Predicting Diagnosis of Breast Cancer Tumor

Jianghui Lin jl5172, Zixu Wang zw2541, Jack Yan xy2395

Introduction

Breast cancer is the second most common cancer diagnosed in women in the United States. The signs and symptoms of breast cancer may include a breast lump or thickening that feels different from the surrounding tissue, change in the size, shape or appearance of a breast, etc. However, most of the breast mass is benign, which means they are not the indicator of breast cancer.

This dataset includes different features that were computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. FNA is a new diagnostic procedure used to investigate lumps or masses. In this technique, a thin, hollow needle is inserted into the mass for a sampling of cells that, after being stained, will be examined under a microscope.

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

K-means clustering was used to partition the observations into 2 clusters. In this dataset, the optimal number of clusters is 2, determined by average silhouette method using the the `fviz_nbclust` function. Interestingly, the number of optimized clusters exactly equals the number of response categories 'B' and 'M'. **Figure 1** shows the distribution of clusters on the first 2 principle components (PCs). Dots are labelled by their response categories, although response information was not used in clustering. The two clusters can be well separated by the first 2 PCs. Benign cases are mainly in cluster 1, and Malignant cases in cluster 2 **(Figure S4)**.

2. Correlation plots

The correlation plot **(Figure 2)** shows that strong correlation exist among a subset of covariates. Intuitively, the high correlation makes sense. For example, the correlation among perimeter, radius, and area are all related to the size of the tumor. However, we did not know which covariates are better, so we decided to keep all covariates at this point, and let the models decide which variables are more important.

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

According to the feature plots **(Figures S1-S3)** that show the one-to-one relation between the response and covariates, 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. The two classes can be well separated on many individual covariates, such as radius_mean, area_se and concave_points_mean.

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, support vector machines, 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 optimal hyperparameter(s) that corresponds to the largest AUC. Plots of CV AUC against the grid of hyperparameters are shown in **Figures S5-S13**.

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 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, the optimal 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 0.9925, sensitivity 0.8784 and specificity 0.9894.

6. Support Vector Machine (linear and radial kernel)

Support vector machines with both linear kernel and radial kernel were fit on the data. The optimal linear kernel has AUC 0.994 and the optimal radial kernel has AUC 0.996. The radial kernel is slightly better.

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 Comparison and selection

The mean cross-validation AUC, sensitivity and specificity are summarized in **Table 1**. SVM (radial kernal) has the highest sensitivity (0.965), while the glmnet model has the highest specificity (0.997). The AUC of all the models are higher than 0.93 **(Figure 3)**.

Model selection was based on cross-validation AUC. The support vector machine (radial kernal) and glmnet model have the largest cross-validation AUC (0.9957). Since the glmnet model can be easier interpreted, we decided to use it as our final model.

Test set performance was also assessed and AUC is reported **(Table 1)**. The ordering of test AUC is similar with cross-validation AUC, so the results of cross-validation is valid.

Final Model Interpretaion

The final model is a mixture of ridge and lasso (alpha = 0.2, lambda = 0.014). Non-zero coefficients of the final regularized logistic regression (glmnet) are summarized in **Table 2**. The 25 non-zero coefficients are arranged in descending order based on their absolute values. SE of radius, worst value of radius, worst value of texture, worst value of concave points, and worst value of perimeter are the top 5 predictors with largest absolute coefficients. Since all the predictors are centered and scaled before fitting the model, the coefficients are on the same scale as well. Therefore, their absolute values can reflect the importance of the predictors.

The train data ROC curve of the final model is shown in **Figure 4**. The AUC for the train data is 0.997, indicating a good fit. Not only does our final model have best prediction performance, but is also easier to interpret as compared to other complex models. The coefficients can be simply interpreted as the log odds ratio of malignant tumor for one-unit change in the corresponding scaled covariate, keeping other covariates constant.

Conclusion

We found that the regularized logistic regression including 25 covariates has the best performance as well as interpretability. This finding is understandable. There is a high correlation among the predictors, so regularization can mitigate the effect of correlation and generate a good prediction on this data. Also, our exploratory analyses **(Figures S1-S3)** found that the two response classes can be well separated, so the high AUC values of all the models are reasonable.

The predictors in the dataset are different from other datasets because for a single characteristic such as smoothness, its mean, standard error, and worst value are all used as predictors. The reason why mean and worst values are included is easy to understand, but it is not obvious why standard errors are included. In

fact, the most important predictor in our final model is the SE of the radius. Our hypothesis is that high standard error reflects the heterogeneity of the cells, so it is positively related to the probability of cancer.

The worst values of the characteristics are in fact among the most important predictors in the final model. This finding may have clinical significance and facilitate a more accurate diagnosis.

Figures and Tables

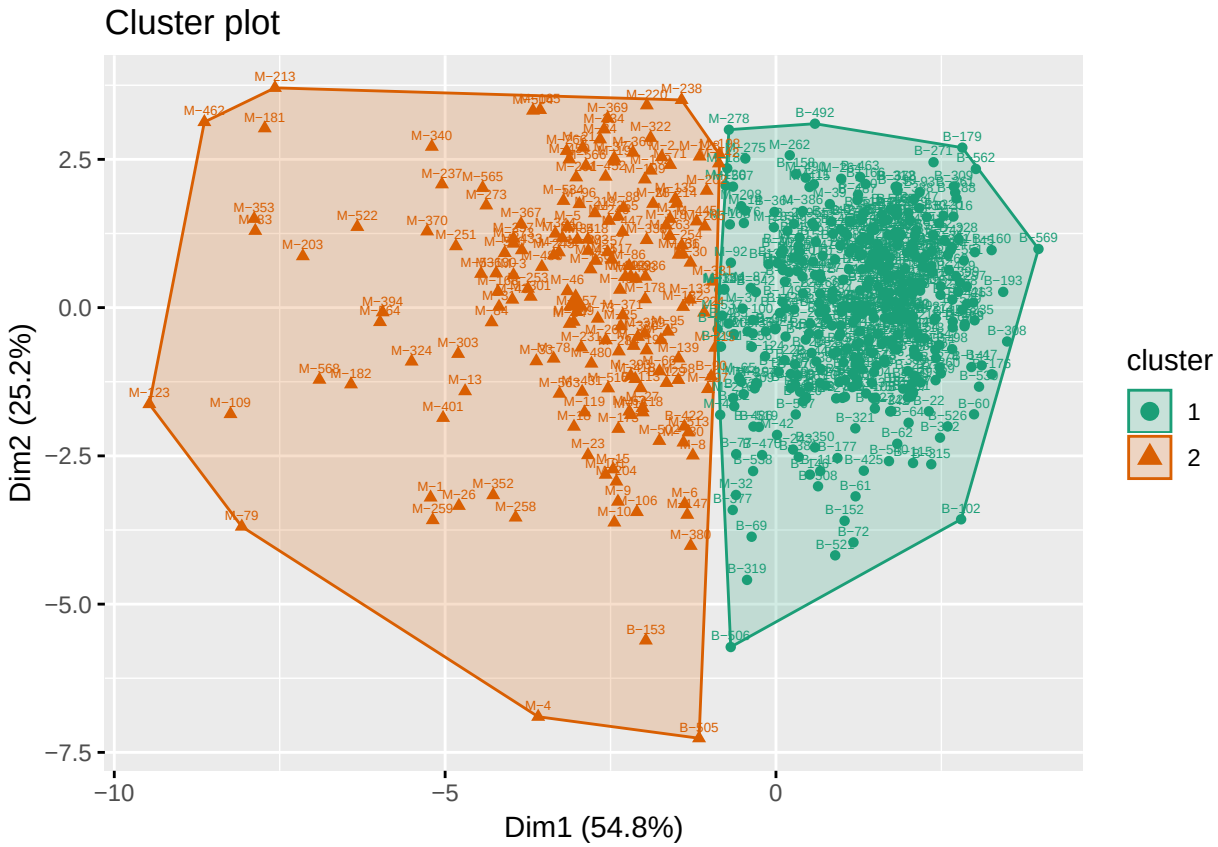

Figure 1 2-means clustering on the first two principle components

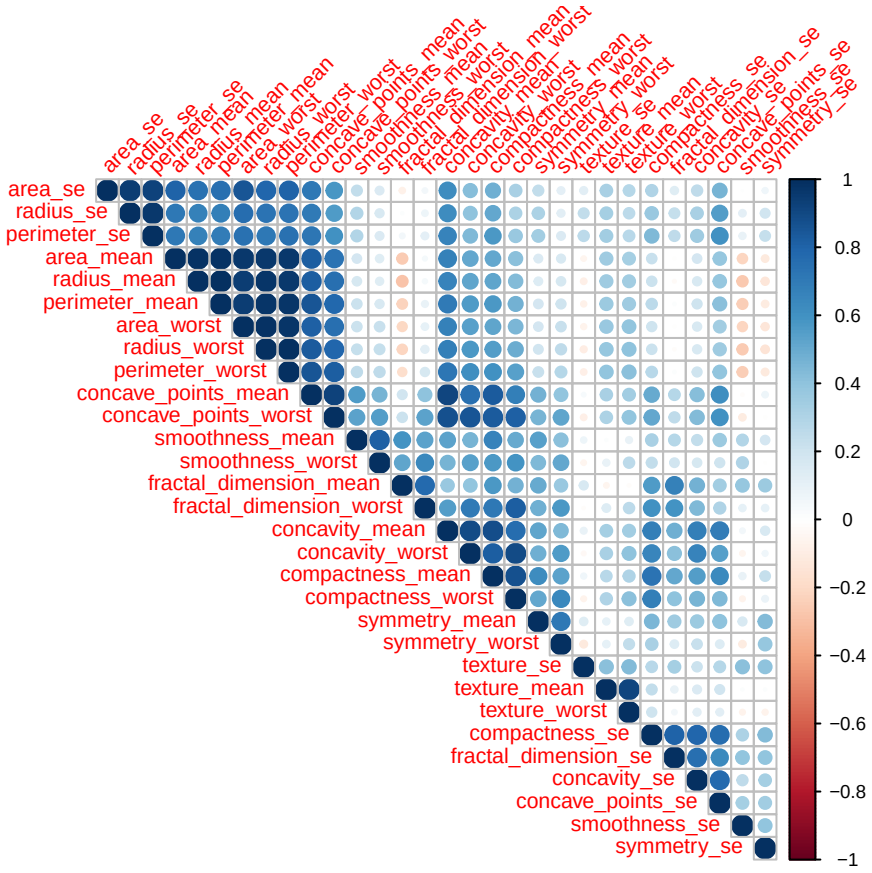

Figure 2 Correlation plot of predictors

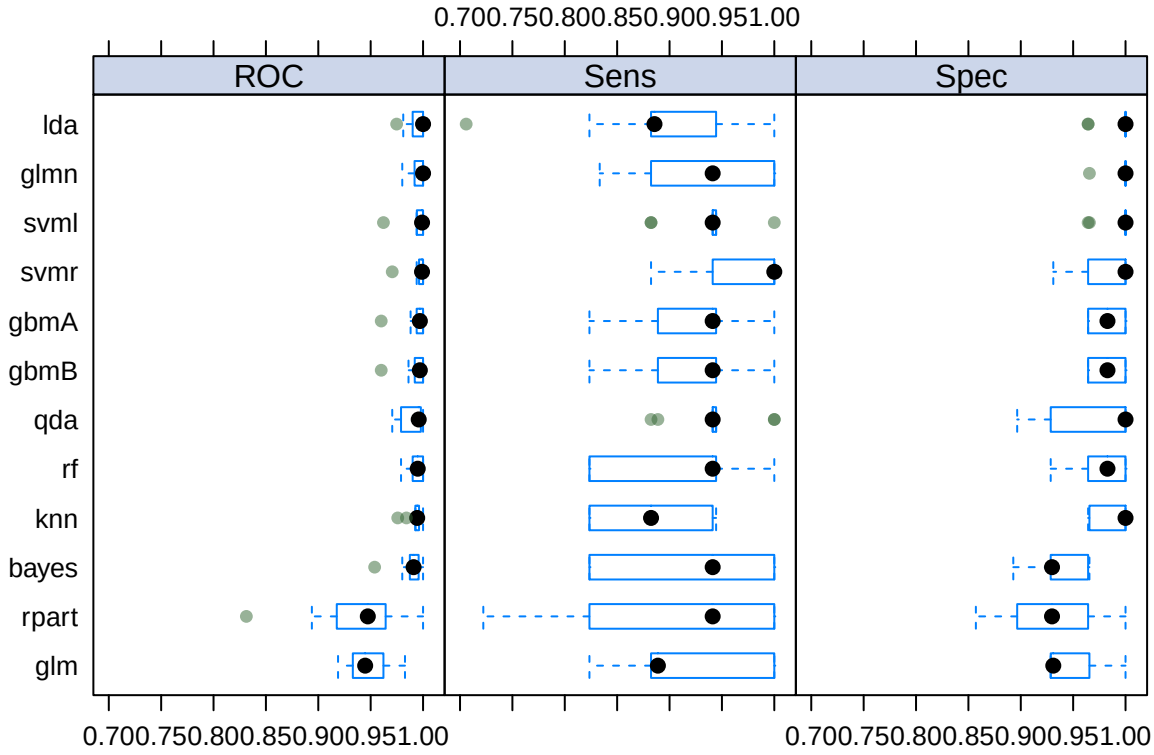

Figure 3 Model comparision on AUC, sensitivity and specificity

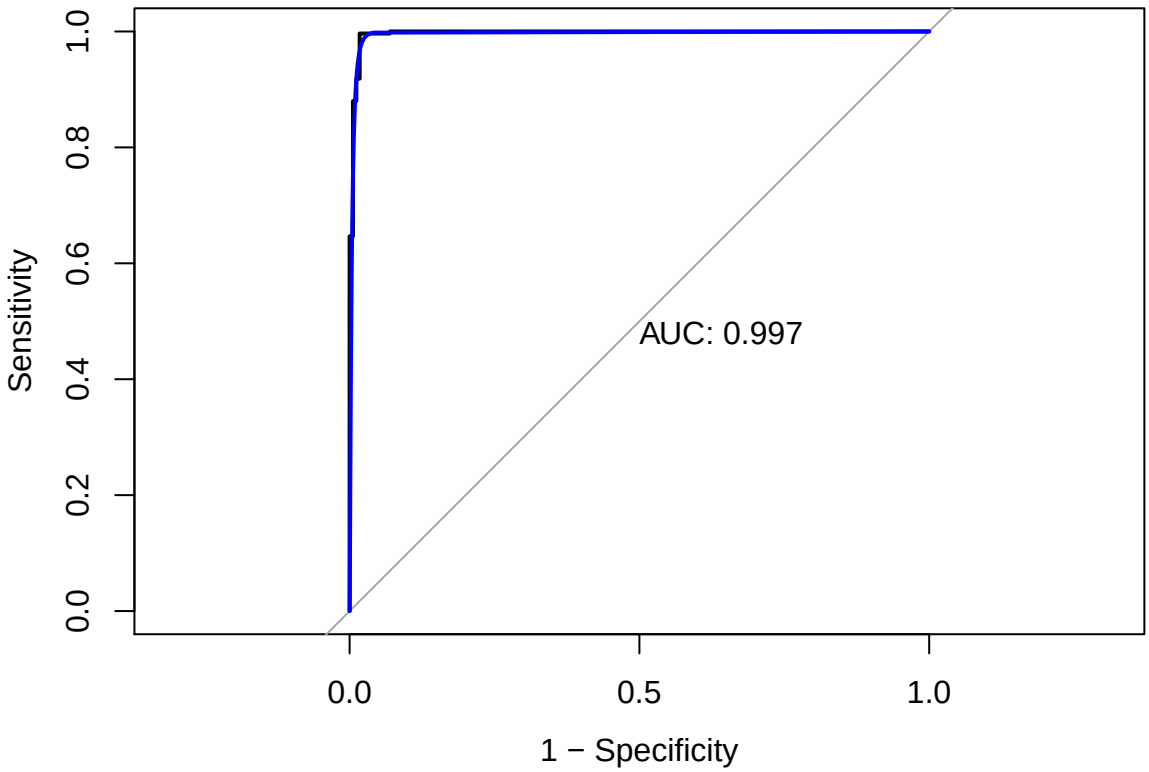

Figure 4 Train data ROC curve of the final model

Table 1: Cross-validation AUC, sensitivity, specificity, and test data AUC of different models

