Jayoung Kang

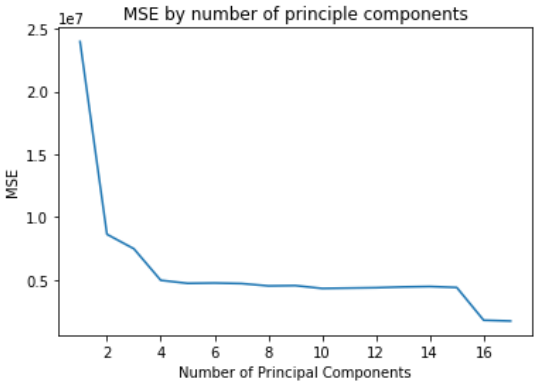
Homework 4

Chapter 6

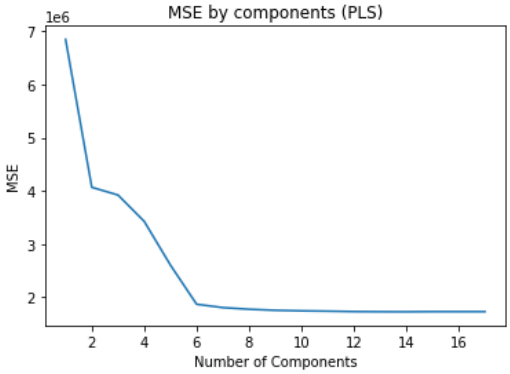
Question 9

b) Fit a linear model using least squares on the training set, and report the test error obtained.
   - Test error: 1384800.4507778855

e) Fit a PCR model on the training set, with M chosen by cross validation. Report the test error obtained, along with the value of M selected by cross-validation.
   - Test error: 1384800.4507778964
   - M: 17



f) Fit a PLS model on the training set, with M chosen by cross validation. Report the test error obtained, along with the value of M selected by cross-validation.
   - Test error: 1382785.9501680527
   - M: 14

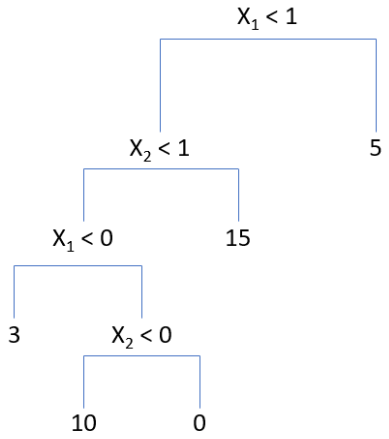


g) Comment on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these five approaches?
   - Based on the MSE we can observe that the three models all perform relatively well in terms of accurately predicting the number of college applications received. We can see that the the PLS model does slightly better than the other two models.

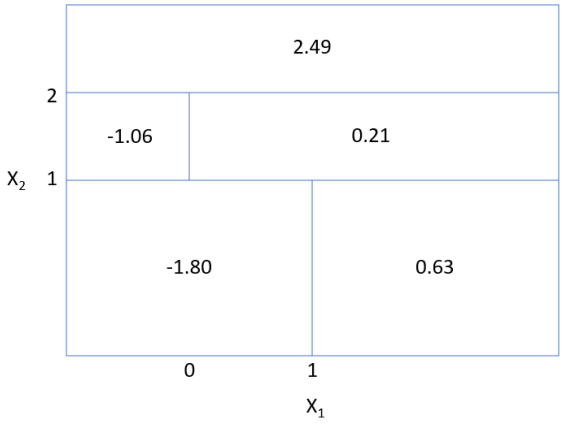
Chapter 8

Question 4

a) Sketch the tree corresponding to the partition of the predictor space illustrated in the left-hand panel of Figure 8.12. The numbers inside the boxes indicate the mean of Y within each region.



b) Create a diagram similar to the left-hand panel of Figure 8.12, using the tree illustrated in the right-hand panel of the same figure. You should divide up the predictor space into the correct regions, and indicate the mean for each region.



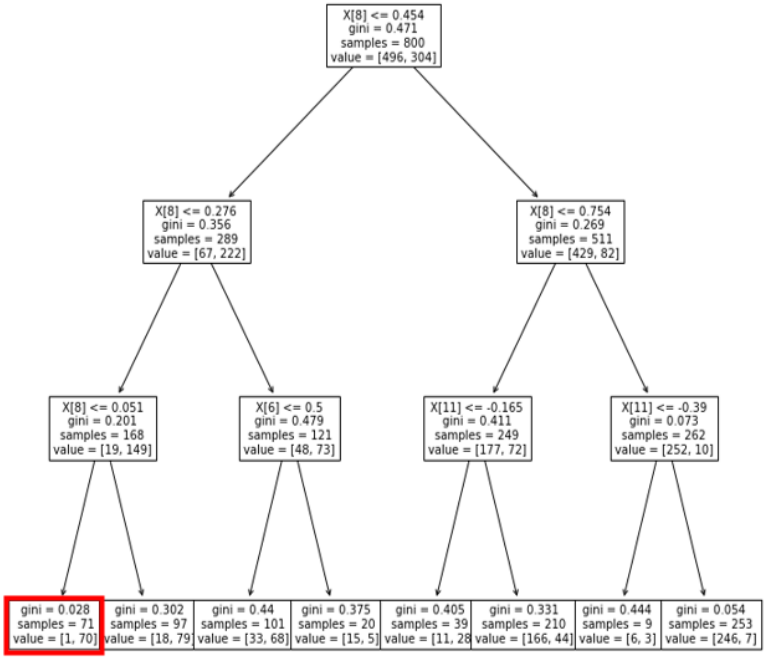
Question 9

ii) Fit a tree to the training data, with Purchase as the response and the other variables as predictors. What is the training error rate? (Set the max_depth parameter to 3 in order to get an interpretable plot)

