Assignment #2 — Due Friday, Nov. 17th

*Submit your homework on Canvas.

*No late homework will be accepted for credit.

*Append the codes you used to your submission.

Problem 1: Basics Knowledge (200 pts)

1. Prove explicitly that how fitting a cubic spline regression can be solved by linear regression with equality constraints. Suppose the given data are $(x_1, y_1), \cdots, (x_n, y_n)$ and the knots are $(\xi_1, \cdots, \xi_K)$.
2. Derive the basis functions of natural cubic spline from the truncated power basis functions $1, x, x^2, x^3, (x - \xi_1)_+^3, \cdots, (x - \xi_K)_+^3$ of cubic spline.
3. Prove that the effective degree of freedom in smoothing splines: $\sum_{i=1}^n \frac{\partial \hat{y}_i}{\partial y_i} = \text{trace}(\mathbf{S}_\lambda)$.
4. What is the difference between *piecewise polynomial regression* and *local polynomial regression*?
5. Why does local polynomial regression has bad interpretability? Explain with details.
6. Does increasing the bandwidth h in polynomial regression will increase or decrease the bias? How about the variance? Explain with details.
7. Regression tree is most similar to which of the following methods:
   (a) K nearest neighbour regression
   (b) Piecewise constant regression
   (c) Logistic regression

   and explain the difference between regression tree and the method you choose.
8. When a tree gets finer, how does the bias and variance behave? Explain with details.
9. Generally speaking, which of the following methods can have the smallest bias?
   (a) Linear regression
   (b) Regression tree
   (c) Natural splines

   and explain why with details.
10. Write down the formal and explicit definition of variable importance measure for random forest using math notations. How can we do variable section/model selection in random forest?

Problem 2: Investigation of the Diameter, Height and Volume for Black Cherry Trees (100 pts)

Dataset: trees.csv
Description: This data set provides measurements of the diameter, height and volume of timber in 31 felled black cherry trees. Note that the diameter (in inches) is erroneously labelled Girth in the data. It is measured at 4 ft 6 in above the ground. The data frame contains 31 observations on 3 variables. The meaning of these variables are:
Girth: Tree diameter (rather than girth, actually) in inches
Height: Height in ft
Volume: Volume of timber in cubic ft

Goal: Find the relation between volume and other variables.

1. Fit four polynomial models (deg=1,2,3,4) to predict the Volume using Girth. Choose the model with the largest adjust R-squared. Plot the polynomial function of the model and also plot the confidence bands with ± 2 standard error. How about choosing the model using 5-CV error?

2. Use a polynomial logistic regression model with deg=2 to predict whether the Volume is larger or not than 30, using the variable Girth. Plot the function $P(\text{Volume} > 30)$ with respect to Girth and the confidence bands with ± 2 standard error.

3. Fit a regression spline with deg=2 to predict the Volume using the variable Girth at knots $10, 14, 18$. Plot the function and also the confidence bands with ± 2 standard error.

4. Fit a smoothing spline to predict the Volume using the variable Girth where the smoothing level is chosen by Cross-Validation. Plot the function. What is the used degrees of freedom?

5. Use both the variable Girth and Height to predict the Volume by a GAM where the individual function on Girth is a smoothing spline with df=4 and the function on Height is a smoothing spline with df=5. Plot the functions and also the confidence bands.

Problem 3: Audit Risk (100 pts)

Dataset: audit_train.csv and audit_test.csv
Description: The goal of the research is to help the auditors by building a classification model that can predict the fraudulent firm on the basis the present and historical risk factors. The information about the sectors and the counts of firms are listed respectively as Irrigation (114), Public Health (77), Buildings and Roads (82), Forest (70), Corporate (47), Animal Husbandry (95), Communication (1), Electrical (4), Land (5), Science and Technology (3), Tourism (1), Fisheries (41), Industries (37), Agriculture (200).

Many risk factors are examined from various areas like past records of audit office, audit-paras, environmental conditions reports, firm reputation summary, on-going issues report, profit-value records, loss-value records, follow-up reports etc. After in-depth interview with the auditors, important risk factors are evaluated and their probability of existence is calculated from the present and past records.

Goal: Use tree-based method to predict the risk, Risk=1 means fraudulent, Risk=0 means unfraudulent.

1. Use the train dataset to fit a classification tree and, plot the tree and report the training error. Test the performance on the test dataset and report the confusion matrix.

2. Use CV to prune the tree in Step 1 on the train dataset. Plot the train error versus the tree size. Plot the pruned tree which has the best train error. Report the test error.

3. Use random forest on the train dataset to build a classifier to predict the risk where setting m=13 and ntree=25. Report the training error.

4. Repeat Step 3 with four difference choices $m = 8, 12, 14, 16, 18$ and choose the one with smallest misclassification error on the train dataset. Test its performance on the test dataset.

5. Compare the above methods and report any findings you observe.

Problem 4: Implementing Regression Tree (100pts)

Write the source codes of regression tree with the following requirements:

(a) Instead of using residual sum of squares $\sum_{i \in R}(y_i - \hat{y})^2$ as the criterion, use the absolute loss $\sum_{i \in R}|y_i - \hat{y}|$ as the criterion of choosing the optimal split.

(b) Enables an option to choose the number of terminal nodes/leaves.

(b) Enables a stopping rule by a threshold v_{gain} of the minimum reduction of absolute loss, i.e., if the reduction of the absolute loss of an optimal split is smaller than v_{gain}, then ignore the split.

Apply your method to the Hitters training (*Hitters_train.csv*) data using the predictor variables *Years, Hits, RBI, Walks, PutOuts, Runs*. Tune your method and report the best test error (*Hitters_test.csv*) achieved.