Machine Learning: Problem Set 3

Jingpeng Hong

February 21, 2022

Chapter 6 Question 1

We perform best subset, forward stepwise, and backward stepwise selection on a single data set. For each approach, we obtain $p + 1$ models, containing 0, 1, 2, . . . , p predictors.

a

Which of the three models with k predictors has the smallest training RSS?

The model with k predictors from best subset selection will have the smallest training RSS. Because best subset selection picks the best model from all $\binom{p}{k}$ models that contain exactly k predictors. The forward (backward) stepwise selection will only pick (remove) the best (worst) additional predictor from the previous models with $k - 1$ ($k + 1$) predictors, therefore, they cannot guarantee to find the best possible model among all possible models with k predictors.

b

Which of the three models with k predictors has the smallest test RSS?

Undetermined. We cannot determine which model will have the smallest test RSS. For best subset model, when p is large, the best model may be overfitting because we search in a large space and are likely to find models that look good on the training data, even though they might not have any predictive power on future data.

c

True or False:

i. The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the $(k + 1)$-variable model identified by forward stepwise selection.

True. Because in the $(k+1)$-variable model, we add the variable that gives the greatest additional improvement to the fit based on the k-variable model. Hence, the $(k+1)$-variable model contains all predictors in the k-variable model plus the best additional one from the remaining $p-k$ predictors.

ii. The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the (k + 1)- variable model identified by backward stepwise selection.

True. Because in the k-variable model, we remove the least useful variable based on the $(k+1)$-variable model. Hence, the k-variable model contains all predictors in the $(k+1)$-variable model except the least useful one among the $k+1$ predictors.

iii. The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the $(k+1)$- variable model identified by forward stepwise selection.

False. The forward stepwise selection and backward stepwise selection can have completely different models because their procedures are independent.

iv. The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the $(k+1)$-variable model identified by backward stepwise selection.

False. Similar to the explanation in (iv)

v. The predictors in the k-variable model identified by best subset are a subset of the predictors in the $(k+1)$-variable model identified by best subset selection.

False. The best subset selection picks the best model among all the k-variable/$k+1$-variable models, so they can be either identical or different.

Chapter 6 Question 2 (a-b)

The lasso/ridge, relative to least squares, is:

i. More flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.

ii. More flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

iii. Less flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.

iv. Less flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

(a) For Lasso, the (iii) is true. When λ increases, the flexibility of the lasso regression fit decreases. The bias will increase and the variance of the predictions will decrease. Since the test MSE is a function of the variance plus the squared bias, we can see prediction accuracy improved when increase in bias is less than its decrease in variance.

(b) For Ridge, the (iii) is true. The explanation is similar to that in Lasso.

Chapter 6 Question 3

3. Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}) \qquad \text{s.t.} \quad \sum_{j=1}^{p}|\beta_j| \leq s$$

for a particular value of s. For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

i. Increase initially, and then eventually start decreasing in an inverted U shape.

ii. Decrease initially, and then eventually start increasing in a U shape.

iii. Steadily increase.

iv. Steadily decrease.

v. Remain constant.

(a) As we increase s from 0, the training RSS will (iv) steadily decrease. When s grows, the constraint will have less effect and the model will tend to the OLS. Hence, the training RSS will steadily decrease.

(b) As we increase s from 0, the test RSS will (ii) decrease initially, and then eventually start increasing in a U shape. Let us consider the test RSS as a function of the variance plus the square bias. When we initially increase s from 0

(c) As we increase s from 0, the variance will (iii) steadily increase. When we restrict the coefficient less, the model will have more flexibility to fit the data, then the variance will increase.

(d)As we increase s from 0, the (squared) bias will (iv) steadily decrease. When we restrict the

coefficient less, the model will have more flexibility to fit the data for a lower bias.

(e) As we increase s from 0, the Bayes error rate will (v) remain constant. Given the definition of the Bayes error rate

$$1 - E(\max_j \Pr(Y = j|X))$$

The Bayes error rate is introduced from the classes overlap in the true population, which is not determined by out model setting. Hence, we cannot change the Bayes error rate by increasing s.

Chapter 6 Question 8

The model is:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$

where we have $\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 3$, $\beta_3 = 4$

c

As shown in Figure 1, we can find that after adding the cubic term, we can approximately get the lowest C_p and BIC, and highest adjusted R^2. Hence, the best model should contain predictors X, X^2 and X^3. (the AIC and BIC differ only in how models of different sizes are compared.)

