make such accurate predictions. Breast Cancer Wisconsin from UCI Machine Learning Repository was chosen as the dataset to implement machine models and predict diagnosis result on practical cases. The target variable in this dataset is the final diagnosis whether the tumor is malignant or benign. This target variable can be transformed to a binary variable, indicating classification should be applied to this problem.

Dataset Description

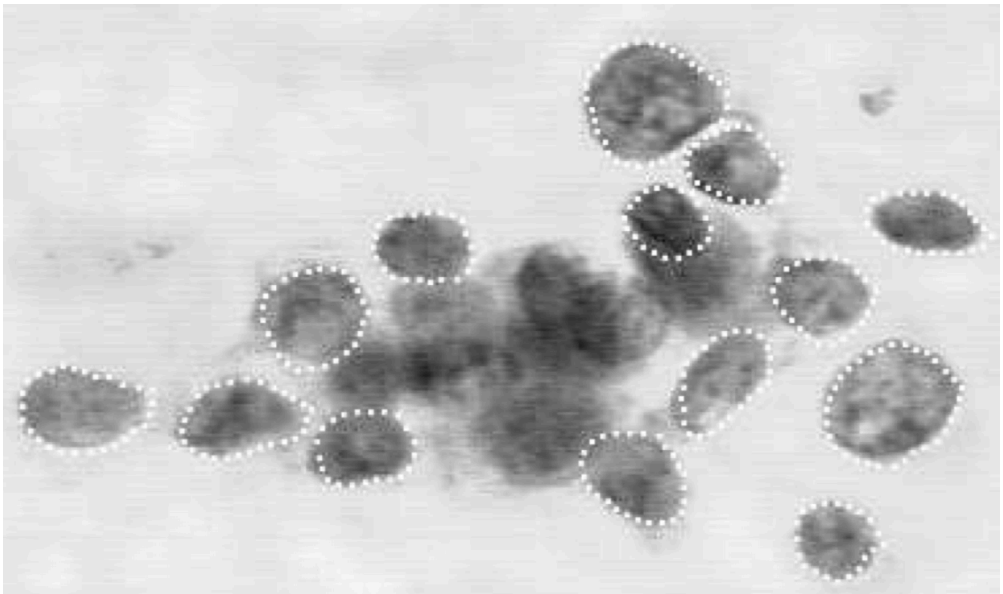


Figure 1. Cells obtained through breast fine needle aspirate. The boundaries of the cells were outlined.

There are 30 real-valued columns as independent variables with three properties, mean, standard error and worst cases, for the 10 features. The 10 features are radius, texture, perimeter, area, smoothness, compactness, concavity, concavity points, symmetry and fractal dimension. Such feature values are processed and computed based on the image of the tissue cells, which were obtained using Fine needle aspirations (FNAs) (Street 1993). Figure 1 shows the cells that were obtained through fine needle aspirate, and the boundaries of the cells were outlined by the contour model called "snake" (Street 1993). While some variables including radius, area and perimeter are relatively intuitive, other variables requires computation based on computer vision diagnostic system. Smoothness and compactness, for example, are calculated by local variation of the radius length and perimeter square divided by area. Table 1 shows how each nuclei feature was obtained specifically.

Variables	Explanation
id	Patient ID
diagnosis	"M" as malignant and "B" as benign
radius	mean of distances from center to points on the perimeter
texture	standard deviation of gray-scale values
perimeter	cell perimeter
area	cell area
smoothness	local variation in radius lengths
compactness	perimeter^2 / area - 1.0
concavity	severity of concave portions of the contour
concave points	number of concave portions of the contour
symmetry	length difference between lines perpendicular to major axis
fractal dimension	"coastline approximation" - 1

Table 1. Extracted nuclear features and the computational processes.

