DS2 Final

M. de Ferrante, K. Maciejewski, P. Batten

April 26, 2018

Intro

For our analysis we used Breast Cancer data from the Wisconsin Breast Cancer Database, compiled by Dr. William Wolberg MD, in the early 1990s, found in the `mlbench` R package. The objective of this dataset is to to use characteristics of tumors such as cell size and texture to classify the tumors as either benign or malignant. The goals of this analysis are to identify the best learning method for classifying tumors.

Methods used include logistic regression, K-means clustering, Support Vector Machine, LDA, Random Forests, and KNN. While it will be helpful to identify one specific method for predicting future classifications of tumor, it is still advisable to consider the output from all methods, albeit with weighted considerations based on how effective the models are. Ultimately, while machine learning is a very powerful tool for making highly important medical classifications, user discretion should be taken into account when considering interpretability, such as potentially using a simpler model that has nearly the same power as a more complex model. Since our research area is highly regulated and in the clinical setting, we chose to use an untouched testing set to evaluate the performance of our methods. We also examine the kappa statistic as a measure of prediction accuracy since our outcome is unbalanced.

Exploratory data analysis

Above, we load in the data and remove the ID column as it is not needed.

Each variable except Class is loaded as 11 numerical factors with values ranging from 0 through 10. Class is benign or malignant, and this is the variable of interest. There are 16 missing values in bare nuclei, as seen in the summary above.

The variables included in the dataset are Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses, and Class. There are 458 subjects with benign growths and 241 with malignant.

Below, we convert the factors into numeric values and remove the NA's. Since there are few missing values compared to the number in the dataset, removing them should not effect our overall analysis power.

Summary plots

The correlation plot below does not give much information. As expected, most measures have correlation with the outcome. Mitoses, however, has almost no corellation.

Below are the density plots for the variables. The blue line signifies benign and pink are malignant. We see that there are few malignant subjects with the following: normal nuclei, mitoses, marginal adhesion, single epithelial cell size, uniformity of cell size, uniformity of cell shape.

There are differences in the density plots of the following varaiables: bland chromatin, bare nucleoli, clump thickness.

The box plots below also show differences in the distributions for bland chromatin, bare nuclei, clump thickness.

Supervised Analyses

Logistic Analysis

Here is a generalized linear model with all variables included.

At $\alpha = 0.05$ the following appear significant :

- Cl.thickness
- Marg.adhesion
- Bare.nuclei
- Bl.cromatin

We rerun the generalized linear model with only the significant values:

(conclusion/ explanation here)

Below are the feature plots for only the significant predictors, so that we may better see the difference in distributions between the variable and the outcome.

Now we split our data into a training and test set so we can make predictions

There were only 13 incorrect predictions; 5 benign were predicted to be malignant and 8 that were truly malignant were predicted to be benign. Logistic has 95% correct response for test, which is pretty good.

The area under the ROC curve is 99%.

Sensitivity is 91%

Specificity is 97%

PPV is 94% and NPV is 96%

LDA and QDA

Linear Discriminant Analysis (LDA) resulted in 95.6% accuracy and a kappa statistic of .9 when comparing predictions from the model generated with the training data to the test data classes. Additionally sensitivity was 90.22% and specificity was 98.34% for LDA. For Quadratic Discriminant Analysis (QDA), there was also 95.6% prediction accuracy for the test data and the kappa statistic was slightly improved with a statistic of .9032. QDA had sensitivity of 96.74% and specificity of 95.03%. With higher sensitivity, we have a higher probability of the model accurately predicting malignancy (versus a tumor being benign) given that someone has a malignant tumor. This is important for being able to treat patients using chemotherapy or performing surgery if necessary. With high specificity, we have a higher probability of predicting that someone has a benign tumor given that they have a benign tumor. This is important because we do not want to subject patients to unnecessary risk by performing unnecessary surgery, or putting them on chemotherapy if they don't need to be since it can be very damaging to a patients health. Since both of these values are extremely important for patient health and safety, QDA would be a better choice for a model since we sacrifice too much in sensitivity with LDA. QDA ends up maximizing these values almost equally.

KNN

After comparing the area under the ROC curve and the percentage of acccuracy in predicting tumor type, the best number of neighbors was chosen to be 3. This value appears to maximize accuracy without losing too much in area under the curve. The chosen number of neighbors has area under the curve equal to .4378 and prediction accuracy against the test set of 96.7%. The kappa statistic is .9256. The sensitivity and specificity are 93.48% and 98.34%, respectively.

Support Vector Machine

The linear kernel with tuning resulted in better performance using cost = .01, and with this model the prediction accuracy on the test data was 95.6% with a kappa statistic of .9011. The sensitivity for the linear kernel was 94.44% and the specificity was 96.17%. Support vector classification with a radial kernel had test set accuracy of 94.87% and a kappa statistic of .8858 (after tuning the parameters). The sensitivity for the radial kernel was 91.49% and the specificity was 96.65%. In comparing prediction accuracy of support vector classification using a linear kernel versus a radial kernel after tuning parameters, the linear kernel performed better on this data.