model	CV_AUC	CV_Sens	CV_Spec	Test_AUC
svmr	0.9957	0.9650	0.9824	0.9986
glmn	0.9957	0.9314	0.9966	0.9965
svml	0.9944	0.9363	0.9930	0.9990
lda	0.9938	0.8954	0.9929	0.9846
rf	0.9933	0.9072	0.9788	0.9941
gbmA	0.9930	0.9193	0.9824	0.9938
gbmB	0.9926	0.9193	0.9824	0.9896
knn	0.9925	0.8784	0.9894	0.9969
qda	0.9899	0.9425	0.9718	0.9896
bayes	0.9880	0.9186	0.9398	0.9771
glm	0.9489	0.9190	0.9506	0.9446
rpart	0.9363	0.9033	0.9329	0.9508

Table 2: Non-zero Coefficients of the Final Model

term	coef
radius_se	0.5796
radius_worst	0.5759
texture_worst	0.5550
concave_points_worst	0.5364
perimeter_worst	0.5242
smoothness_worst	0.4886
area_worst	0.4783
concave_points_mean	0.4758
concavity_worst	0.4594
texture_mean	0.4250
area_se	0.4041
perimeter_se	0.3707
radius_mean	0.3549
symmetry_worst	0.3437
perimeter_mean	0.3436
concavity_mean	0.3422
area_mean	0.3185
fractal_dimension_se	-0.2383
compactness_se	-0.1616
symmetry_se	-0.1496
fractal_dimension_mean	-0.1288
smoothness_mean	0.1223
compactness_worst	0.1195
fractal_dimension_worst	0.0672
smoothness_se	0.0088

APPENDIX

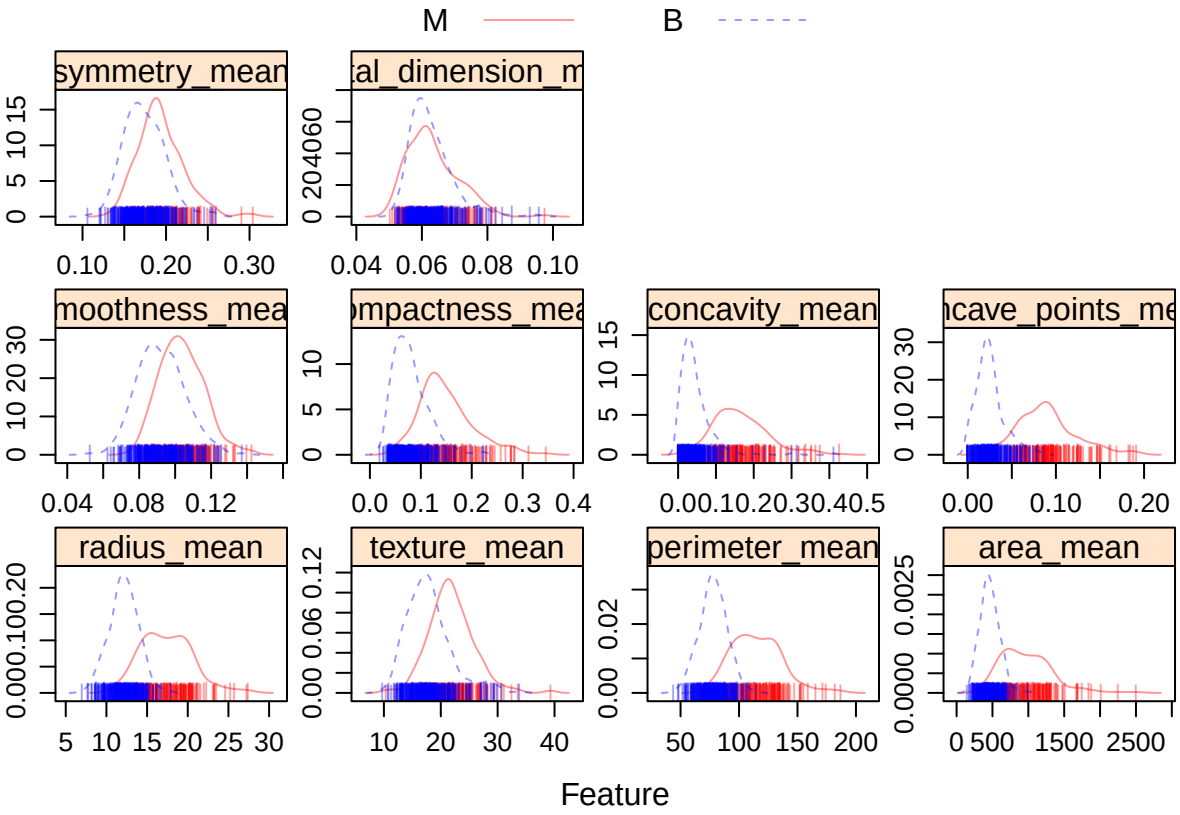

Figure S1 One-to-one relation between response classes and covariates (means)