There are 569 tumor cell observations, 212 among which were marked as malignant and the other 357 observations were marked as benign. There is no missing value. Previous studies have focused on how logistic regression and inductive machine learning such as decision trees, Cart and C4.5 could achieve the classification of the diagnosis (Mangasarian 1995). The studies used all features to fit models to obtain high cross validation accuracy. The sensitivity and specificity of the models, however, was not deeply investigated and discussed in previous studies. Features selections in previous studies also tended to be similar in collecting all cell features.

Method

This project will conduct different feature selections and apply machine learning algorithms to dataset. Sensitivity and specificity for such machine learning methods will be analyzed along the process of hyper parameter tuning. Generally, the studies aim to increase the true positive rate at a reasonable cost of true negative rate. Appropriate machine learning methods and corresponding selected features will also be determined along this investigation. In specific, several basic machine learning algorithms including logistic regression, K-Nearest neighbors, random forest and support vector machine are considered to be utilized to the scaled and reduced dataset. The dataset is expected to be reduced by selecting particular feature measurements such as the feature means/worst/variance groups and conduct dimension reduction using PCA to decrease feature numbers.

Data Preprocessing

	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal dimension_mean
count	569	569	569	569	569	569	569	569	569	569
mean	14.12729174	19.28964851	91.96903339	654.8891037	0.096360281	0.104340984	0.088799316	0.048919146	0.181161863	0.06279761
std	3.524048826	4.301035768	24.29898104	351.9141292	0.014064128	0.052812758	0.079719809	0.038802845	0.027414281	0.007060363
min	6.981	9.71	43.79	143.5	0.05263	0.01938	0	0	0.106	0.04996
25%	11.7	16.17	75.17	420.3	0.08637	0.06492	0.02956	0.02031	0.1619	0.0577
0.5	13.37	18.84	86.24	551.1	0.09587	0.09263	0.06154	0.0335	0.1792	0.06154
75%	15.78	21.8	104.1	782.7	0.1053	0.1304	0.1307	0.074	0.1957	0.06612
max	28.11	39.28	188.5	2501	0.1634	0.3454	0.4268	0.2012	0.304	0.09744

Table 2. Descriptive statistics of feature means.

Table 2 shows the descriptive statistics of the dataset. Since the mean, standard error and worst cases measurements were computed for all ten features, there are several reasonable ways to calculate and select feature values for the models. Literature reviews show that most of the previous studies took the mean of the features of the nuclei. To increase model sensitivity, however, this study might also fit models using worst feature values cases among the cells. This feature selection process will be conducted after determining one or two most appropriate models.