The coefficients are $\hat{\beta}_0 = 1.0615$, $\hat{\beta}_1 = 1.9753$, $\hat{\beta}_2 = 2.8762$, $\hat{\beta}_3 = 4.0176$

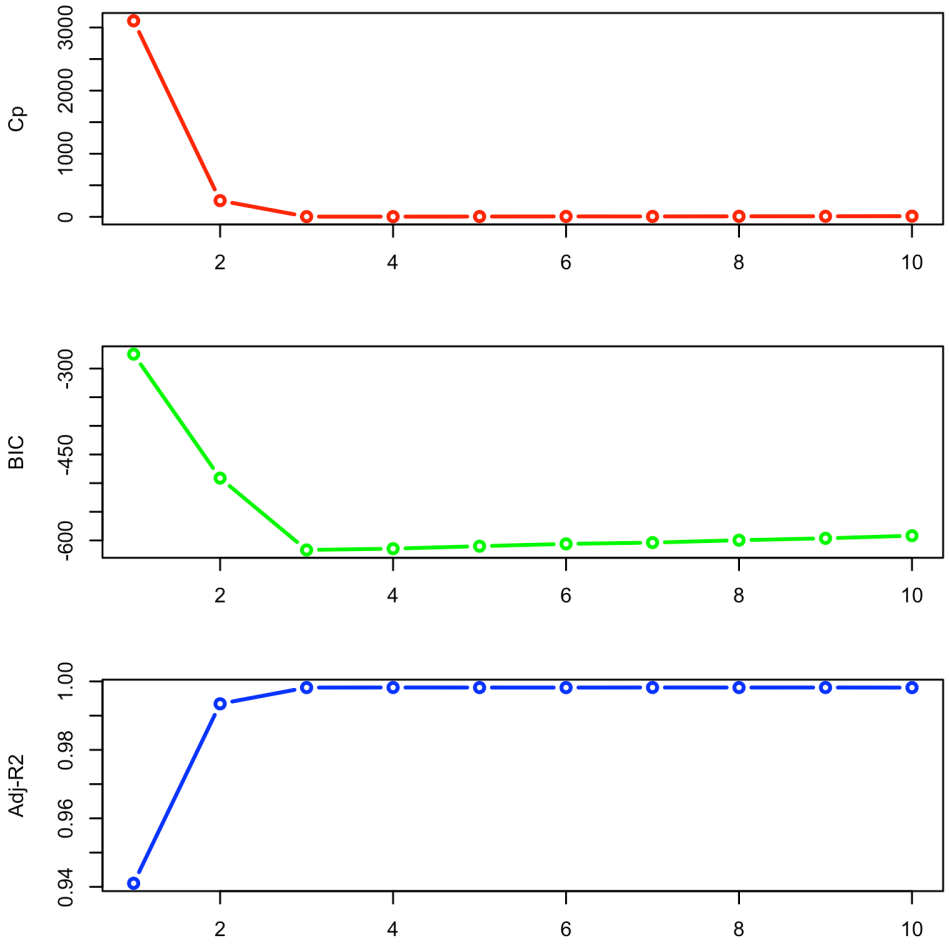

Figure 1: C_p, BIC, and adjusted R^2 (Best Subset Selection)

d

Use forward stepwise selection and backward stepwise selection.

As shown in Figure 2, the results from Forward stepwise selection and Backward stepwise selection are quite similar to that in the best subset selection. We also obtain the model with predictors X, X^2, and X^3 as our best selection. Hence, the coefficients of the best model obtained are exactly identical as the model from best subset selection.