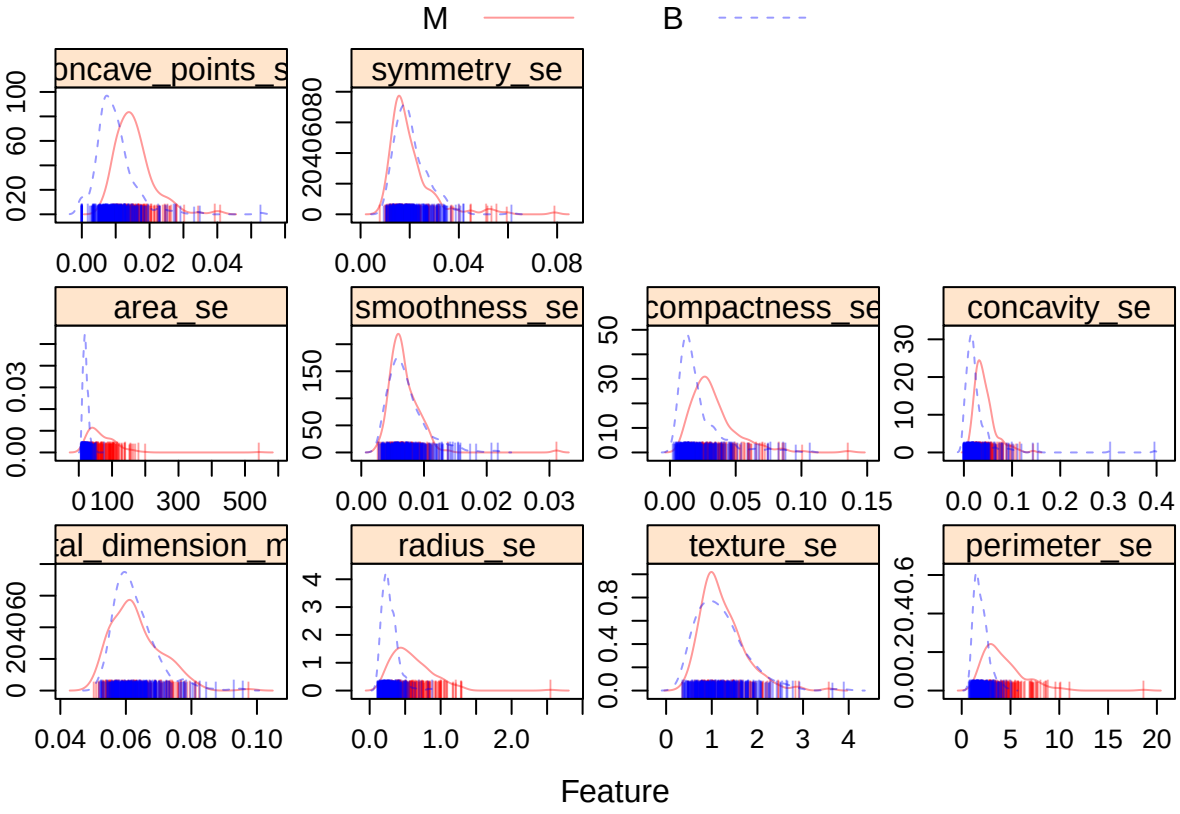

Figure S2 One-to-one relation between response classes and covariates (standard errors)

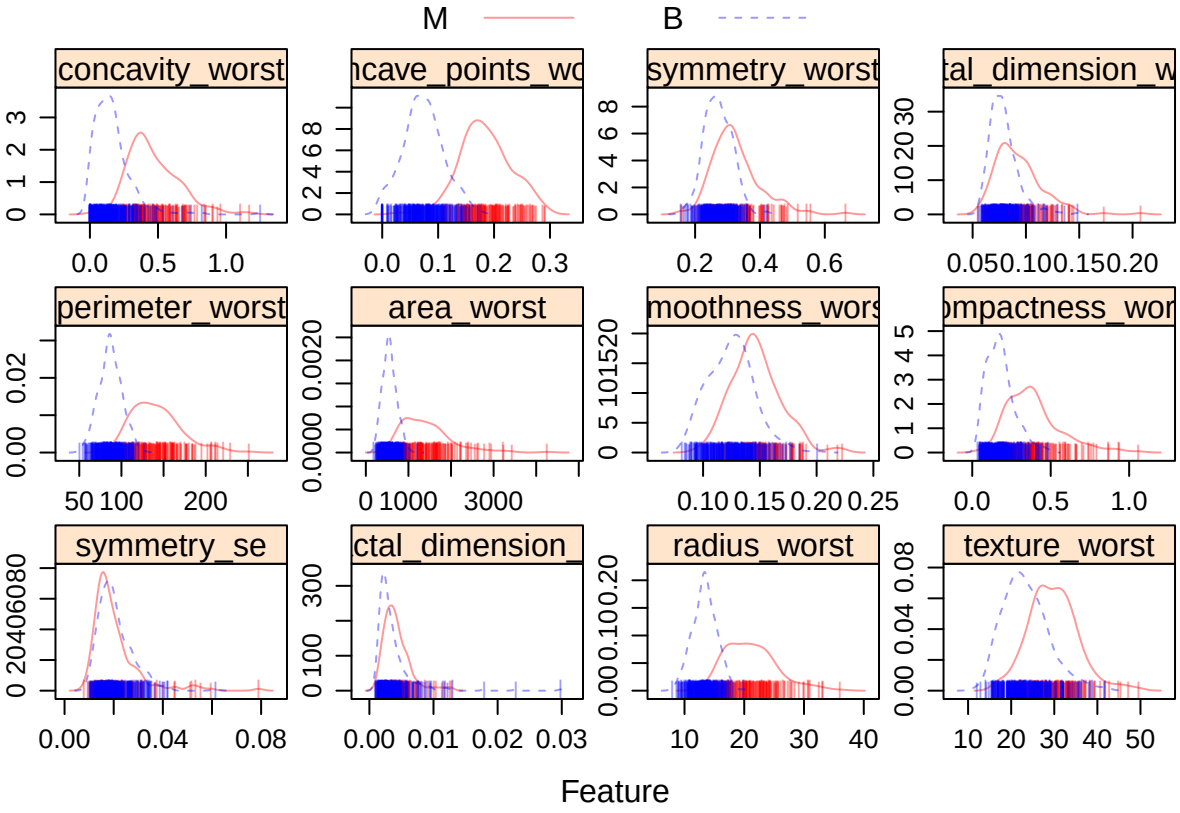

Figure S3 One-to-one relation between response classes and covariates (worst)

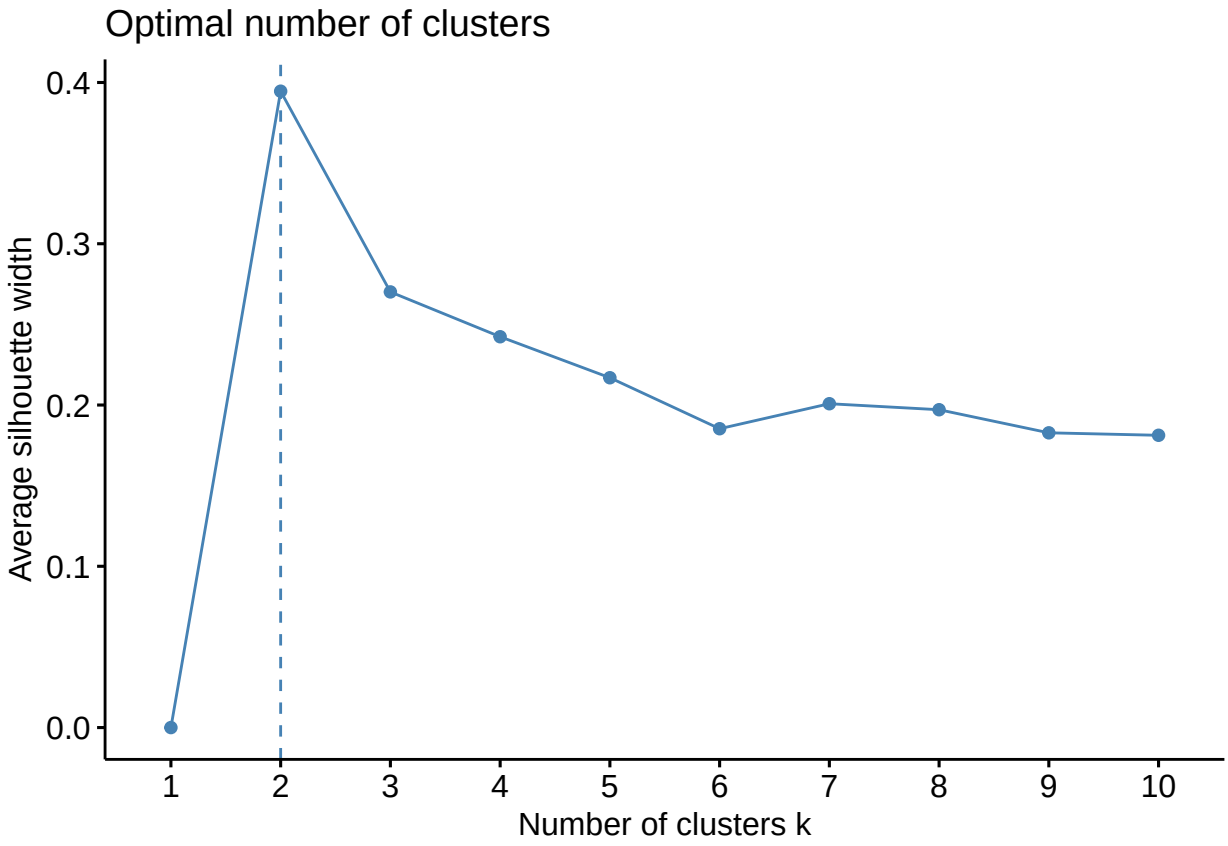

Figure S4 Optimal number of clusters determined by average silhouette method

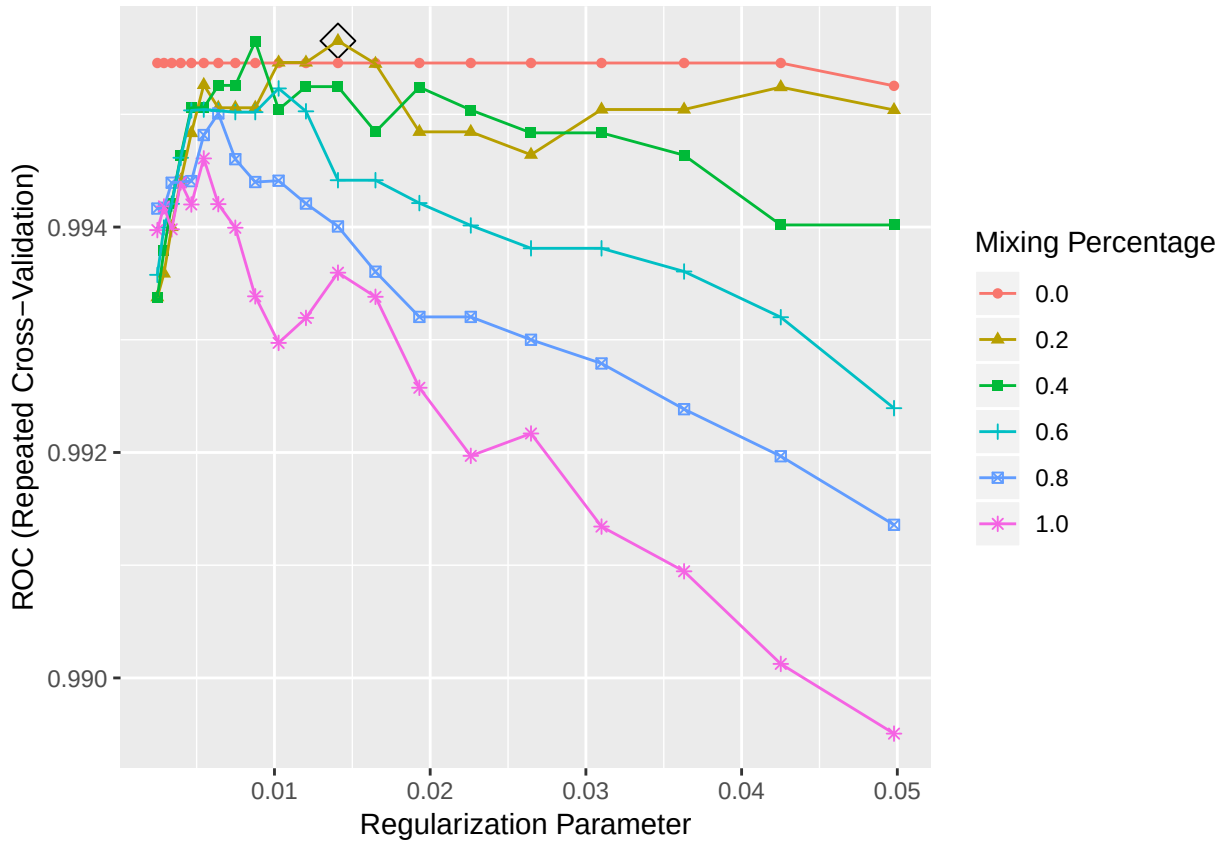

Figure S5 Regularized logistic regression model tuning parameter selection

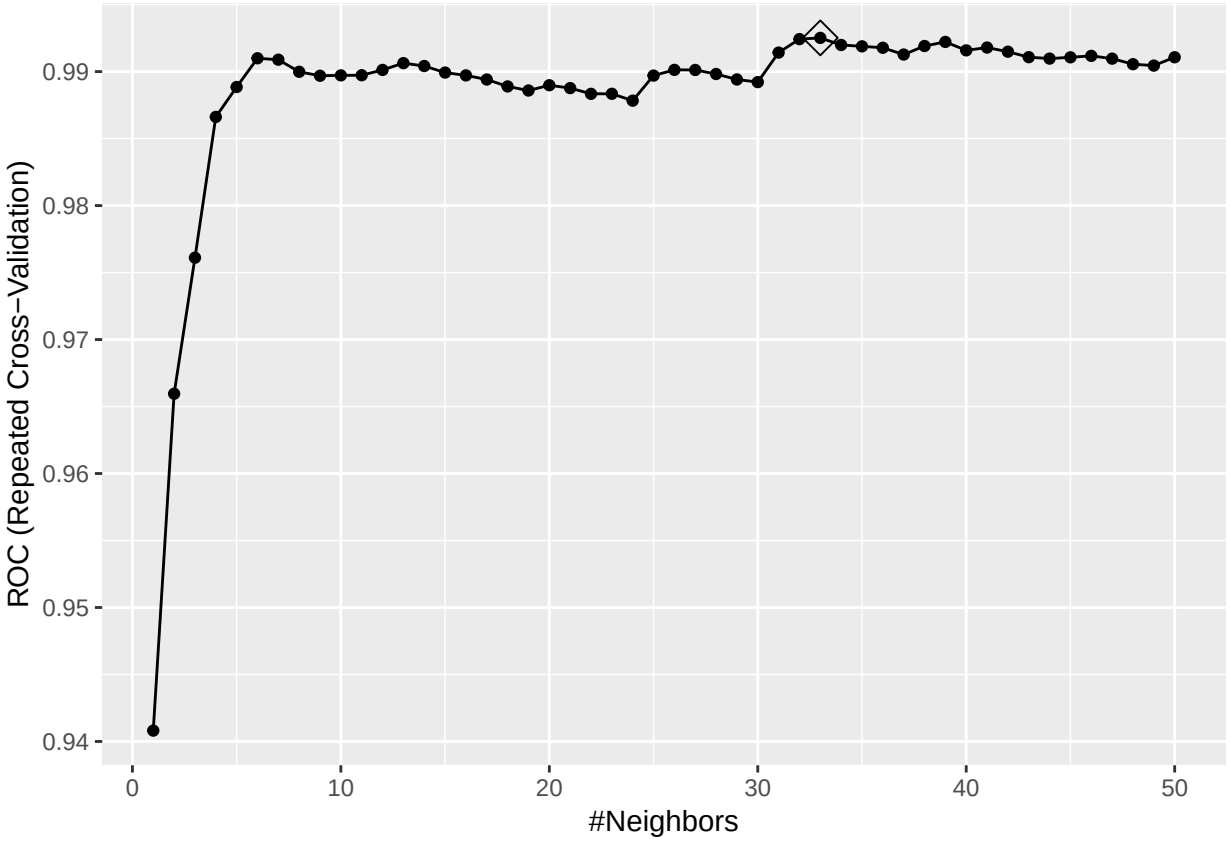

