STAT 239 Homework 4

Tarek Tohme

June 7th, 2018

Data Analysis 1

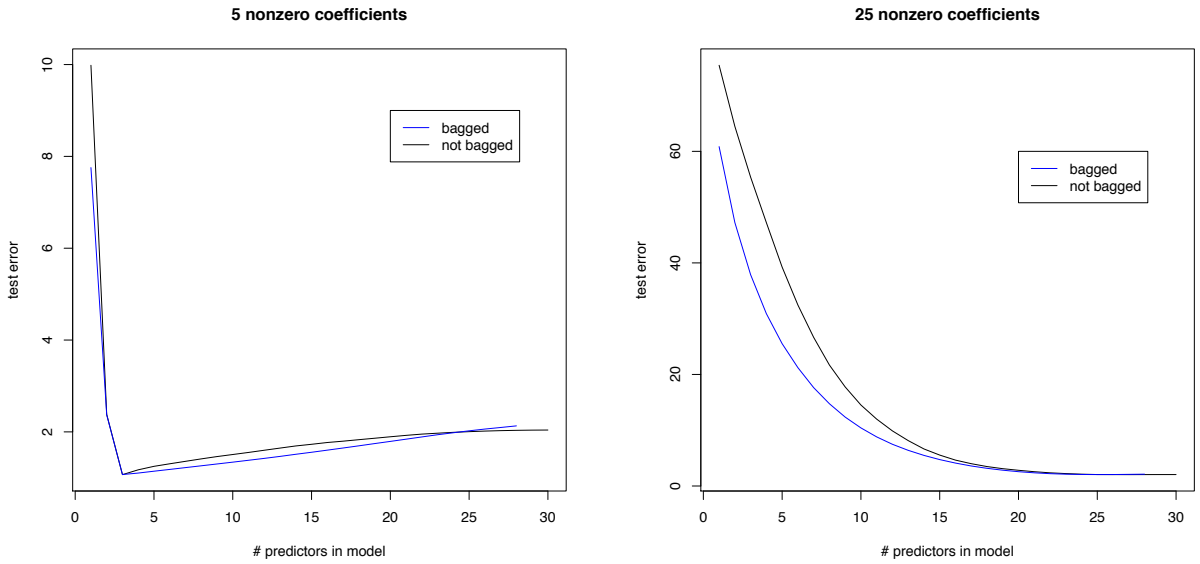

Figure 1

The minimum error is achieved at the same point for both plots. The bagged prediction error is lower for models with few predictors in both situations. As the number of predictors in the model increases, the difference between the bagged error and the regular error shrinks until the bagged error becomes greater. This will be explained in the context of the decomposition of the error, reproduced from the notes4 file below,

$$Error = E(\epsilon^2) + E_X\left[\left(F(X) - E_T\hat{F}(X,T)\right)^2\right] + E_{X,T}\left[\left(\hat{F}(X,T) - E_T\hat{F}(X,T)\right)^2\right]$$

Since we are averaging over 300 simulated samples for both the bagged and non-bagged models, the irreducible error is shrunken by the same amount in both. Therefore it cannot be causing a difference in the

prediction error. The bias term also cannot be causing a difference in the error between both models since, as outlined in the notes, the limit of the bootstrapped model approximates the expected value of the learner over the training data,

$$E_T(h(T)) \approx \lim_{B\to\infty} \frac{1}{B}\sum_{b=1}^{B} h(T_b)$$

which has the same bias as the learner itself. Thus, the difference is due to the variance term. The bagged model's error is reduced most when the bootstrapped learners are least correlated among each other. This happens when there are fewer coefficients in the model being fitted. In a model with a smaller number of coefficients, each coefficient has to account for greater variability in the data. This results in less correlation between coefficients of different learners. An extreme example would be a model with a single nonzero coefficient, fitted on a dataset with many principal components. Such a model would have to account for all the variability of the data with a single predictor, and thus the correlation between two such bootstrapped learners would be directly related to the correlation of the data samples, and so would be very small since the samples are supposedly random. This is the situation where bagging is beneficial, since it reduces this variability in the predictor. In the other extreme, if there were too many predictors in the model, each one would capture a specific part of the variability of the data, and hence would vary less across samples, resulting in a higher correlation among the learners fitted on these samples. If there are more coefficients in the model than needed to capture the meaningful variability in the data (as it happens in the figure on the left after 25 predictors) then eventually the added variance of the useless predictors that capture only noise outweighs the reduction in variance granted by bagging.