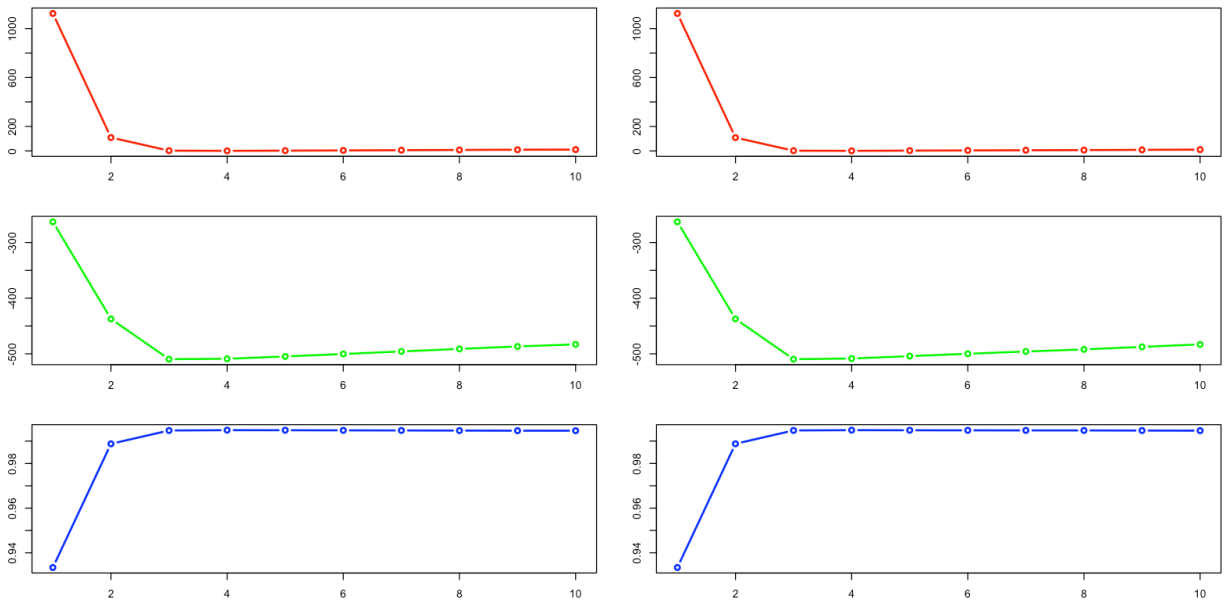

Figure 2: Forward(Left) / Backward(Right) Stepwise Selection

e

From the cross-validation, we find the optimal value of λ is 0.0579. Figure 3 shows the cross-validation error as a function of λ.

With $\lambda = 0.0579$, the results of coefficient estimation are shown in Table 1. First, we can find that the estimates of $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ are closed to the model selection above and our true model. Second, with the sparsity assumption, we see that four of the eleven coefficient estimates are exactly zero. So the lasso model with our best λ by cross-validation contains only X to X^6 and the intercept.

$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$
1.147	2.312	2.718	3.614	0.002	0.066	0.004	0	0	0	0

Table 1: Coefficient estimates from Lasso

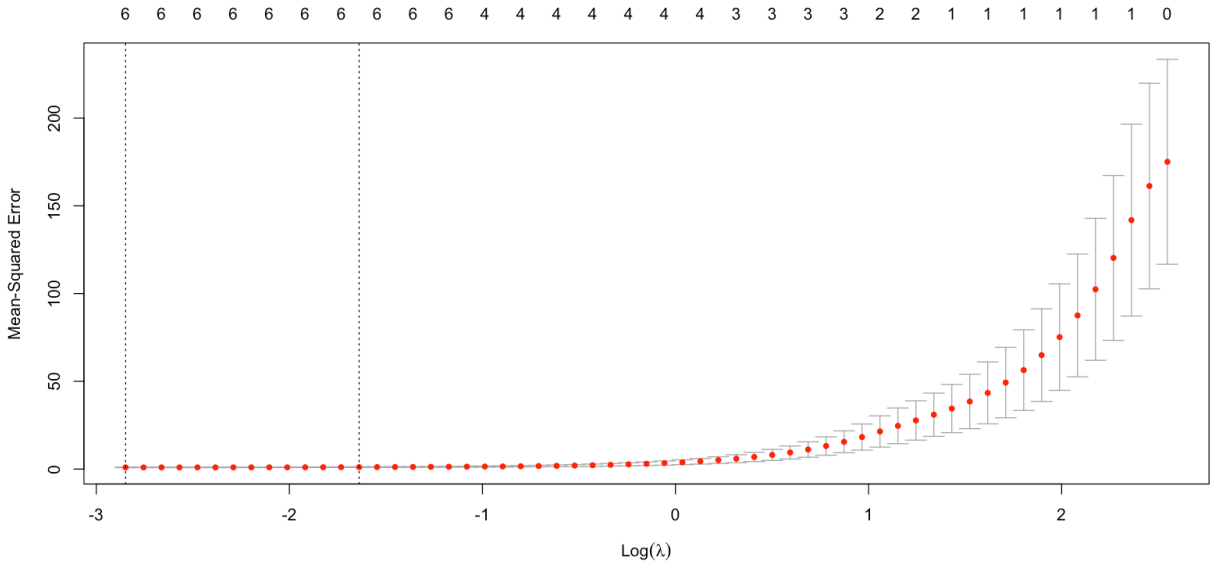

Figure 3: cross-validation error v.s. λ

f

The model is:

$$Y = \beta_0 + \beta_7 X^7 + \epsilon$$

where we have $\beta_0 = 1$ and $\beta_7 = 8$

From the best subset selection, we can find the model $Y = \beta_0 + \beta_7 X^7$ has the smallest BIC. The coefficients estimated are $\hat{\beta}_0 = 0.9589$ and $\hat{\beta}_7 = 8.0008$

From the Lasso, we first find the best $\lambda = 14.14$. With this best λ, the coefficients estimated are shown in Table 2. This results also align with our true model.