- Training error rate (unpruned): 0.010000000000000009
- Training error rate (with max_depth parameter as 3): 0.15249999999999997

iii) Create a plot of the tree. How many terminal nodes does the tree have? Pick one of the terminal nodes, and interpret the information displayed

- The terminal node marked in red is a result from the split criterion X[8] <= 0.051 which is LoyalCH <= 0.051 and since it is on the left side it means that these samples have values of LoyalCH <= 0.051. There are 71 samples in this terminal node. The value shows how the samples to test for information gain are split up. So the 71 samples are divided into 1 and 70.
- The Gini index or Gini impurity measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen. Since 0 denotes that all elements belong to a certain class, and 1 denotes that the elements are randomly distributed across various classes. Since the Gini Index is 0.028 it denotes that there is a high probability that all elements belong to a certain class.

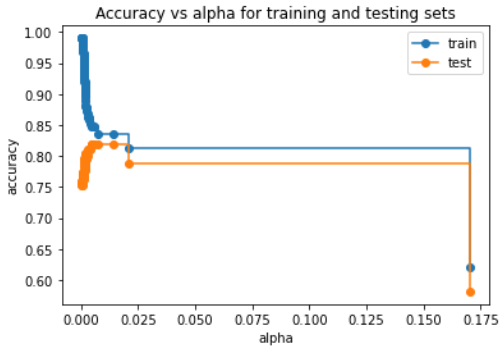


iv) Predict the response on the test data, and produce a confusion matrix comparing the test labels to the predicted test labels. What is the test error rate?

- Test error rate (unpruned): 0.2407407407407407
- Test error rate (with max_depth parameter as 3): 0.18888888888888888

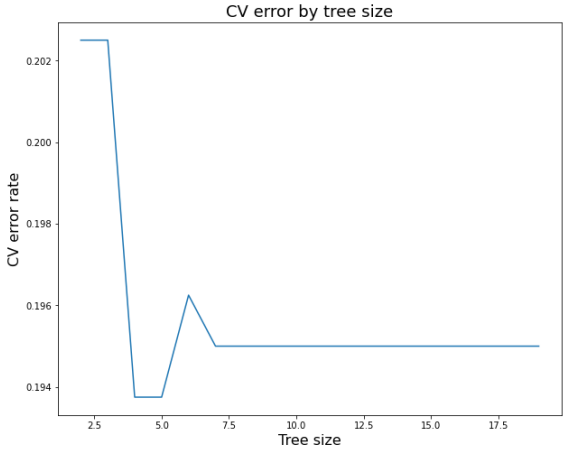
v) Determine the optimal tree size by tuning the ccp_alpha argument in scikit-learn's DecisionTreeClassifier. 1 You can use GridSearchCV for this purpose.

- The alpha that gives the highest test accuracy are: 0.00480852, 0.00481055, 0.00581514, 0.00747714, 0.0141431
- All 5 values give equal test accuracies so I chose 0.00747714 to choose the optimal tree size.

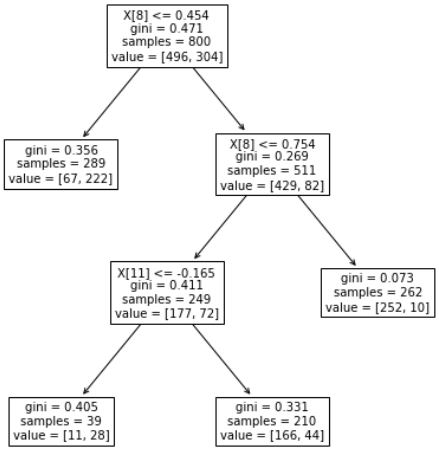


vi) Produce a plot with tree size on the x-axis and cross-validated classification error rate on the y-axis calculated using the method in the previous question. Which tree size corresponds to the lowest cross-validated classification error rate?

- Tree size with lowest CV error rate: 4



vii) Produce a pruned tree corresponding to the optimal tree size obtained using cross validation. If cross-validation does not lead to selection of a pruned tree, then create a pruned tree with five terminal nodes.



viii) Compare the training error rates between the pruned and unpruned trees. Which is higher?

- Unpruned: 0.010000000000000009
- Pruned: 0.16500000000000004
- The training error rate of the pruned tree is higher

ix) Compare the test error rates between the pruned and unpruned trees. Which is higher?

- Unpruned: 0.24444444444444446
- Pruned: 0.18148148148148147
- The test error rate of the unpruned tree is higher