Random Forest

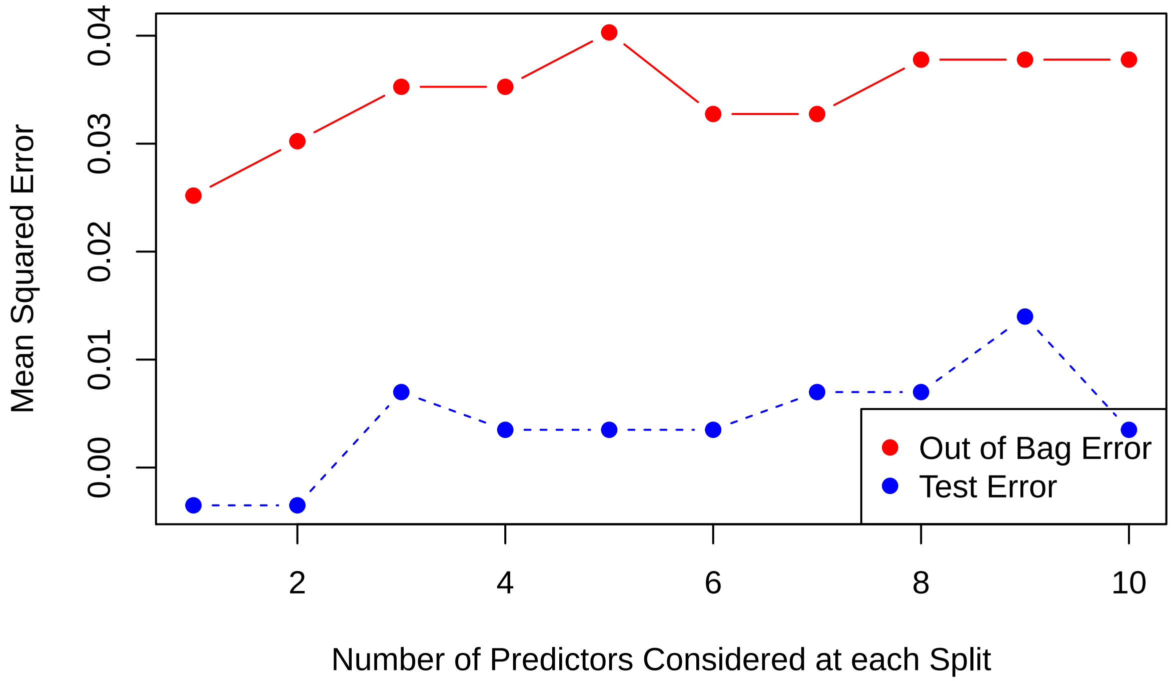
```
confusionMatrix(pred, BreastCancer.test$Class, positive = 'malignant')
```

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction  benign malignant
##   benign       176         7
##   malignant      6        97
##
##                Accuracy : 0.9545
##                  95% CI : (0.9235, 0.9756)
##     No Information Rate : 0.6364
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9016
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.9327
##             Specificity : 0.9670
##          Pos Pred Value : 0.9417
##          Neg Pred Value : 0.9617
##              Prevalence : 0.3636
##          Detection Rate : 0.3392
##    Detection Prevalence : 0.3601
##       Balanced Accuracy : 0.9499
##
##        'Positive' Class : malignant
##
```

```
importance(rf, type = 2)
```

```
##                 MeanDecreaseGini
## Id                      1.463590
## Cl.thickness            3.111236
## Cell.size             130.173519
## Cell.shape              5.334126
## Marg.adhesion           3.292916
## Epith.c.size            1.036150
## Bare.nuclei            11.093291
## Bl.cromatin            11.833287
## Normal.nucleoli         9.643289
## Mitoses                 1.030384
```

```
matplot(1:mtry , cbind(oob.err,test.err), pch=19 , col=c("red","blue"),type="b",ylab="Mean Squared Erro
legend("bottomright",legend=c("Out of Bag Error","Test Error"), pch=19, col=c("red","blue"))
```



Applying random forests techniques to predict tumor classification results in 3 variables being deemed important in contributing to this classification, as suggested by the Oob error being minimized at 3 predictors and the relatively low test error at 3 predictors. Restricting the number of variables used at each split per tree to three yields the best out-of-bag error, as well as the (tied-for) best test error. 3 variables is also relatively simple, which is important to consider when building classifier models. The 3 variables that most effectively decrease impurity at each decision node are cell size, bland chromatin (which measures texture of thhe nucleus), and bare nuclei, a term used to describe the composition of the nucleus. The accuracy of the random forests in predicitng tumor type are as follows: Accuracy = 95.45%, Kappa statistic = 0.9012, Sensitivity = 0.9231, Specificity = 0.9725 (where sensitivity predicts Malignant tumors). Comparisons to other learning methods can be seen in the table in the Conclusion section.

Conclusions

Overall Comparison of Methods

```
library(knitr)

accuracy <- c("95.24%" , "95.45%", '96.7%', "95.6%", "95.6%", "95.6%")
kappa_stat <- c(.8926 , .9012, .9256, .9, .9032, .9011)
sensitivity <- c("91.30%", '92.31%', "93.48%", "90.22%", "96.74%", "94.4%")
specificity <- c("97.24%", "97.25%", "98.34%", "98.34%", "95.03%", "96.17%")


summary <- data_frame(accuracy, kappa_stat, sensitivity, specificity)

row.names(summary) <- c("Logistic Regression", "Random Forest", "KNN","LDA","QDA", "Support Vector Clas
```

```r
colnames(summary) <- c("Test Accuracy", "Kappa Statistic", "Sensitivity", "Specificity")

kable(summary, align = "c")
```

	Test Accuracy	Kappa Statistic	Sensitivity	Specificity
Logistic Regression	95.24%	0.8926	91.30%	97.24%
Random Forest	95.45%	0.9012	92.31%	97.25%
KNN	96.7%	0.9256	93.48%	98.34%
LDA	95.6%	0.9000	90.22%	98.34%
QDA	95.6%	0.9032	96.74%	95.03%
Support Vector Classification	95.6%	0.9011	94.4%	96.17%