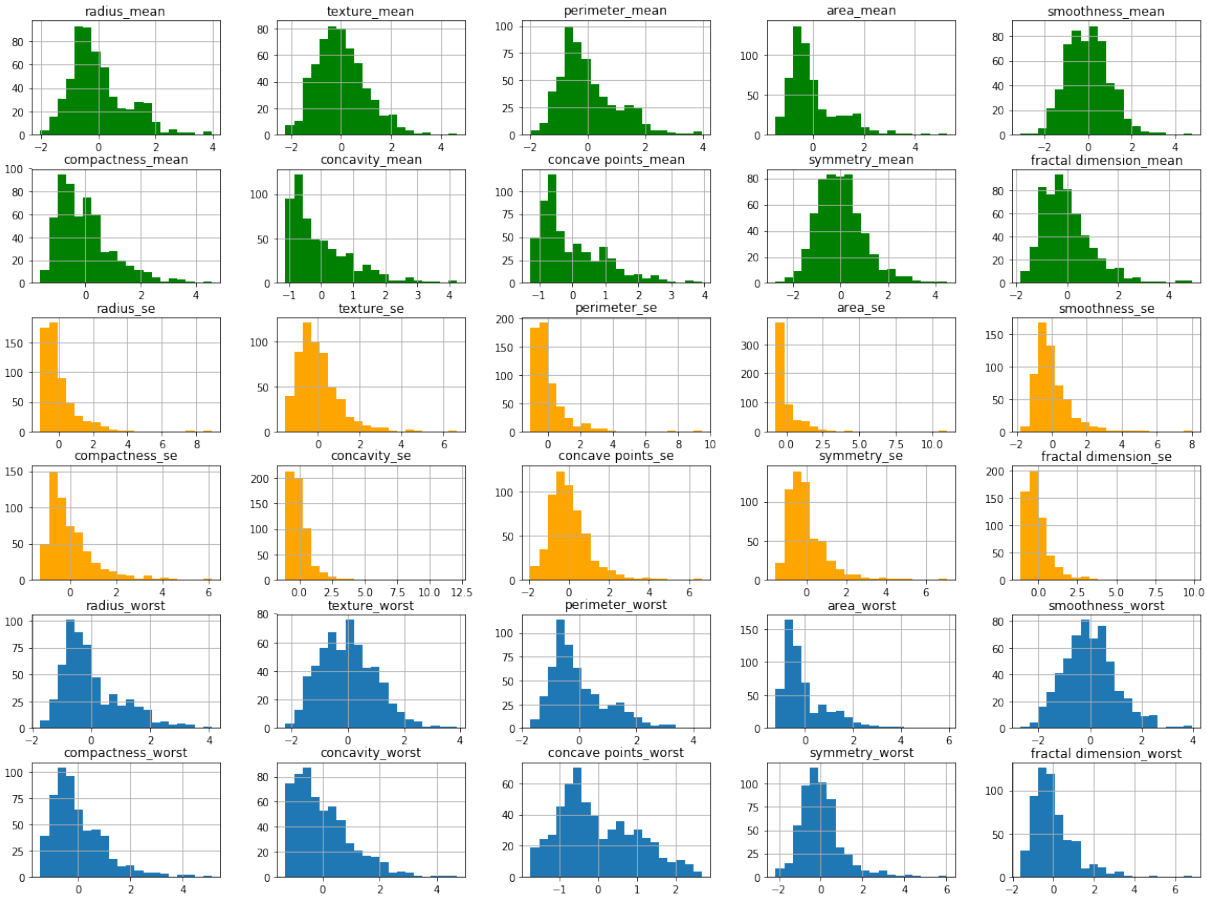


Figure2: The histogram of each feature measurements. The top two rows show the mean of feature measurements; the top two rows show the standard error of feature measurements; the top two rows show the worst cases of feature measurements,

The variables are all numerical. As shown in figure 2, it can be observed from the histogram of each feature that there is no apparent outlier for the feature means. Therefore, both Min-max and standard scaling shall be applied to such numerical variables. The data preprocessing process temporarily used standard scaling for the independent variables. The target values whether the tumor is malignant or benign were then labeled as 0 for benign and 1 for malignant. Figure 3 shows the dataset with standardly scaled numerical independent variables and labeled target variable.

```
# Standard Data Scaling
scaler = StandardScaler(copy=True, with_mean=True, with_std=True)
wdbc[name_list[2:]] = scaler.fit_transform(wdbc[name_list[2:]])

# Label target variable
leb = LabelEncoder()
wdbc[["diagnosis"]] = leb.fit_transform(wdbc["diagnosis"])
wdbc.head()
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	conca
0	842302	1	1.097064	-2.073335	1.269934	0.984375	1.568466	3.283515	
1	842517	1	1.829821	-0.353632	1.685955	1.908708	-0.826962	-0.487072	
2	84300903	1	1.579888	0.456187	1.566503	1.558884	0.942210	1.052926	
3	84348301	1	-0.768909	0.253732	-0.592687	-0.764464	3.283553	3.402909	
4	84358402	1	1.750297	-1.151816	1.776573	1.826229	0.280372	0.539340	

5 rows × 32 columns

Figure2. Breast cancer Wisconsin dataset with scaled numerical independent variables for the 30 feature measurements and labeled target variable for "diagnosis".

References:

W.N. Street, W.H. Wolberg and O.L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, volume 1905, pages 861-870, San Jose, CA, 1993.

O.L. Mangasarian, W.N. Street and W.H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), pages 570-577, July-August 1995.