Data Analysis 2

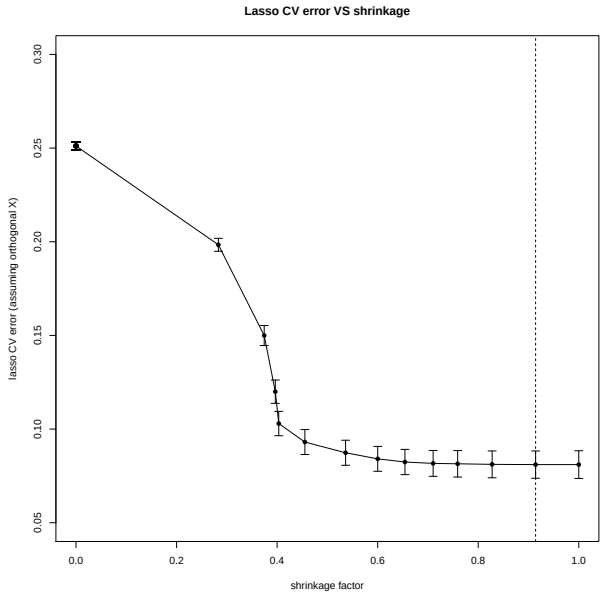

Figure 2

Lasso The optimal shrinkage parameter was obtained by cross validation, as shown in figure 2 (ignoring the one-standard-deviation rule). Then, predictions were made using this shrinkage value and the predicted values were rounded to integers to produce classifications.

Random Forest Using tuneRF the optimal number of variables to select was 1. Figure 3 shows the partial dependence plots.

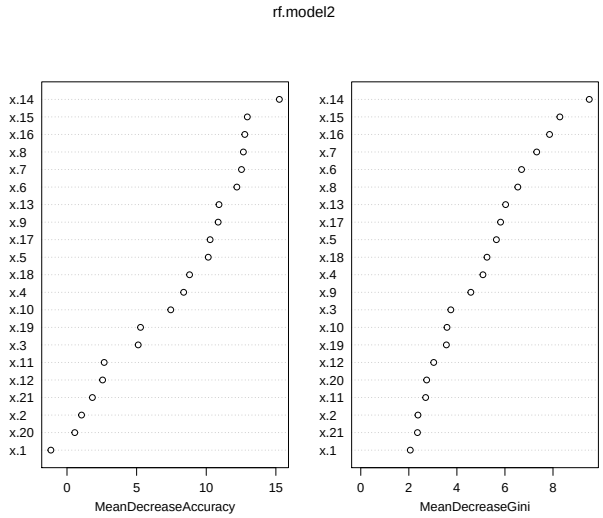

Figure 3

Boosted Trees Four GBM models were tried out having interaction depths of 4,3,2 and 1. The model with interaction depth 1 turned out to have the smallest CV error, which means that stumps seem to be the best choice of base learners. Figure 4 shows a comparison of the CV errors of RF and GBM as a function of the number of trees used in the models. We see that random forests always performs a little better than boosting. The choice of loss for the cross validated error was the average number of misclassifications.

