05_Trees_Hitters_Task

May 5, 2025

1 Preliminary setup

[ ]:
```
import numpy as np
import pandas as pd
from ISLP import load_data
from matplotlib.pyplot import subplots, show
import matplotlib.pyplot as plt


# Load and preprocess data
Hitters = load_data('Hitters').dropna()
```

2 Task 1

1. Use the Hitters data and remove all rows that contain missing values. Create a new variable that is the log of Salary and provide histograms for Salary and Log(Salary). Interpret.

[ ]:

2. Split the sample into a training dataset consisting of the first 200 observations and a test dataset containing the remaining observations.

[ ]:

3. Fit a large, unpruned regression tree to predigt Log(Salary). Which features are used to construct the tree, which features are the most important and how many terminal nodes does the tree have? You might want to plot the tree for this exercise.

[ ]:

4. Compute the mean squared prediction error for the test data.

[ ]:

5. Let's try to improve predictions using k-fold CV. Set the seed to 2 and run 5-fold cross validation. Plot the mean squared cross validation error against the tree size and report the tree size and the pruning parameter that minimize the mean squared cross validation error.

[ ]:

6. Use the pruning parameter from the previous task to prune the tree. Plot the tree and report the most important variables.

[ ]:

7. Compute the test mean squared prediction error for pruned tree and compare to the results from Task 4.

[ ]:

8. Use random forest to improve the predictions. Fit 500 trees using $m = \sqrt(p)$ (round to the nearest integer).

[ ]:

9. Do you think it was necessary to fit 500 trees or would have fewer trees be sufficient? Determine the number of trees that provides the lowest OOB error.

[ ]:

10. Compute the OOB estimate of the out-of-sample error and compare it to best pruned model from CV of Task 5. Interpret the outcomes.

[ ]:

11. Which are the most important variables used in the random forest?

[ ]:

12. Let's try to improve the random forest by trying out different values for m. Set up a grid for m going from 1 to p. Write a loop that fits a random forest for each m. Explain which model you would choose.

[ ]:

13. For the best model, compute the test errors and compare them to the best pruned model from Task 7.

[ ]:

14. What is the OOB error obtained from bagging (you can infer the answer from the previous task).

[ ]: