Wisconsin Diagnostic Breast Cancer Project

Ziyan Lin, Xiao Sun, Han Wang, Mengnan Zhao, Siyao Zhu

Abstract

Breast cancer was the form of cancer most often described in ancient documents because of its visibility. Before the 20th century, breast cancer was feared and discussed in hushed tones, as if it were shameful. As little could be safely done with primitive surgical techniques, women tended to suffer silently rather than seeking care. When surgery advanced, and long-term survival rates improved, women began raising awareness of the disease and the possibility of successful treatment. The first noticeable symptom of breast cancer is typically a lump that feels different from the rest of the breast tissue. More than 80% of breast cancer cases are discovered when the woman feels a lump. Of course, if possible, find out if a person has breast cancer or not as early as possible will be the best for the treatment successfully rate. But it is also important to determine a breast cancer is malignant or benign in an early stage before people can feel it. The project using classification methods to do the breast cancer prediction. Breast Cancer Wisconsin (Diagnostic) Data Set is used to do the analysis. Five classification methods are used and their accuracy is compared. Results shows the best classification method are SVM and Logistic Regression.

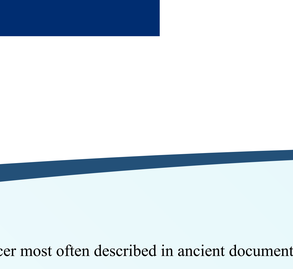
Introduction

If people have been diagnosed with a breast cancer, the first step doctors will take is to find out whether it is malignant or benign, as this will affect the treatment plan. Therefore, it is significant for analyst to improve prediction accuracy. The objective of our project is to predict whether a breast cancer is malignant or benign based on ten real-valued features: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each feature, resulting in 30 features.

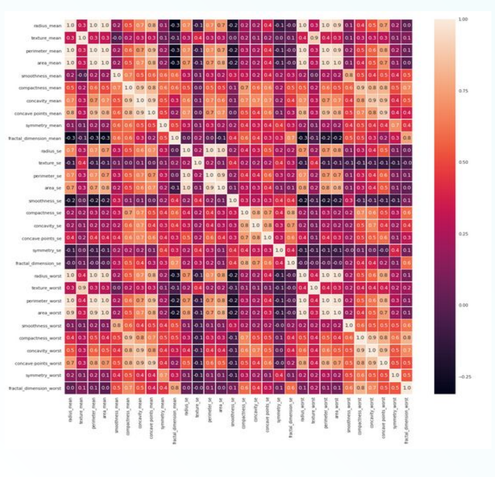
Our dataset includes 569 observations. For each observation, it has one response variable with “malignant” or “benign”, one ID number, and 30 predictor variables. Hence, the size is 569 ✕ 32.

Data Visualization

The two classes are **biased** with 357 (62.7%) benign and 212 (37.3%) malignant.



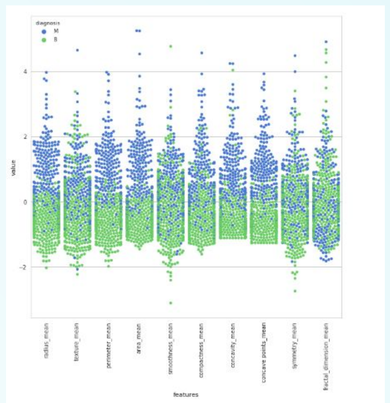
Next, the data is normalized by extracting the mean and then divided by the standard deviation.

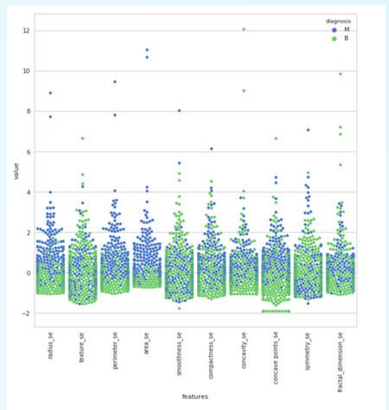


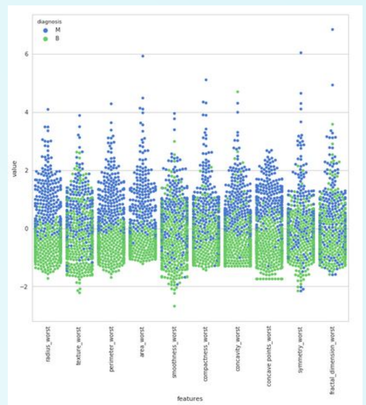
Feature Selection

Since there exists correlations among features, the first step is variable selection. Based on the heatmap of all 30 features, it is obvious that there are some larger number in the map except for diagonal line. After removing all high-correlated features, we find there are only 15 features left. Then we check again with heatmap and find that no high correlation any more.

