IEOR 242: Applications in Data Analysis, Fall 2023

Homework Assignment #3

October 4, 2023

Problem 1: True/False and Multiple Choice Questions (10 points)

1. Which of these does not apply to pruning decision trees?
   (a) Can be a key part of a Random Forest model
   (b) Reduces chances of overfitting
   (c) Potentially improves interpretability. be a key part of a CART model
   (d) None of the above

2. If we use k-fold cross validation on a training set to select a final model, then there is no need to evaluate the performance of this model on a test set since it is impossible for this model to overfit the training set.
   (a) True
   (b) False

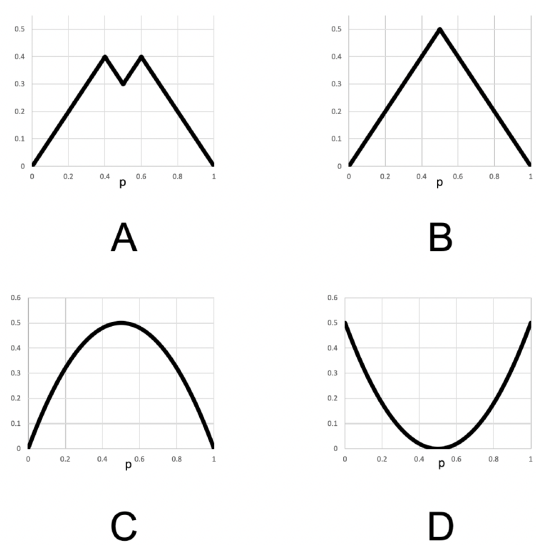
3. One of the main reasons that boosting is effective is because, at every iteration, the algorithm ends a new decision tree that is very large (i.e., its depth is very big).
   (a) True
   (b) False

4. Suppose we have 5 positive values and 1 negative value, what is the Gini impurity for this set?
   (a) 0.24
   (b) 0.28
   (c) 0.32
   (d) 0.65
   (e) None of the above

5. Consider the CART algorithm for binary classification. A desirable property of an impurity function for the CART algorithm is that a split never increases the total impurity cost of the tree (i.e., using the notation from class we have Δ 0). Figure 1 below depicts four potential impurity functions that might be used in CART, each as a function of the proportion p of observations with Y = 0 in the current bucket. Which of these impurity functions have the desirable property mentioned above?

   (a) All four
   (b) Only A, B, and C
   (c) Only B and C
   (d) Only C

Figure 1



Problem 2: (30 points)

Consider the algorithm for building a CART model **in the case of regression**. Following and expanding on the notation from class, suppose that our current tree, denoted by T_{old}, has $|T_{\text{old}}| = M$ terminal nodes/buckets. For each bucket $m = 1, \ldots, M$, let:

1. N_m denote the number of observations in bucket m ,
2. $Q_m(T_{\text{old}})$ denote the value of the impurity function at bucket m , and
3. R_m denote the region in the feature space corresponding to bucket m .

Also let N be the overall total number of observations. Recall that, in the case of regression we have that:

$$Q_m(T_{\text{old}}) = \frac{1}{N_m} \sum_{i:x_i \in R_m} (y_i - \hat{y}_m)^2 \ ,$$

where $\hat{y}_m = \frac{1}{N_m} \sum_{i:x_i \in R_m} y_i$ is the mean response in bucket m.

Then the total impurity cost of the tree T_{old} is defined as:

$$C_{\text{imp}}(T_{\text{old}}) = \sum_{m=1}^{M} N_m Q_m(T_{\text{old}}) \ .$$

Consider a potential split at the final bucket M (we're using M just for ease of notation), which results in a new tree T_{new}. This new tree has $|T_{\text{new}}| = M + 1$ terminal nodes/buckets, and for this new tree we let

1. $\tilde{N}_m$ denote the number of observations in bucket m ,
2. $\tilde{Q}_m(T_{\text{new}})$ denote the value of the impurity function at bucket m , and
3. $\tilde{R}_m$ denote the region in the feature space corresponding to bucket m .

The total impurity cost of the tree T_{new} is defined analogously as:

$$C_{\text{imp}}(T_{\text{new}}) = \sum_{m=1}^{M+1} \tilde{N}_m \tilde{Q}_m(T_{\text{new}}) \ .$$

Please answer the following:

a) (10 points) Let $\Delta = C_{\text{imp}}(T_{\text{old}}) - C_{\text{imp}}(T_{\text{new}})$ be the absolute decrease in total impurity resulting from the split. Let $\tilde{R}_M$ and $\tilde{R}_{M+1}$ denote the newly created region in T_{new}. Please write the explicit expression of Δ, consisting of data points (x_i, y_i) in regions R_M, $\tilde{R}_M$ and $\tilde{R}_{M+1}$.

b) (10 points) Show that $\Delta \geq 0$. *(Hint: you can use the fact that, given a sequence of real numbers $z_1, z_2, \ldots, z_n$, we have $\bar{z} = \arg\min_z \sum_{i=1}^{n}(z_i - z)^2$, where $\bar{z}$ is defined as $\frac{1}{n}\sum_{i=1}^{n} z_i$)*

c) (10 points) Let R^2_{old} be the training set R^2 value for the model defined by T_{old}. Let R^2_{new} be the training set R^2 value for the model defined by T_{new}. Let $\text{SST} = \sum_{i=1}^{N}(y_i - \bar{y})^2$, where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

For a given value of the complexity parameter (CP) $\alpha \geq 0$, recall that we have the *modified cost function*

$$C_\alpha(T) = C_{\text{imp}}(T) \ + \ \alpha \cdot \text{SST} \cdot |T|$$

Show that $C_\alpha(T_{\text{new}}) \leq C_\alpha(T_{\text{old}})$ if and only if $R^2_{\text{new}} - R^2_{\text{old}} \geq \alpha$.