$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$
1.9474	0	0	0	0	0	0	7.7505	0	0.0029	0

Table 2: Coefficient estimates from Lasso

Chapter 8 Question 3

Gini Index:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) = \hat{p}_{m1}(1 - \hat{p}_{m1}) + \hat{p}_{m2}(1 - \hat{p}_{m2}) = 2(\hat{p}_{m1}(1 - \hat{p}_{m1}))$$

Cross-entropy:

$$D = -\sum_{k=1}^{K} \hat{p}_{mk}\log(\hat{p}_{mk}) = -\hat{p}_{m1}\log(\hat{p}_{m1}) - (1 - \hat{p}_{m1})\log(1 - \hat{p}_{m1})$$

Classification error rate:

$$E = 1 - \max(\hat{p}_{mk}) = 1 - \max(\hat{p}_{m1}, 1 - \hat{p}_{m1}) = 1 - (\hat{p}_{m1}\mathbb{1}\{\hat{p}_{m1} \geq 0.5\} + (1 - \hat{p}_{m1})\mathbb{1}\{\hat{p}_{m1} < 0.5\})$$

The plots are shown in Figure 4

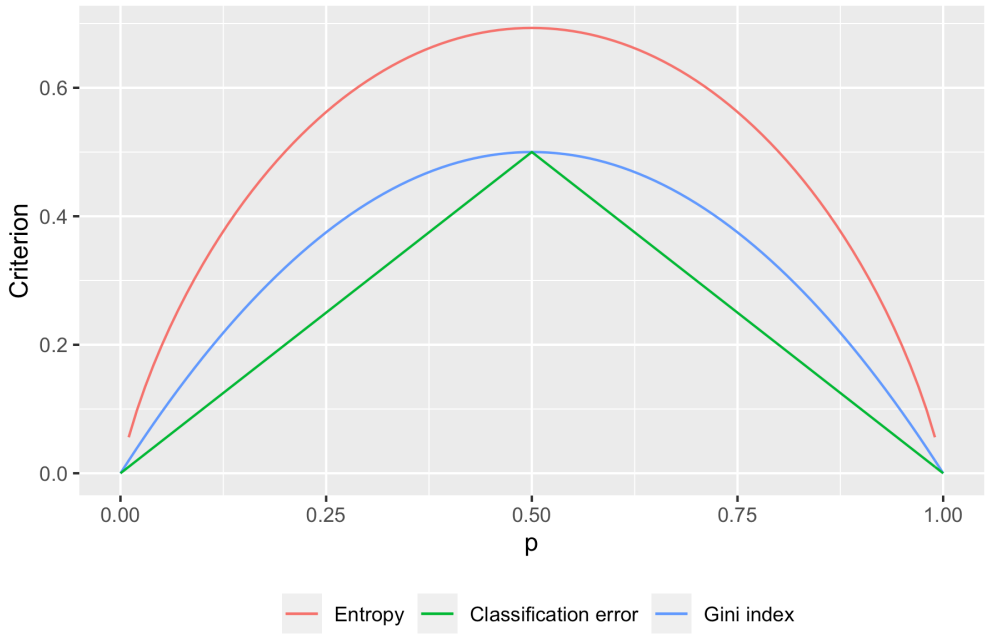

Figure 4: Gini index, Classification error, and Entropy as a function of $\hat{p}_{m1}$

Chapter 8 Question 8

b

The tree is shown in Figure5. First, from the tree, we can find that the most important indicator of Sales is shelving location, since the first branch differentiates Good locations from Bad and Medium locations. Second, other variables used in the construction of the tree are "Price", "Age", "Advertising", "CompPrice" and "US".

The test set MSE associated with this regression tree is 4.922.

