Predicting Diagnosis of Breast Cancer Tumor

Jianghui Lin jl5172, Zixu Wang zw2541, Jack Yan xy2395

Introduction

This dataset includes different features which were computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. The FNA describes the characteristics of the cell nuclei present in the image. Ten real-valued features are computed for each cell nucleus. The features include radius, texture, perimeter, area, compactness, concavity, concave points, symmetry, and fractal dimension. In this project, we aim to use the lab results to determine which features contribute to a better diagnosis result of breast mass (M = malignant, B = benign). In the dataset, we have 357 benign observations and 212 malignant observations, so the two classes are balanced.

Dataset Partitioning

The original dataset contains 569 observations, among which 456 were randomly partitioned into the training set, and 113 into the test set. The training set was used to conduct model selection based on 10-fold cross-validation, and the test set was used to verify the model selection result.

Exploratory Data Analysis

1. K Means Clustering (Unsupervised learning)

In 2-means clustering **(Figure 1)**, we seek to partition the observations into a pre-specified number of clusters, in this case, we have specified the number of clusters as 2 using the `fviz_nbclust` function to determine the optimized the number of clusters. The number of optimized clusters exactly equals to the number of our response categories 'B' and 'M'.

2. Correlation plots

Although there are some strong correlations among several covariates, we considered them as intuitive ones **(Figure 2)**. For example, the correlations among perimeters, radius, and area are all related to the size of the tumor. We decided to keep all covariates at this point for further analysis.

3. One-to-one relation between classes and covariates

According to the feature plot **(Figure 3)** which shows the one-to-one relation between classes and covariate. The 'M' (malignant tutor) class generally has larger radius mean, texture mean, perimeter mean, area mean, compactness mean, smoothness mean and concavity mean compare to those of 'B' (benign) tumors. We will circle back after the thorough analysis to check if the conclusions we obtain are identical.

Models

All the 30 covariates were included in all the models. The 30 predictors are the means, standard errors and worst values of the 10 characteristics measured on the cell nuclei. Ordinary and regularized Logistic regression, Linear and quadratic discriminant analysis, Naive Bayes, KNN, classification trees, random forests, and ensemble methods were used to fit the data. The area under the ROC curve (AUC) were calculated from repeated cross-validation and used to compare models. For models that require tuning, repeated cross-validation was used to decide the best hyperparameter(s) that corresponds to the largest AUC.

1. Logistic regression and regularized logistic regression (glmnet)

Both logistic regression and regularized logistic regression models (glmnet) were built to fit the 30 predictors on the response. Data were centered and scaled before fitting into the glmnet model. Cross-validation showed that the logistic regression has AUC 0.9544, sensitivity 0.9821 and specificity 0.8856. The best parameter of glmnet is alpha 0.2 and lambda 0.014, with AUC 0.9961, which is better than the logistic regression model.

2. LDA

The assumption of LDA is that all the variables follow a multivariate normal distribution. Therefore, each covariate has its own mean and shares a common variance-covariance matrix. All 30 predictors were included in this model. By fitting the model with the cross-validation method, this LDA model has AUC 0.9917, sensitivity 0.8958 and specificity 0.9930.

3. QDA

QDA assumes that all the variables are following a Gaussian distribution. Therefore, different from LDA, each covariate has its specific mean and different variance-covariance matrix. All 30 predictors were included in this QDA model and the AUC is 0.9910, sensitivity is 0.9718, specificity is 0.9425.

4. Naive Bayes

Naive Bayes assumes that the features are conditionally independent given the class instead of modeling their full conditional distribution given the class, and it is an approximation to the Bayes classifier. All 30 predictors were included in the model and the best AUC is 0.9873.

5. KNN

All 30 predictors were included in this KNN model, and the best tuning parameter is 33. This model has AUC equals to 0.9925, sensitivity equals to 0.8784 and specificity equals to 0.9894.

6. Support Vector Machine (linear and radial kernel)

7. Classification Tree and random forests

A classification tree was built on the 30 predictors. Cp was tuned using cross-validation. The best cp turned out to be 0.0062, corresponding to 5 terminal nodes and AUC = 0.9412. In the random forest model, both interaction depth (mtry) and minimal node size were tuned by cross-validation. The best model random forest model had interaction depth 1 and minimal node size 1, with AUC 0.9931, which is much better than that of a single tree.