Figure S6 KNN model tuning parameter selection

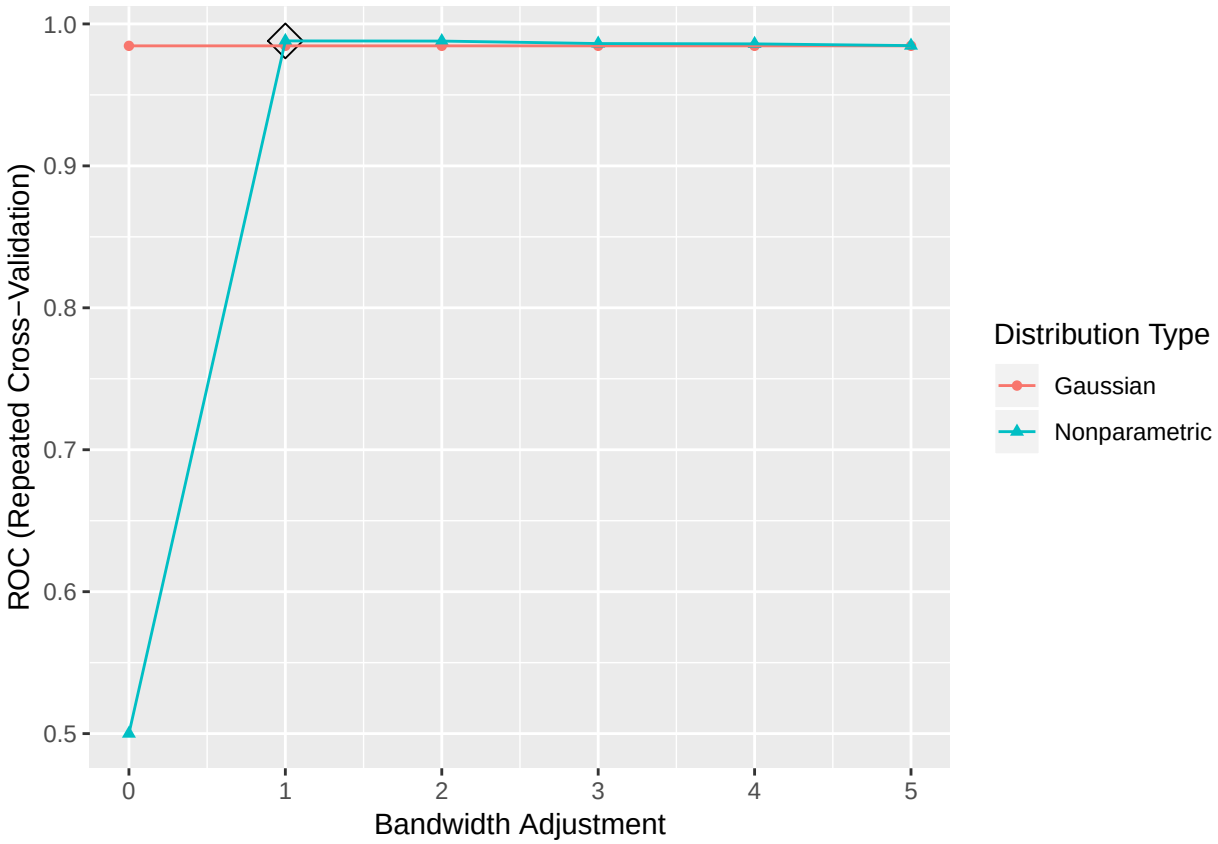

Figure S7 Naive Bayes model tuning parameter selection

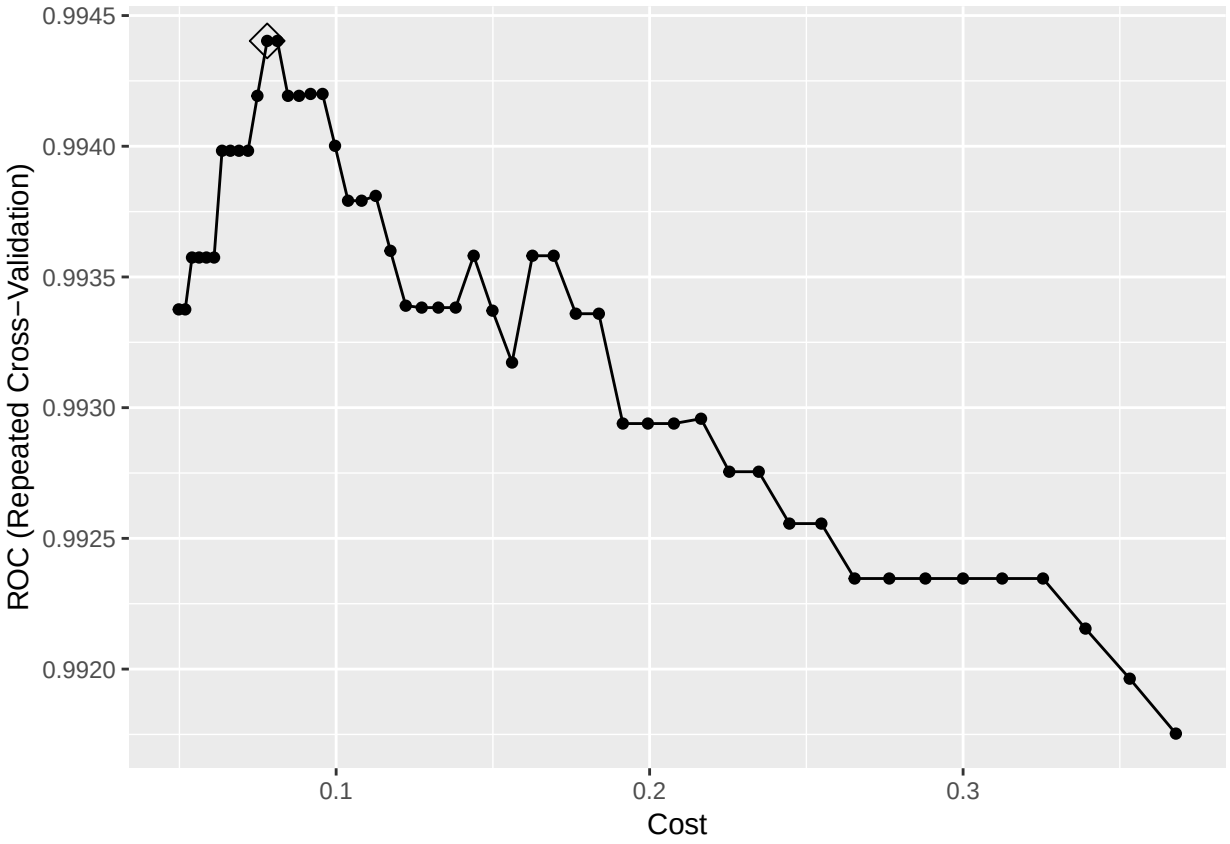

Figure S8 Support Vector Machine (linear kernel) tuning parameter selection