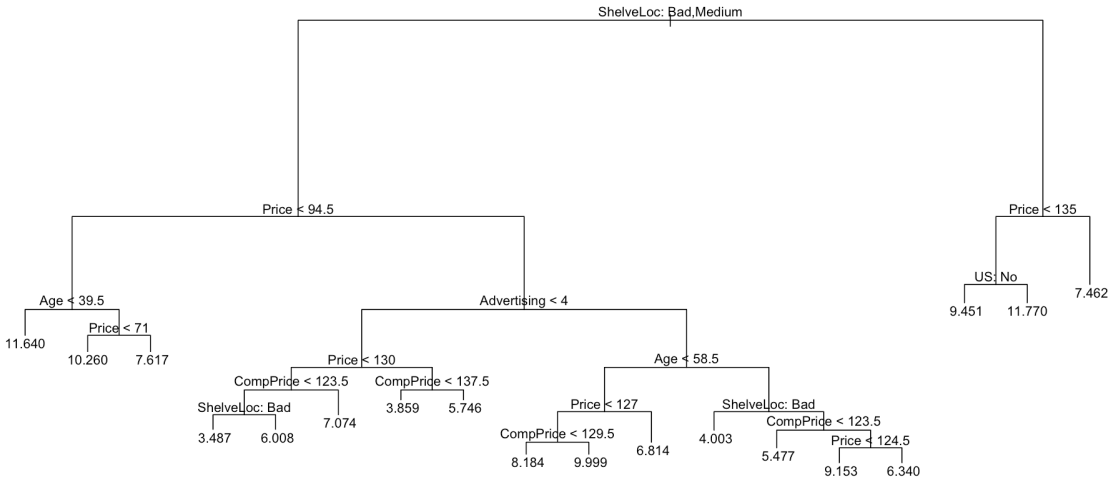

Figure 5: The unpruned tree for Sales in Carseats

c

From the cross-validation, we find that the most complex tree is selected (shown in Figure 6). Hence, pruning the tree cannot improve the test error rate.

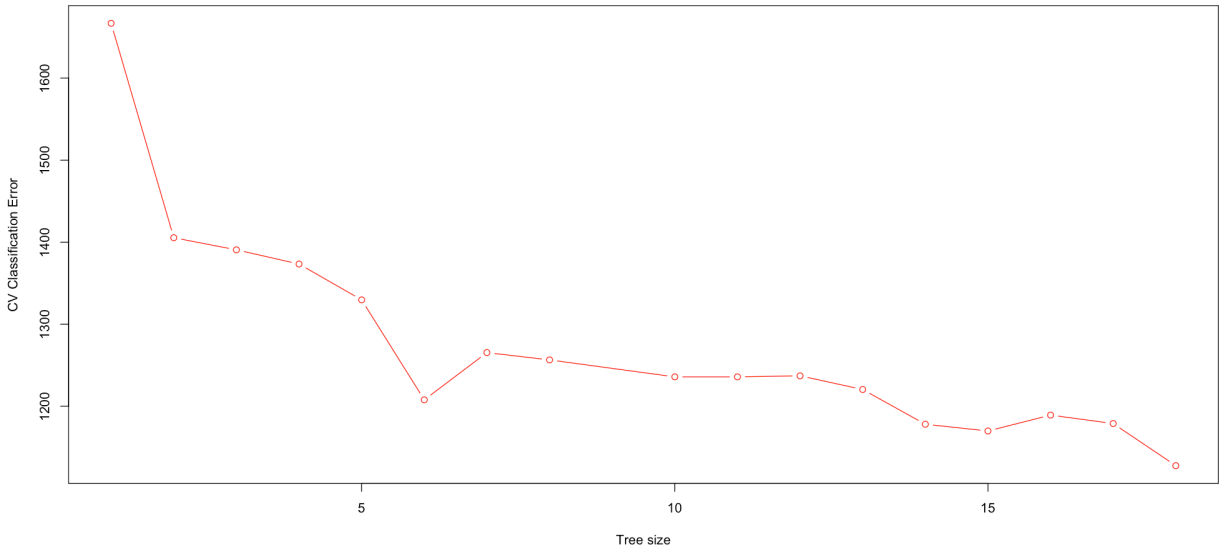

Figure 6: Cross validation

d

With the bagging approach, the test error rate is 2.585. The most important variables are "Price" and "ShelveLoc".

e

First, we consider $m = \sqrt{p} = 3$. The test MSE is 3.03, and the most important variables are "Price" and "ShelveLoc". Second, we consider the range of m from 1 to 9, the relationship between m and the test MSE is shown in Figure 7.

