Due Date: Friday, October 28, 2022, 11:59 PM

Abstract

In this assignment, you will learn about linear regression and classification. The questions below can be answered with the help of the book. Show all calculations done step-by-step. In problem 3, you need not worry about being "mathematically" correct. Please follow the same notation as in the book. You are allowed to work on teams of up to three people. If working as a team, every member of the team should submit the (same) report. All reports and supplemental material must be zipped as `team17_Assignment2.zip` and uploaded on Blackboard.

Problem 1. *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, 0.75. There are two common ways to combine these results together into a single class prediction. One is the majority vote approach. 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?*

Solution:
For average probability the P($Red|X$) = 0.45 and P($Green|X$) = 0.55 which is clearly greater than the threshold '0.5', hence final classification for this approach is 'Green'. Whereas in majority vote approach it's final classification is 'Red' because the number of estimates for class Red is 6 more when compared to class Green, 4.

Both the approaches produced two different results as their final classifications. The reason over here is that computing mean is sensitive to outliers in the data. As average probability approach uses this, though the number of estimates below the threshold are '4', yet they are close to '0' rather than to the threshold value, due to this mean value is less than the threshold resulting in a different output.

Link to notebook with code and explanation

Problem 2. *Provide a detailed explanation of the algorithm that is used to fit a regression tree.*

Solution:
Lets first look at the algorithm that's is used to fit a regression tree.

1. We go for a top to bottom approach and use a recursive binary splitting on the data.

2. We do this by recursively find the best single partitioning of the data where the reduction of residual sum of squares is highest. This will be a greedy approach.

3. We keep in repeating the above step to each of the split parts individually until we find some minimal number of observations present in each of the leaves.

4. Cost complexity pruning is then applied to this larger tree to obtain a sequence of optimal sub-trees.

For each value of α there corresponds α sub-tree T $\subset T_0$ such that

$$\sum_{m=i}^{|T|} \sum_{i:x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha|T|$$

Here $|T|$ represents no. of terminal nodes. The tuning parameter α controls a trade-off between the sub-tree's complexity and its fit to the training data.

Next we can use K-fold cross-validation to choose α.
1. For each value of K we evaluate the mean squared prediction as a function of α on the fold.
2. Average the results, and pick α to minimize the average error.
3. Taking the α selected in previous step, we return the tree calculated using the formula on the entire dataset with that chosen value of α

Problem 3. *In the general case, imagine that we have d binary features, and we want to count the number of features with value 1. How many leaf nodes would a decision tree need to represent this function? If we used a sum of decision stumps, how many terms would be needed?*

Solution:
a. 2^d leaf nodes, for one feature we have two leaf nodes, for two features we have four leaf nodes.....so on.

b. 'd' terms are required when sum of decision stumps is used. Sum of decision stumps uses one feature at a time, so if there are 'd' binary features then 'd' terms are required.

GPA	Studied	Passed
L	F	F
L	T	T
M	F	F
M	T	T
H	F	T
H	T	T

Problem 4. *Based on the table above (L: Low; M: Medium; H: High; F: False; T: True), what is the entropy H(Passed)?*

Solution:
Complete Set - S(Passed) - (4+, 2-) - Total 6

Entropy H(Passed) = $\frac{-4}{6} \log_2 \frac{4}{6}$ - $\frac{2}{6} \log_2 \frac{2}{6}$

Entropy H(Passed) = $\frac{-2}{3}$ $\log_2 \frac{2}{3}$ - $\frac{1}{3}$ $\log_2 \frac{1}{3}$

Entropy H(Passed) = -0.666666 * -0.58496250072 - 0.333333 * -1.5849625007

Entropy H(Passed) = 0.92

Problem 5. *What is the entropy H(Passed|GPA)? Hint. Give an entropy value for each conditioned value.*

Solution:
Values(GPA) = [L, M, H]

Entropy for each value,

S_L = (1+, 1-) = Entropy H($Passed|S_L$) = $\frac{-1}{2}$ $\log_2 \frac{1}{2}$ - $\frac{1}{2}$ $\log_2 \frac{1}{2}$ = $\log_2 2 = 1$

S_M = (1+, 1-) = Entropy H($Passed|S_M$) = $\frac{-1}{2}$ $\log_2 \frac{1}{2}$ - $\frac{1}{2}$ $\log_2 \frac{1}{2}$ = $\log_2 2 = 1$

S_H = (2+, 0-) = Entropy H($Passed|S_H$) = $\frac{-2}{2}$ $\log_2 \frac{2}{2}$ - $\frac{0}{2}$ $\log_2 \frac{0}{2}$ = 0

Entropy H($Passed|GPA$) = $\frac{2}{6}$ * 1 + $\frac{2}{6}$ * 1 + $\frac{2}{6}$ * 0

Entropy H($Passed|GPA$) = $\frac{1}{3}$ * 1 + $\frac{1}{3}$ * 1 + $\frac{1}{3}$ * 0

Entropy H($Passed|GPA$) = $\frac{2}{3}$ = 0.66

Gain(S, GPA) = Entropy H(Passed) - Entropy H($Passed|GPA$) = 0.92 - 0.66 = 0.26

References

[1] Textbook

[2] Reference to calculate entropy https://www.youtube.com/watch?v=coOTEc-0OGwtReference to calculate entropy

[3] Fit Decision Tree https://epurdom.github.io/Stat131A/book/regression-and-classification-trees.html

[4] Fit Decision Tree reference 2 for Problem 2 https://daviddalpiaz.github.io/stat430fa17/slides/isl/trees.p