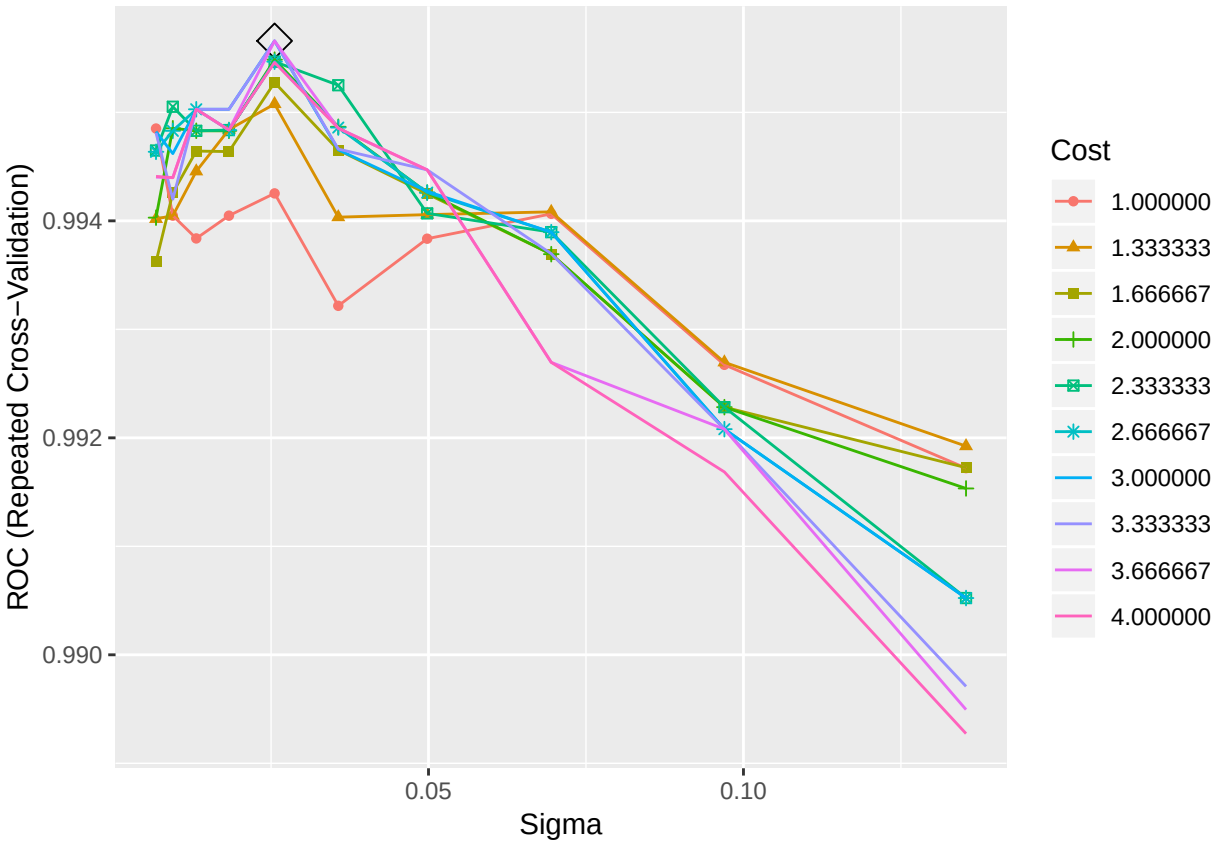

Figure S9 Support Vector Machine (radial kernel) tuning parameter selection

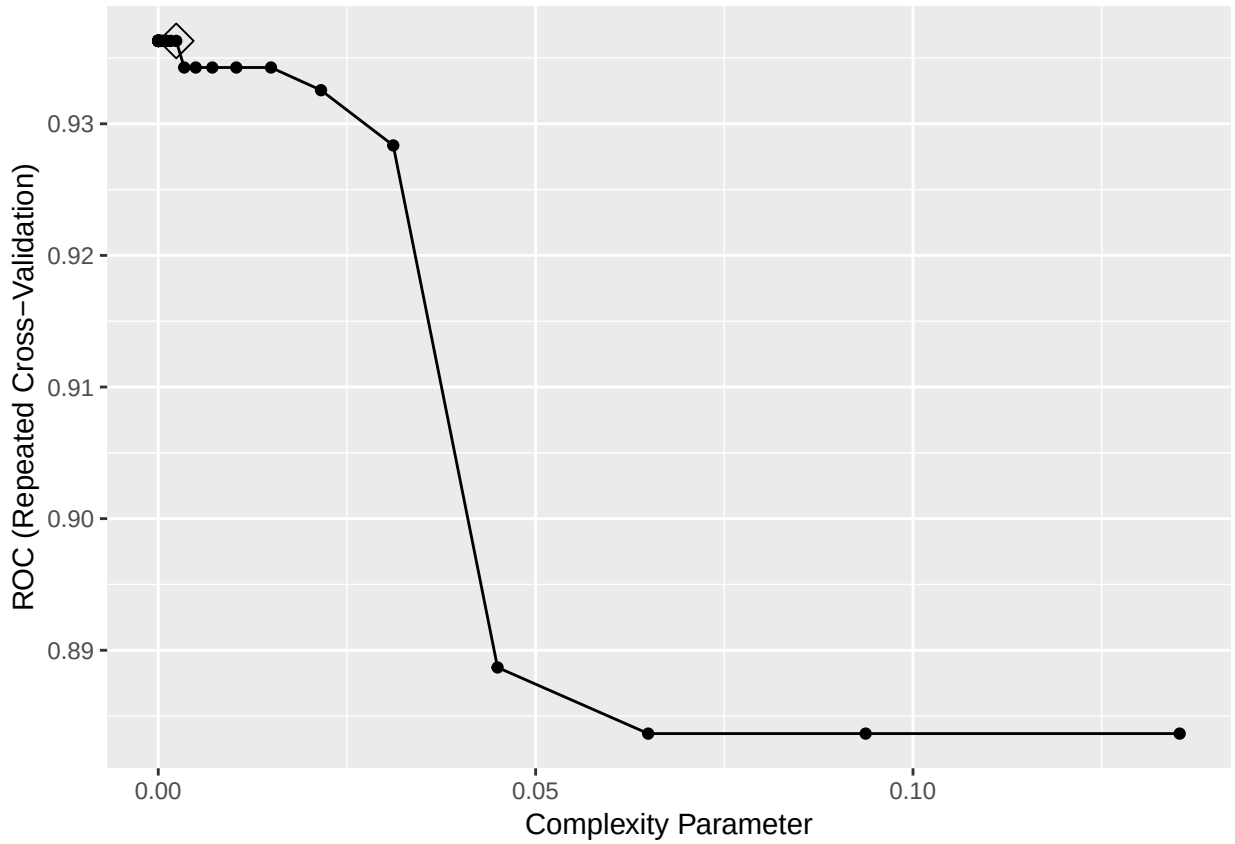

Figure S10 Classification Tree tuning parameter selection

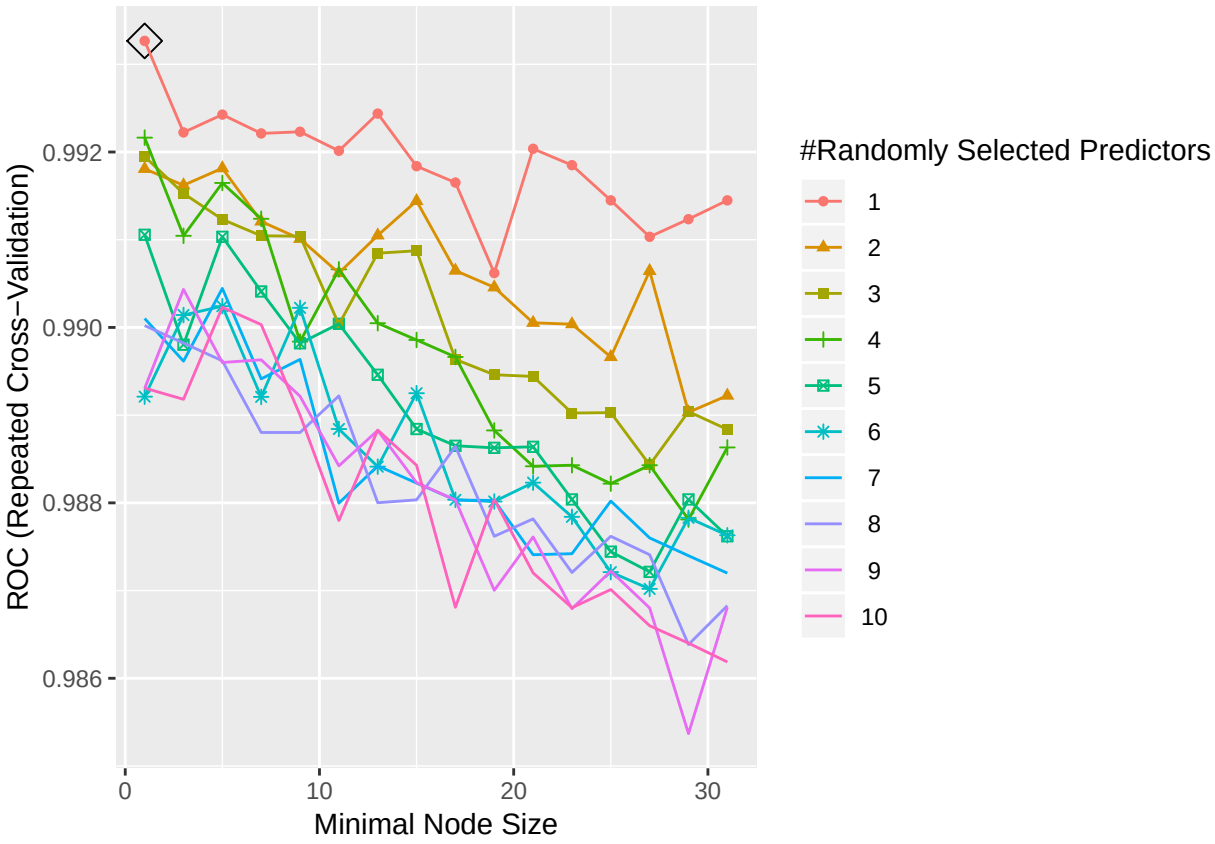