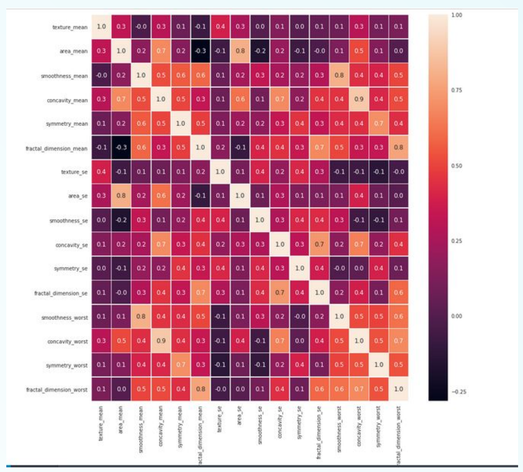
Visualization of the distribution of 30 features to detect most discriminative features.



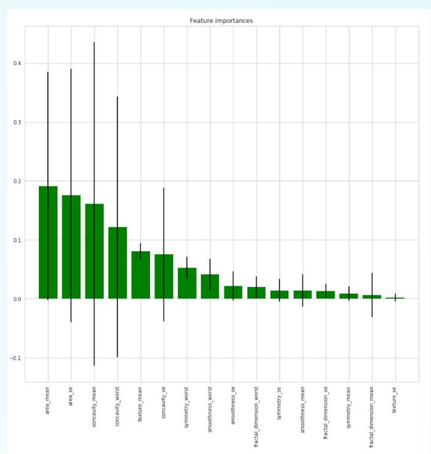




For most of the features, medians of the Malignant and Benign are well separated so it can be good for classification. However, for some features, medians of the Malignant and Benign are close so it does not gives good information for classification.

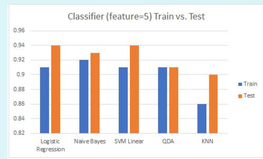


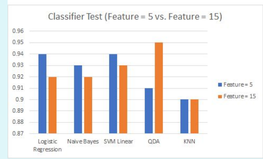
Feature selection by excluding features with high correlation.



Feature importance in random forest classification. Therefore, we can use the most informative 5 features to do prediction..

Results





Classifier (feature=5, train)	Accuracy mean (cv=10)
Logistic Regression	0.91
Naive Bayes	0.92
SVM Linear	0.91
QDA	0.91
KNN	0.86

Classifier (feature=15, train)	Accuracy mean (cv=10)
Logistic Regression	0.93
Naive Bayes	0.92
SVM Linear	0.94
QDA	0.94
KNN	0.86

Classifier(feature=5,test)	Accuracy mean (cv=10)
Logistic Regression	0.94
Naive Bayes	0.93
SVM Linear	0.94
QDA	0.91
KNN	0.90

Classifier(feature=15, test)	Accuracy mean (cv=10)
Logistic Regression	0.92
Naive Bayes	0.92
SVM Linear	0.93
QDA	0.95
KNN	0.90

When using five features, Logistic regression and SVM linear method perform well. And generally, the accuracy of testing data is better.
While using fifteen features, the accuracy of training data tends to be larger than that of testing data, which indicates the probability of overfitting problem. Comparing the accuracy of using five and fifteen features based on testing data, five features seem to be enough. And SVM linear and Logistic regression method can classify the data well. QDA has better performance if we use fifteen features. However, QDA has weaker interpretability as logistic regression.

Conclusion

(1) For breast cancer classification, area_mean, area_se, texture_mean, concavity_worst and concavity_mean are the five most important features.

(2) If we use more features, overfitting problems may occur. In reality, five features are enough to predict breast cancer diagnosis.

Acknowledgements

550.436 Data Mining Lecture Notes by Professor Tamas Budavari

Feature Selection and Data Visualization: https://www.kaggle.com/kanncaa1/feature-selection-and-data-visualization

Wisconsin Diagnostic Breast Cancer Dataset: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data