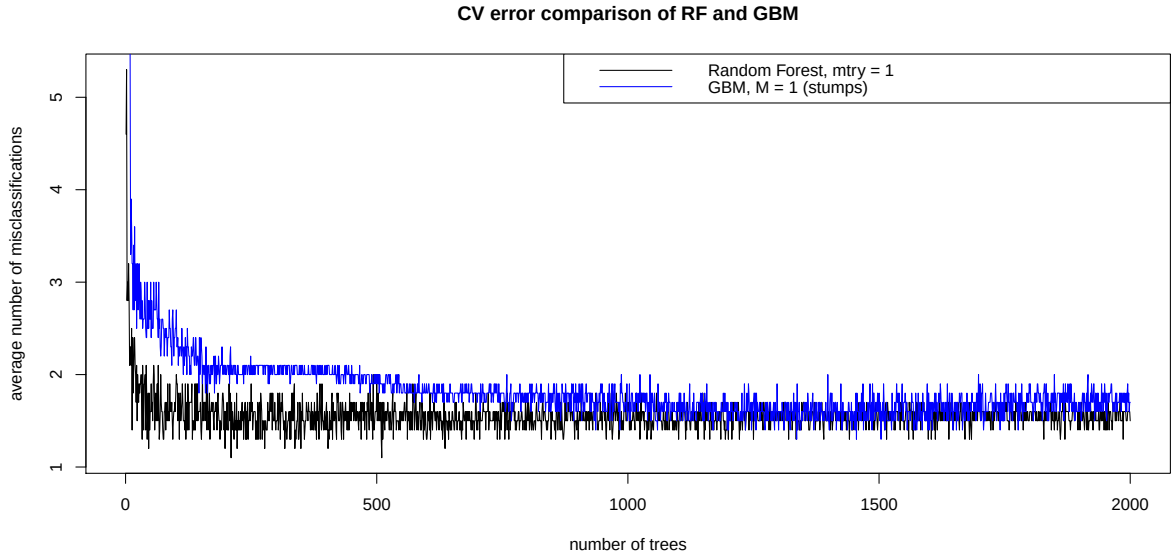

Figure 4

SVM The optimal value of the cost parameter C (figure 5) and the best degree for the polynomial kernel were both determined using cross-validation. Figure 5 (right side) shows a comparison of the number of misclassifications of all the techniques tried out.