8. Boosting (AdaBoosting and binomial loss function)

Binomial loss boosting and AdaBoosting were tuned on 3 hyperparameters: number of trees, interaction depth, and shrinkage. Minimal node size was set to be 1. The best binomial loss boosting model and the best AdaBoosting model had AUC 0.9926 and 0.9930, respectively. Their AUC are comparable.

Model Selection

Model selection was based on repeated cross-validation. The glmnet model had the largest AUC, so we decided to use it as our final model.

APPENDIX

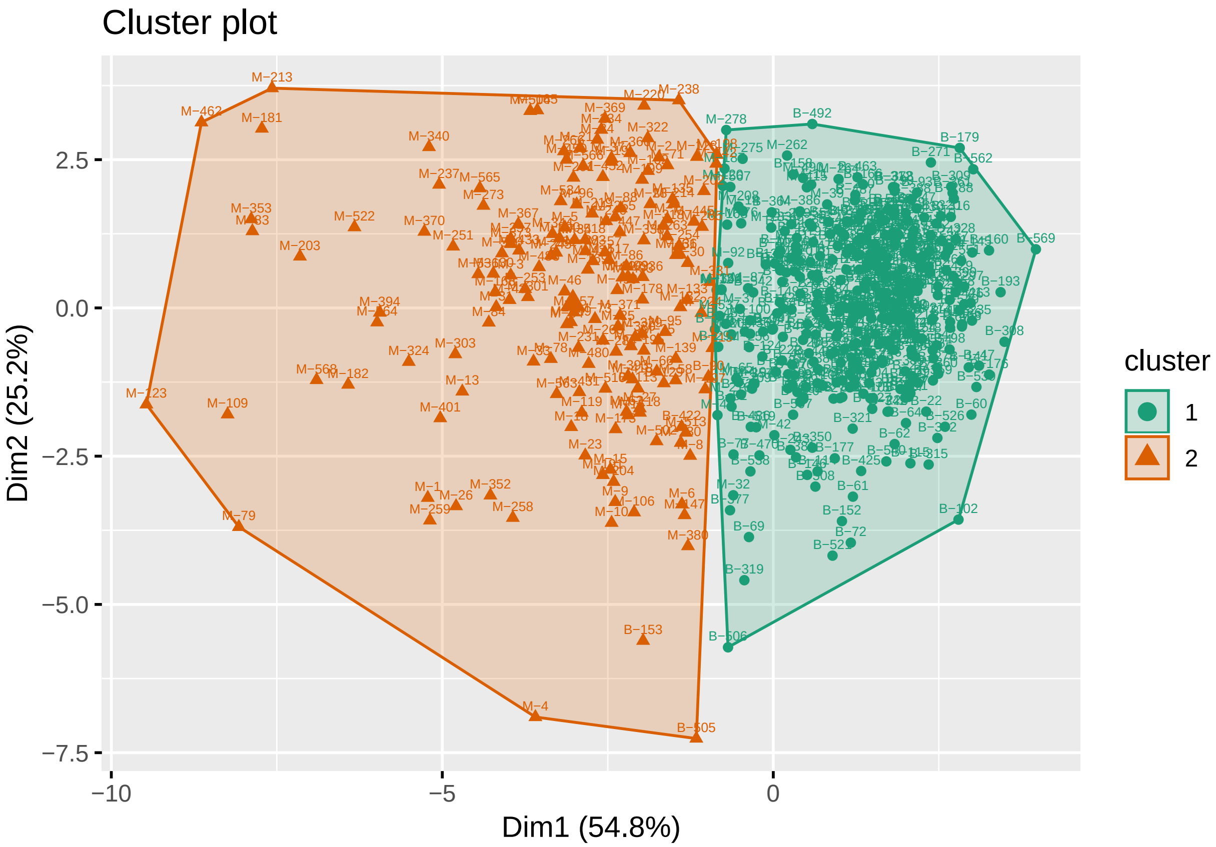

Figure 1 2-means clustering on the first two principle components

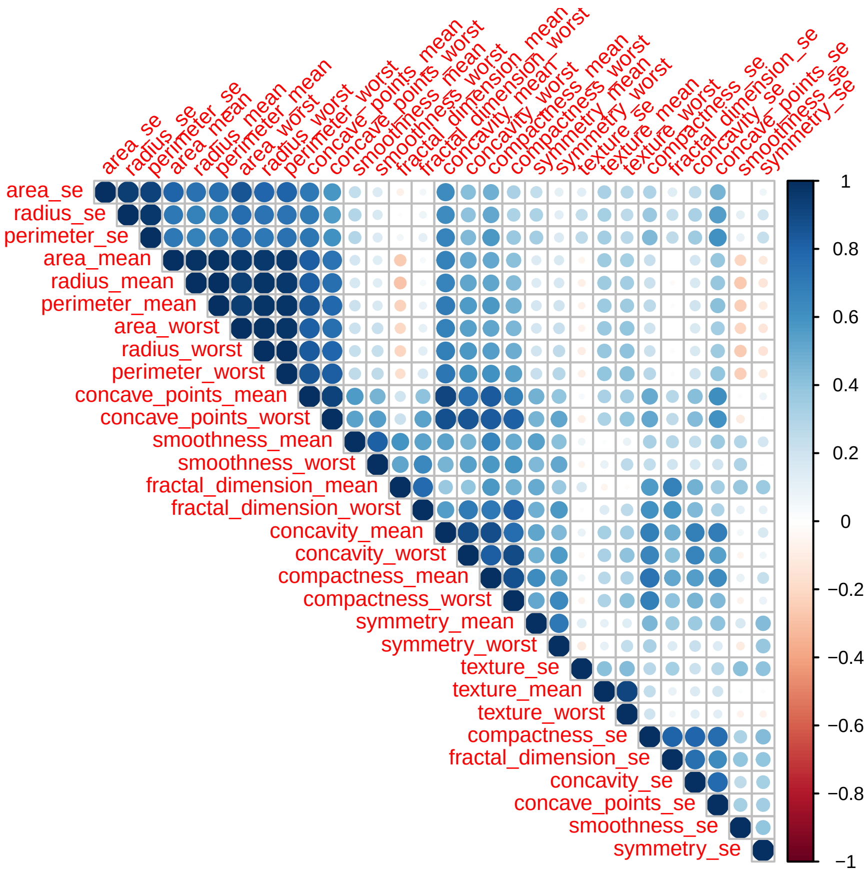

Figure 2 Correlation plots

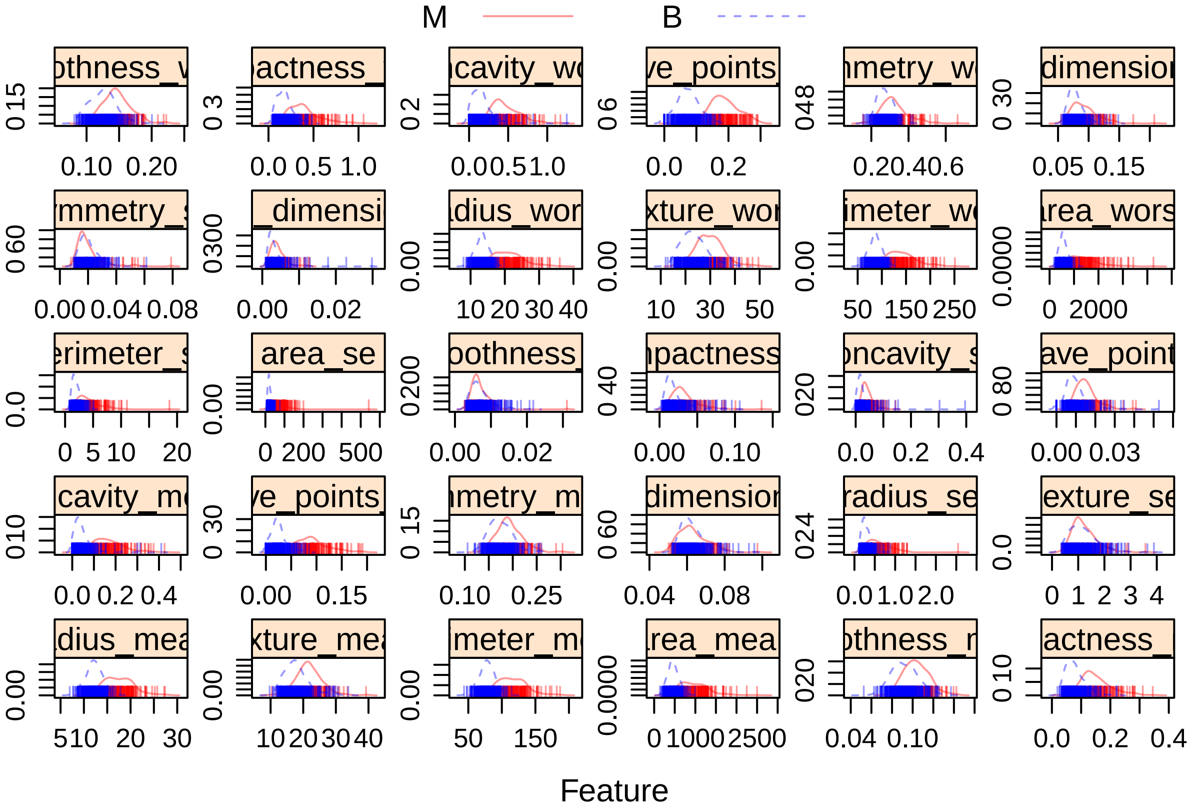

Figure 3 One-to-one relation between classes and covariates

```
## Warning: Setting row names on a tibble is deprecated.

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: Setting row names on a tibble is deprecated.
```

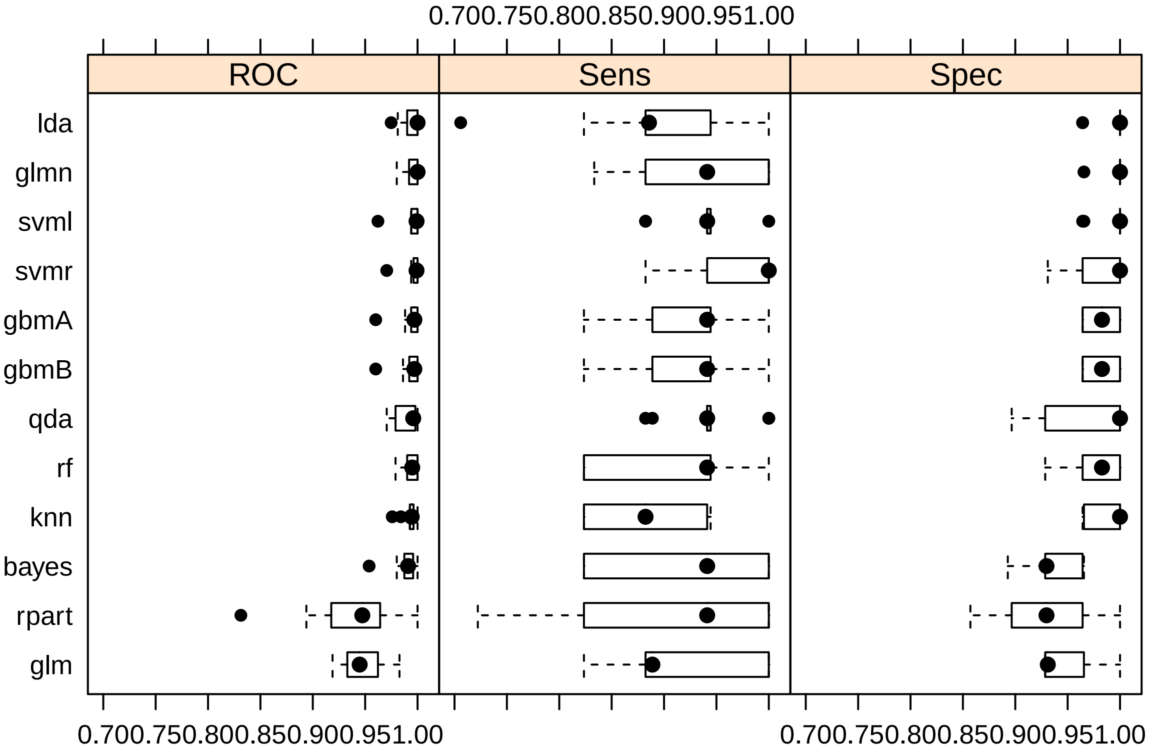

Figure Model comparision on AUC, sensitivity and specificity

Table 1: Mean of AUC, sensitivity and specificity of different models

model	AUC	Sensitivity	Specificity
svmr	0.9956587	0.9650327	0.9823892
glmn	0.9956514	0.9313725	0.9965517
svml	0.9944030	0.9362745	0.9929803
lda	0.9938025	0.8954248	0.9928571
rf	0.9932668	0.9071895	0.9788177
gbmA	0.9929940	0.9192810	0.9823892
gbmB	0.9926016	0.9192810	0.9823892
knn	0.9925138	0.8784314	0.9894089
qda	0.9898975	0.9424837	0.9717980
bayes	0.9880409	0.9186275	0.9397783
glm	0.9488751	0.9189542	0.9506158
rpart	0.9362977	0.9032680	0.9328818