Figure S11 Random Forests tuning parameter selection

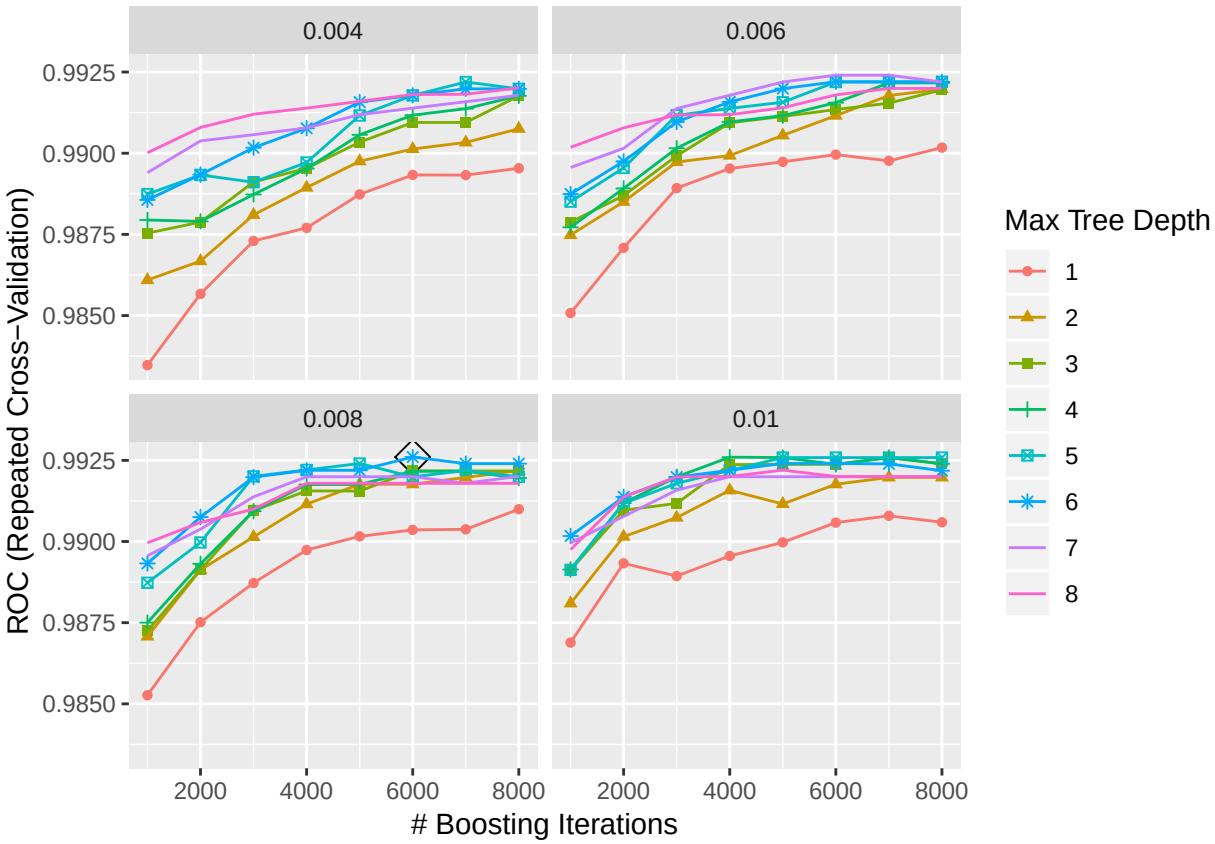

Figure S12 Binomial loss boosting tuning parameter selection

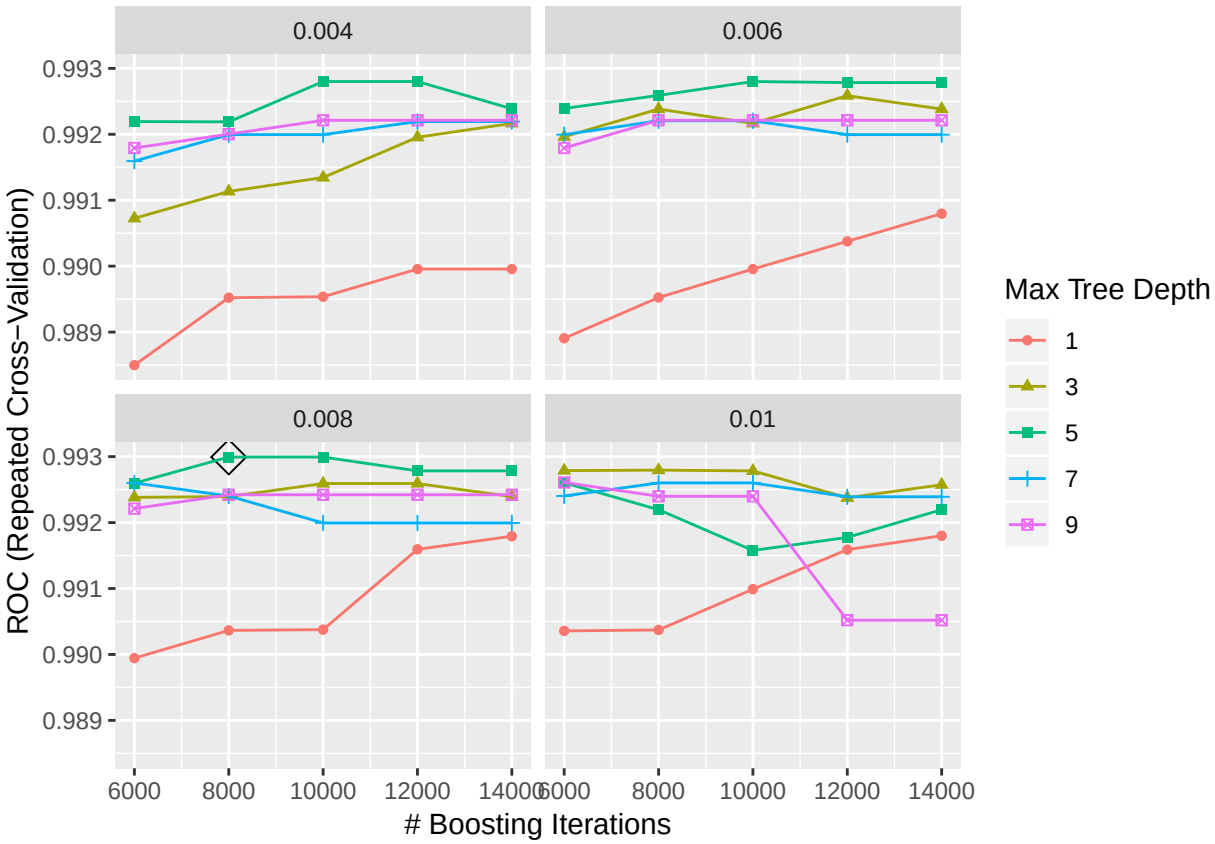

Figure S13 AdaBoosting tuning parameter selection