Data Science and Social Inquiry: HW5

Yu-Chang Chen and Ming-Jen Lin

December 30, 2022

Question 1: Benjamini-Hochberg Procedure

In this problem, we will simulate data from $m = 1000$ fund managers. Generate the data using

```
np.random.seed(5)
np.random.normal(size=(n,m))
```

These data represent each fund manager's percentage returns for each of $n = 20$ months. We wish to test the null hypothesis that each manager's percentage returns have population mean equal to zero. To perform the hypothesis testing, one should conduct a one sample t-test and calculate the p-value for each manager.

(a) (1pt) If we control Type I error for each null hypothesis at level $\alpha = 0.05$, then how many null hypotheses do we reject.

Solution: By doing a one sample t test, we can calculate the p-value for each manager. We can find that there are 55 null hypotheses rejected. See HW5_G9.ipynb.

(b) (1pt) If we control the FDR at level 0.2 using Benjamini-Hochberg Procedure, how many null hypotheses do we reject?

Solution: By sorting the p-values, and by using Benjamini-Hochberg Procedure, we can find that there are 2 null hypotheses rejected. See HW5_G9.ipynb.

Question2: Tree

This is a coding question. Write your own code to solve the following questions, and do not use the scikitlearn package.

Given the data set

$$\begin{aligned}(Y, X) = \{&(-0.53, 0.33),\\ &(-1.07, 0.32),\\ &(2.16, 0.88),\\ &(-1.25, 0.01),\\ &(-2.32, 0.71)\}\end{aligned}$$

The mean square loss of a node m is defined as

$$\frac{1}{n_m} \sum_{i \in Q_m} (y_i - \bar{y}_m)^2,$$

where Q_m is the set of node m, and $n_m = |Q_m|$. The optimal regression tree is the tree that minimizes the mean of the mean square loss, weighted by each node's size.

(c) (2pt) Show all 25 possible regression trees with maximum depth 2 and calculate the weighted mean square loss for each tree. What Y would you predict given $X = 0.5$ using the optimal tree.

Solution: Between two data points, the splitting point can be at any point in the interval to result in the same tree. For simplicity, in our code, we use 5 data points as our cut points and obtain 25 possible regression trees. The optimal tree is the tree with depth 2, the first split is on $X \geq 0.71$, and the left node splits again on $X \geq 0.33$, and the right node splits again on $X \geq 0.88$.

Thus, if $X = 0.5$, there are two scenarios:
1. If the cut point c is $0.33 \leq c < 0.50$, then $Y = -2.32$.
2. If the cut point c is $0.50 \leq c < 0.71$, then $Y = -0.53$.

See HW5_G9.ipynb.

(d) (2pt) Which regression tree you found in the previous question is the recursive split optimum tree? What Y would you predict given $X = 0.5$ using this tree?

Solution: The recursive optimal tree is the tree with depth 2, the first split is on $X \geq 0.88$, and the left node splits again on $X \geq 0.71$, and the right node splits on $X \geq 0.88$.

Thus, if $X = 0.5$, there are two scenarios:
1. If the cut point c is $0.34 \leq c < 0.50$, then $Y = -2.32$.
2. If the cut point c is $0.50 \leq c < 0.71$, then $Y = -0.95$.

See HW5_G9.ipynb.

Given the data set

$$\begin{aligned}(Animal, X) = \{&(Cat, 0.33),\\ &(Cat, 0.32),\\ &(Dog, 0.88),\\ &(Dog, 0.01),\\ &(Cat, 0.71)\}\end{aligned}$$

If the target is a classification outcome taking on values $1, 2, \ldots, K$ (in this case $K = 2$), let

$$p_{mk} = \frac{1}{n_m} \sum_{y \in Q_m} I(y = k)$$

be the proportion of class k observations in node m. The Gini index of a given node m is then defined as

$$\sum_{k=1}^{K} p_{mk}(1 - p_{mk}),$$

which measures the impurity of the node. The optimal classification tree is the tree that minimizes the mean of the Gini index, weighted by each node's size.

(e) (2pt) Show all 25 possible classification trees with maximum depth 2 and calculate the weighted mean Gini index for each tree. Which animal would you predict given $X = 0.1$ using this optimal tree.

Solution:
There are four optimal trees that have the same mean of Gini Index. The first optimal tree is the tree with depth 2, the first split is on $X \geq 0.32$, and the left node splits again on $X \geq 0.01$, and the right node splits again on $X \geq 0.88$.
The second optimal tree is the tree with depth 2, the first split is on $X \geq 0.33$, and the left node splits again on $X \geq 0.32$, and the right node splits again on $X \geq 0.88$.

The third optimal tree is the tree with depth 2, the first split is on $X \geq 0.71$, and the left node splits again on $X \geq 0.32$, and the right node splits again on $X \geq 0.88$.
The last optimal tree is the tree with depth 2, the first split is on $X \geq 0.88$, and the left node splits again on $X \geq 0.32$, and the right node splits again on $X \geq 0.88$.

Thus, if $X = 0.1$, there are two scenarios:
1. If the cut point c is $0.01 \leq c < 0.1$, then $Animal = Cat$.
2. If the cut point c is $0.1 \leq c < 0.33$, then $Animal = Dog$.

See HW5_G9.ipynb.

(f) (2pt) Which classification tree you found in the previous question is the recursive split optimal tree? Which animal would you predict given $X = 0.1$ using this tree?

Solution: The recursive optimal tree is the tree with depth 2, the first split is on $X \geq 0.32$, and the left node splits again on $X \geq 0.01$, and the right node splits again on $X \geq 0.88$.

Thus, if $X = 0.1$, there are two scenarios:
1. If the cut point c is $0.01 \leq c < 0.1$, then $Animal = Cat$.
2. If the cut point c is $0.1 \leq c < 0.33$, then $Animal = Dog$.

See HW5_G9.ipynb.