CS4342 Assignment #5

Conceptual and Theoretical Questions

1. Suppose we fit a curve with basis functions $b_1(X) = X, b_2(X) = (X-1)^2 I(X \geq 1)$. (Note that I(X ≥ 1) equals 1 for X ≥ 1 and 0 otherwise.) We fit the linear regression model

$$Y = \beta_0 + \beta_1 b_1(X) + \beta_2 b_2(X) + \epsilon,$$

and obtain coefficient estimates $\hat{\beta}_0 = 1, \hat{\beta}_1 = 1, \hat{\beta}_2 = -2$. Sketch the estimated curve between X = −2 and X = 2. Note the intercepts, slopes, and other relevant information.

Points: 5

2. Suppose we fit a curve with basis functions $b_1(X) = I(0 \leq X \leq 2) - (X-1)I(1 \leq X \leq 2), b_2(X) = (X-3)I(3 \leq X \leq 4) + I(4 < X \leq 5)$. We fit the linear regression model

$$Y = \beta_0 + \beta_1 b_1(X) + \beta_2 b_2(X) + \epsilon,$$

and obtain coefficient estimates $\hat{\beta}_0 = 1, \hat{\beta}_1 = 1, \hat{\beta}_2 = 3$. Sketch the estimated curve between X = −2 and X = 2. Note the intercepts, slopes, and other relevant information.

Points: 5

3. Draw an example (of your own invention) of a partition of two dimensional feature space that could result from recursive binary splitting. Your example should contain at least six regions. Draw a decision tree corresponding to this partition. Be sure to label all aspects of your figures, including the regions R1, R2, . . ., the cutpoints t1, t2, . . ., and so forth.

Points: 5

4. It is mentioned in Section 8.2.3 that boosting using depth-one trees (or stumps) leads to an additive model: that is, a model of the form

$$f(X) = \sum_{j=1}^{p} f_j(X_j),$$

Explain why this is the case. You can begin with (8.12) in Algorithm 8.2.

Points: 5

5. This question relates to the plots in the below figures.

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

(b) Create a diagram similar to the left-hand panel of the figure, 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.

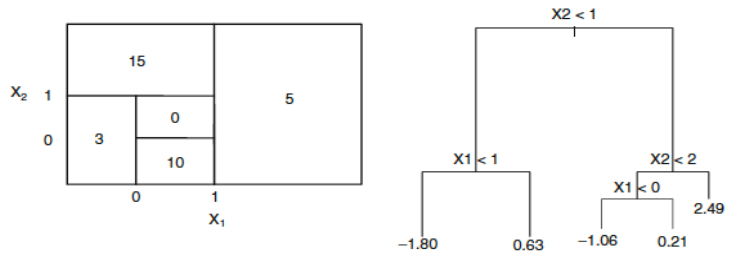

FIGURE 8.12. Left: *A partition of the predictor space corresponding to Exercise 4a.* Right: *A tree corresponding to Exercise 4b.*

Points: 5

6. Suppose we produce ten bootstrapped samples from a data set containing red and green classes. We then apply a classification tree to each bootstrapped sample and, for a specific value of X, produce 10 estimates of P(Class is Red|X):

0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.6, 0.65, 0.7, and 0.75.

There are two common ways to combine these results together into a single class prediction. One is the majority vote approach discussed in this chapter.

The second approach is to classify based on the average probability. In this example, what is the final classification under each of these two approaches?

Points: 5

Applied Questions

1. In this exercise, we will further analyze the Wage data set.

(a) Perform polynomial regression to predict wage using age. Use cross-validation to select the optimal degree d for the polynomial. Make a plot of the resulting polynomial fit to the data.

(b) Fit a step function to predict wage using age, and perform cross-validation to choose the optimal number of cuts. Make a plot of the fit obtained.

Hints:

- Check https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html

Points: 15

2. The Wage data set contains a number of other features not explored in Chapter 7, such as marital status (maritl), job class (jobclass), and others. Explore the relationships between some of these other predictors and wage, and use non-linear fitting techniques in order to fit flexible models to the data. Create plots of the results obtained, and write a summary of your findings.

Points: 10

3. This question uses the variables dis (the weighted mean of distances to five Boston employment centers) and nox (nitrogen oxides concentration in parts per 10 million) from the Boston data. We will treat dis as the predictor and nox as the response.

(a) Fit a cubic polynomial regression to predict nox using dis. Report the regression output, and plot the resulting data and polynomial fits.

(b) Plot the polynomial fits for a range of different polynomial degrees (say, from 1 to 10), and report the associated residual sum of squares.

(c) Perform cross-validation or another approach to select the optimal degree for the polynomial, and explain your results.

(d) Fit a regression spline to predict nox using dis. Report the output for the fit using four degrees of freedom. How did you choose the knots? Plot the resulting fit.

(e) Now fit a regression spline for a range of degrees of freedom, and plot the resulting fits and report the resulting RSS. Describe the results obtained.

(f) Perform cross-validation or another approach in order to select the best degrees of freedom for a regression spline on this data. Describe your results.

Hints:

- Check this https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html
- Check https://www.analyticsvidhya.com/blog/2018/03/introduction-regression-splines-python-codes/

Points: 10

4. Apply random forests to predict mdev of the Boston data after converting it into a qualitative response variable – values above the median of mdev is set 1 and others are set to zero. Use all other predictors in preditction of the qualitative data using 25 and 500 trees. Create a plot displaying the test error resulting from random forests on this data set for a more comprehensive range of values of number of predictors and trees. Describe the results obtained.

Hints:

- Check https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
- Check https://scikit-learn.org/stable/modules/ensemble.html

Points: 10

5. We want to predict Sales in the Carseats data set using regression trees and related approaches.

(a) Split the data set into a training set and a test set.

(b) Fit a regression tree to the training set. Plot the tree, and interpret the results. What test MSE do you obtain?

(c) Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test MSE?

(d) Use the bagging approach in order to analyze this data. What test MSE do you obtain? Determine which variables are most important (variable impooratnce measure).

(e) Use random forests to analyze this data. What test MSE do you obtain? Determine which variables aremost important (variable importance measure). Describe the effect of m, the number of variables considered at each split, on the error rate obtained.

Hints:

- Check https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
- Chec https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html#sklearn.tree.DecisionTreeRegressor
- Check https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingRegressor.html
- Check https://scikit-learn.org/stable/modules/ensemble.html
- Check https://machinelearningmastery.com/calculate-feature-importance-with-python/

Points: 15

6. We now use boosting to predict Salary in the Hitters data set.

(a) Remove the observations for whom the salary information is unknown, and then log-transform the salaries.

(b) Create a training set consisting of the first 200 observations, and a test set consisting of the remaining observations.

(c) Perform boosting on the training set with 1,000 trees for a range of values of the shrinkage parameter λ. Produce a plot with different shrinkage values on the x-axis and the corresponding training set MSE on the y-axis.

(d) Produce a plot with different shrinkage values on the x-axis and the corresponding test set MSE on the y-axis.

(e) Compare the test MSE of boosting to the test MSE that results from applying two of the regression approaches seen in Chapters 3 and 6.

(f) Which variables appear to be the most important predictors in the boosted model?

(g) Now apply bagging to the training set. What is the test set MSE for this approach?

Hints:

- Check https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html

Points: 10