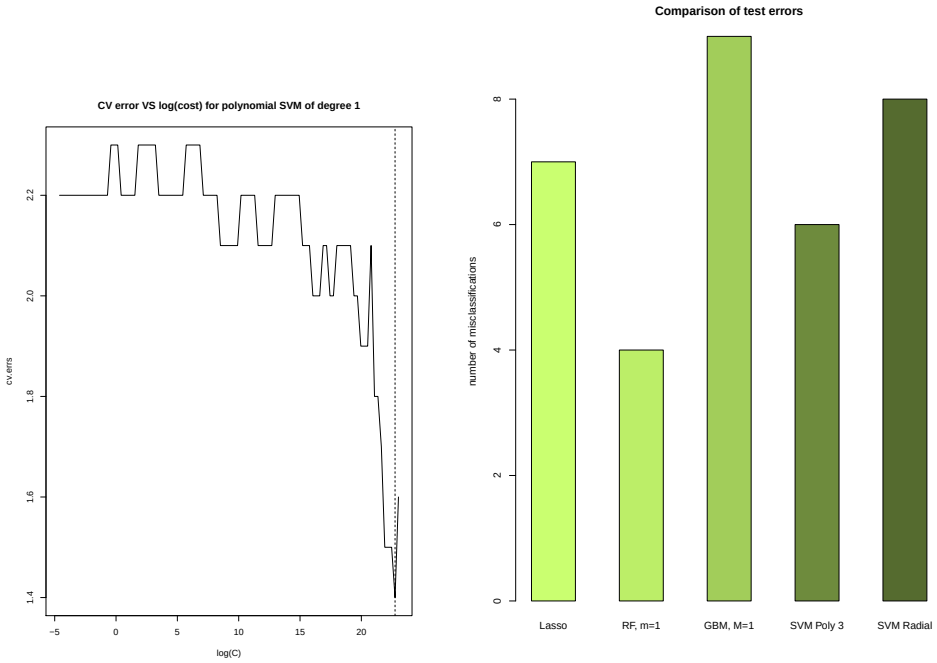

Figure 5

Data Analysis 3

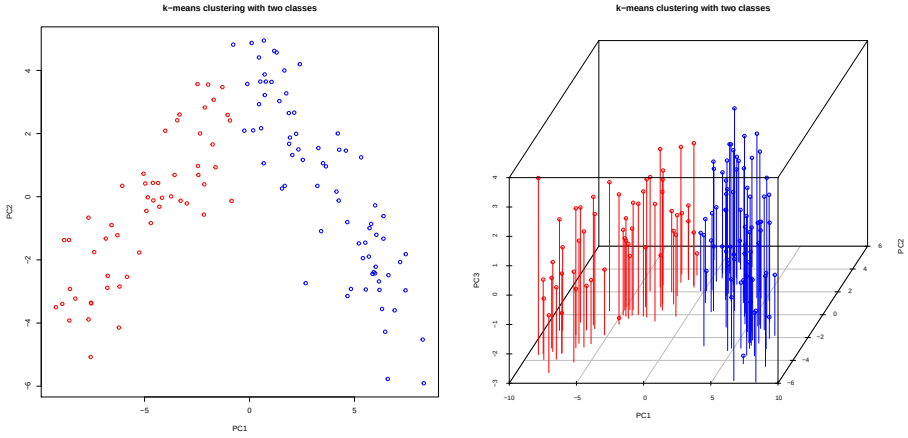

Figure 6

K-Means clustering Figure 6 shows plots of the first principal components of the data. Clearly, k-means clustering successfully identifies the two classes in the data. They can be more or less separated along the first principal component, as most points of class 1 have negative values of PC1, and most class 2 observations

have positive values. The 3d plot shows that the 3rd principal component doesn't play a role in determining the class.