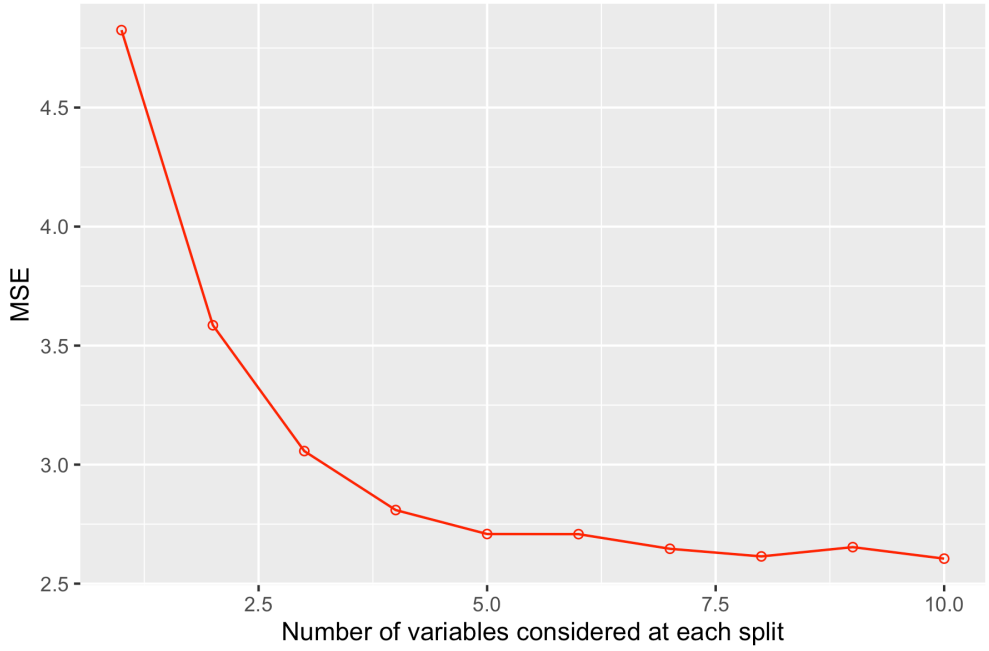

Figure 7: m and the test error rate

Chapter 8 Question 9

b

From the summary, we can find that variables used as internal nodes in the tree showin in Table 3. The number of terminal nodes is 9, and the training error rate is 15.88% (127 / 800)

"LoyalCH"	Customer brand loyalty for CH
"PriceDiff"	Sale price of MM less sale price of CH
"SpecialCH"	Indicator of special on CH
"ListPriceDiff"	List price of MM less list price of CH
"PctDiscMM"	Percentage discount for MM

Table 3: Variables in internal nodes

c

Consider the terminal node 21: For those 0.280875 < LoyalCH < 0.5036, PriceDiff < 0.05 and SpecialCH > 0.5, the overall prediction for this branch is "CH". The fraction of observations in this branch that take on values of CH and MM is 0.60 and 0.40, repectively. The number of observations in this branch is 15 and the deviance is 20.19.

d

The plot of the tree is shown in Figure 8. From the plot, we can find the most important indicator of the choice of purchase is the customer brand loyalty. When LoyalCH>0.5036, the customer will always choose CH except that the price of MM is low enough (ListPriceDiff<0.236 and PctDiscMM>0.196). Except for the customer brand loyalty, the price and the special are also important indicators for customers' choice.

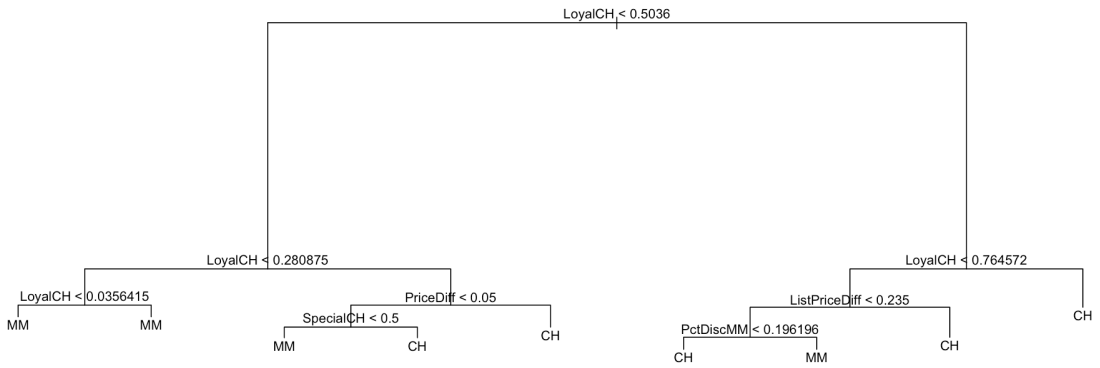

Figure 8: The unpruned tree for Purchase in OJ

e

The confusion matrix comparing the test labels to the predicted test labels is shown in Figure 4. The test error rate is 17.04% $((38 + 8)/(160 + 64 + 38 + 8))$

	CH	MM
CH	160	38
MM	8	64

Table 4: Confusion matrix

f-h

The plot with tree size on the x-axis and cross-validated classification error rate on the y-axis is shown in Figure 9. The tree with 7 terminal nodes results in the lowest cross-validation error rate, with 152 cross-validation errors.

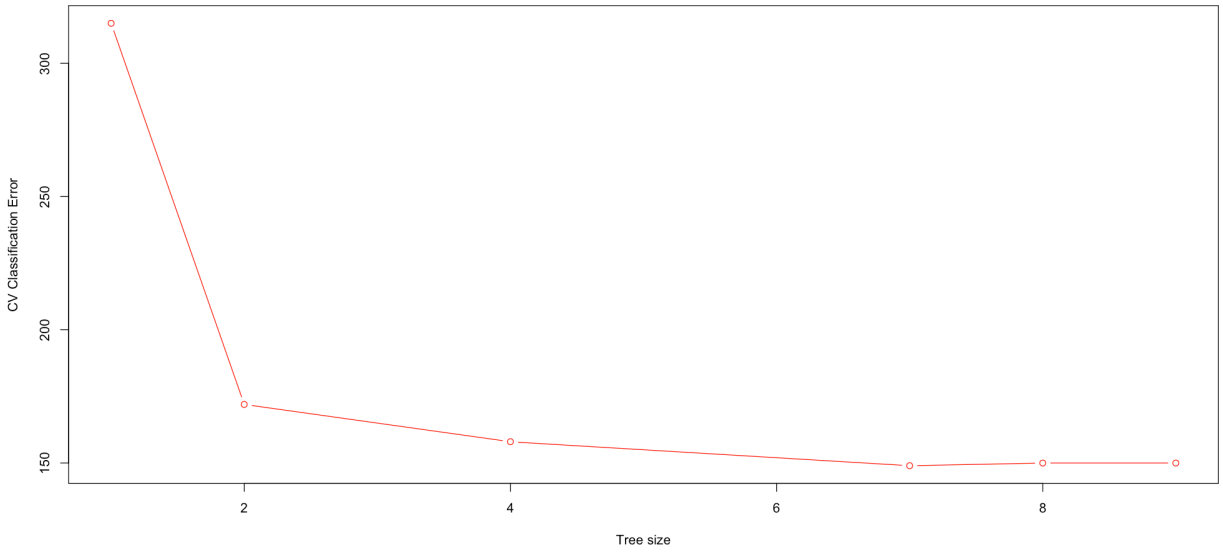

Figure 9: Tree size & Cross-validated classification error rate

i-k

We prune the tree to obtain the seven-node tree shown in Figure 10.

The training error rate of the pruned tree is 16.25%, which is higher than that of the unpruned tree, which is 15.88%

The test error rate of the pruned tree is 16.30%, which is lower than that of the unpruned tree, which is 17.04%