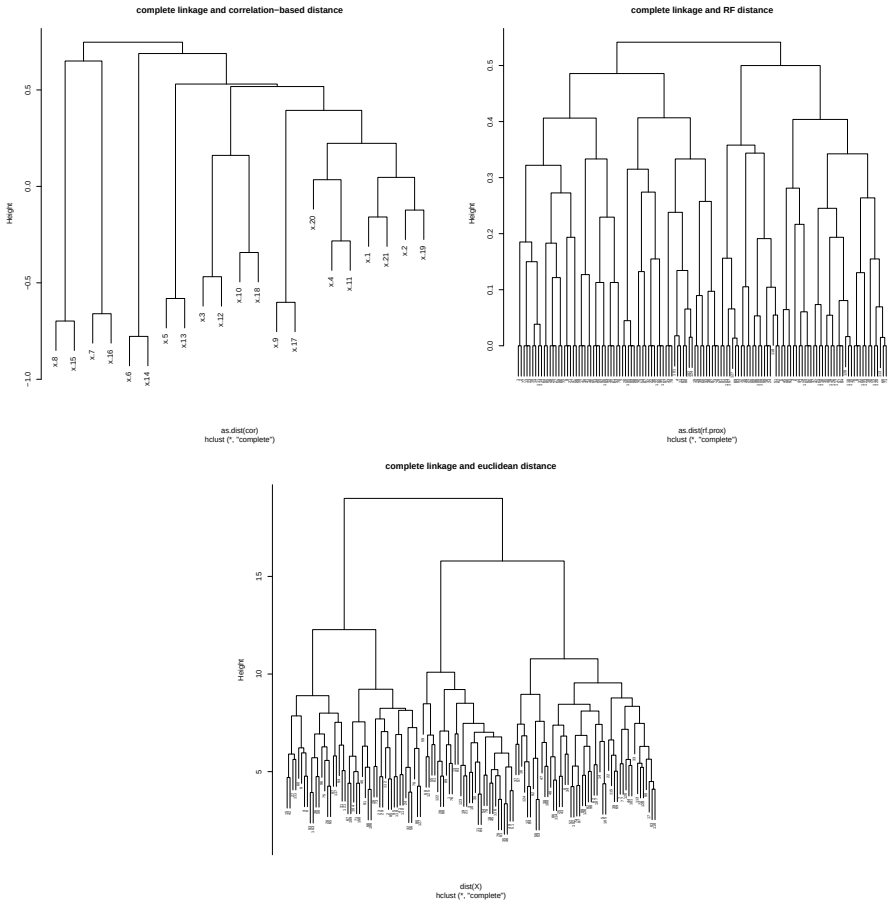

Figure 7

Hierarchical Clustering Examining the dendrograms in figure 7, it seems that correlation-based clustering doesn't give good insight as to how the classes are separated: the separation that gives two classes appears at the very top of the plot. This might be because the predictors that most influence the outcome (class) are not strongly correlated with each other. If they were, then it would show on this dendrogram as two "tight" clusters. The euclidean distance dendrogram tells a different story: there are clearly four clusters at around height 10, which eventually merge to 2 at about height 15. The appearance of the plot suggests that the data is well described as belonging to two clusters, three clusters or four clusters. The RF-distance dendrogram is more "evenly distributed" across height, it doesn't convince us that there are two classes, just like the correlation-based plot. This means that trees don't really capture the class difference in this dataset. If they did, i.e. if observations of class 1 tended to end up at the same terminal node often with other observations of class 1, it would show up on the plot as if the observations were close.

Theoretical Answers

12.1) The problem asks to prove that

$$\min_{\beta_0,\beta} \sum_{i=1}^{N} [1 - y_i f(x_i)]_+ + \frac{\lambda}{2}||\beta||^2 = \min_{\beta_0,\beta} \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N} \xi_i$$

ξ_i is the distance by which the point i is on the wrong side of its margin. The margin is represented by

$$\begin{aligned} f(x) = x^T\beta + \beta_0 &= \pm M \\ y_i f(x_i) &= M \end{aligned}$$

points that are inside the margin by an amount ξ_i are represented as

$$y_i f(x_i) = M(1 - \xi_i) \quad \xi_i \geq 0$$

normalizing as was done in ESL page 418 eq. 12.4, we get

$$\begin{aligned} y_i f(x_i) &= 1 - \xi_i \\ 1 - y_i f(x_i) &= \xi_i \quad \xi_i \geq 0 \\ \xi_i &= (1 - y_i f(x_i))^+ \end{aligned}$$

Choosing $\lambda = \frac{1}{C}$ and multiplying the expression by C (which doesn't affect the equation since we're finding the minimum), the equation is verified.

12.2) The problem asks to prove that

$$\min_{\beta_0,\beta} \sum_{i=1}^{N} [1 - y_i f(x_i)]_+ + \frac{\lambda}{2}||\beta||^2 = \min_{\beta_0,\alpha} \sum_{i=1}^{N} [1 - y_i f(x_i)]_+ + \frac{\lambda}{2}\alpha^T \mathbf{K}\alpha$$

We know from ESL equations 12.10 and 12.20 that

$$\beta = \sum_{i=1}^{N} \alpha_i y_i h(x_i)$$

Writing out the summations, we get

$$\begin{aligned} \alpha^T \mathbf{K}\alpha &= \sum_{i=1}^{N}\sum_{i'=1}^{N} \alpha_{i'} \mathbf{K}(x_i, x_{i'})\alpha_i \\ \beta^T\beta &= \sum_{i=1}^{N}\sum_{i'=1}^{N} \alpha_{i'}\alpha_i h(x_i)h(x_i)y_{i'}y_i \\ &= \sum_{i=1}^{N}\sum_{i'=1}^{N} \alpha_{i'}\alpha_i \mathbf{K}(x_i, x_{i'})y_{i'}y_i \end{aligned}$$

we can remove $y_i y_{i'}$ from the second equation since when a point is misclassified (outside the margin), $\alpha_i = 0$ and when it is correctly classified, $y_i y_{i'} = 1$.