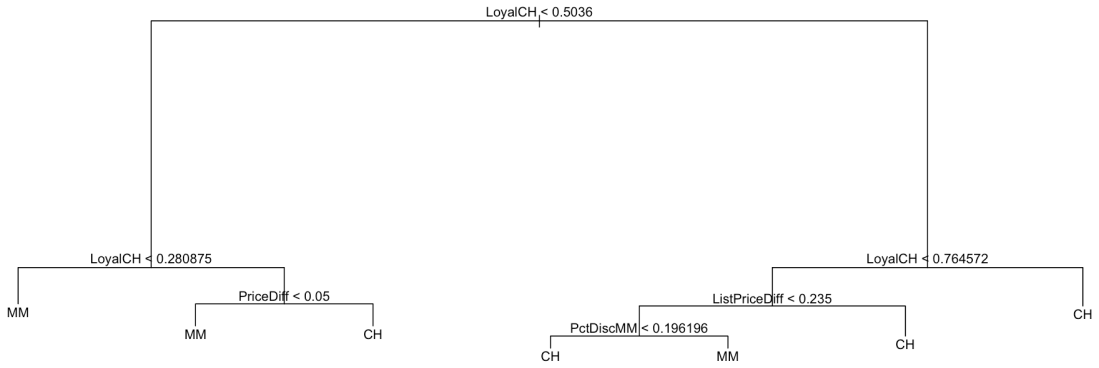

Figure 10: Pruned tree with 7 terminal nodes

Chapter 8 Question 10

c-d

The plot with different shrinkage values on the x-axis and the corresponding training/test set MSE on the y-axis is shown in Figure 11. We find that the shrinkage value $\lambda = 0.1$ can lead to the smallest test set error 0.263.

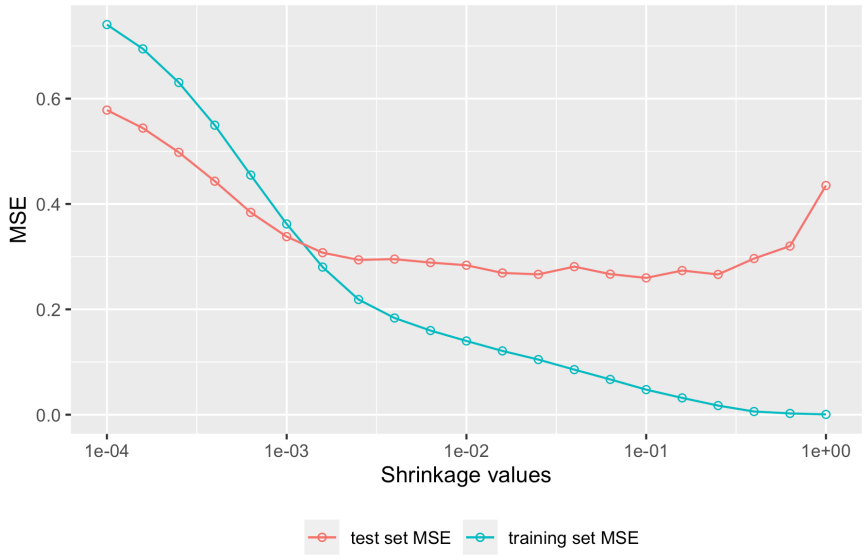

Figure 11: Training/Test set MSE with different shrinkage values

e

First, we consider the linear regression, of which the test MSE is 0.492; Second, we consider the Lasso, of which the test MSE is 0.476. Hencer, we can see that the MSE from boosting is smaller than that from linear regression or Lasso.

f

With $\lambda = 0.1$, "CAtBat" appears to be the most important predictors in the boosted model so far.

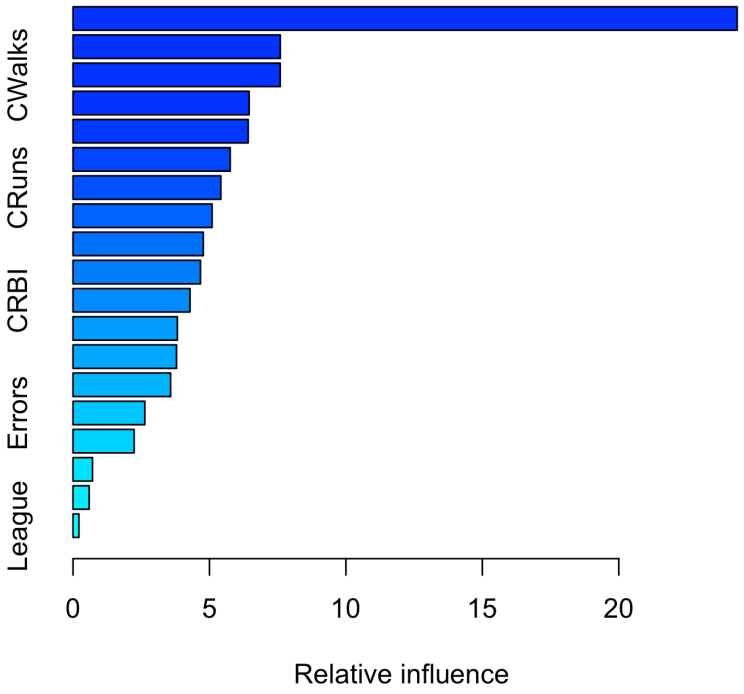

Figure 12: Variable Importance Plot

g

When applying bagging to the training set, the test set MSE for this approach is 0.23.