Ch. 8 Trees - Problem Bank Questions

October 27, 2020

1. Trees true and false. For each, explain your answer.
(a) T/F: Trees consider every possible partition of the feature space.
(b) T/F: Let X_j be some feature in your data set and suppose you wish to build a regression tree. At each split in the tree, we define the pair of half-planes

$$R_1(j,s) = \{X|X_j < s\} \text{ and } R_2(j,s) = \{X|X_j \geq s\},$$

and we want to find the value of j and s such that the following is minimized:

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2,$$

where $\hat{y}_{R_k}$ is the mean response for the training observations in region k.
(c) T/F: Binary decision trees partition the feature space.
(d) T/F: For regression trees, each leaf node predicts the sum of data points that are in that region of feature space as the prediction for the response variable.
(e) T/F: Deep trees with many splits tend to have low bias.
(f) T/F: Cost complexity pruning considers every possible subtree to decide which one to select.

2. Classification Trees: Let $\hat{p}_{mk}$ represent the proportion of training examples in the m^{th} region that are from the k^{th} class. Two measures of node purity are the Gini index,

$$G = \sum_{k=1}^{K} \hat{p}_{mk} \left(1 - \hat{p}_{mk}\right),$$

and the cross-entropy,

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}.$$

(a) Explain why the Gini index is a measure of the total variance across all K classes.

(b) Show that the Gini index takes on a smaller value the more "pure" the tree is splitting each class; that is, the more examples that are assigned to the same node and are in the same class, the smaller the Gini index.

(c) What is the largest possible value for the Gini index? What does this correspond to in terms of the behavior of the tree?
(d) Show that the cross-entropy also takes on a smaller value the more pure a node is.
(e) What is the maximum value of the cross-entropy?

(f) Plot the cross entropy and the Gini index as a function of $\hat{p}_{mk}$ for a fixed k.

3. Trees: For the true and false, explain your answer.
(a) T/F: Regression trees assume a model of the form

$$f(X) = \sum_{m=1}^{M} c_m \mathbf{1}_{X \in R_m},$$

where $R_1, \ldots, R_m$ is a partition of the feature space.
(b) T/F: ~~Regression trees can fit linear models.~~ 
(c) T/F: A disadvantage of tree models is that they are hard to explain and interpret.
(d) T/F: Trees cannot handle qualitative predictor variables without special coding.
(e) T/F: Trees are very robust in general.

4. Bagging, Random Forests and Boosting.

(a) What is the goal of bagging for trees?
(b) What is the effect of averaging a set of independent observations on the variance? Include a numerical example.
(c) Explain conceptually what the bootstrap does and describe a basic procedure for bootstrapping in the case of trees.
(d) How are random forests different from bagging? What improvement do random forests offer?
(e) What is a case in which bagging will not lead to a very substantial reduction in variance?
(f) How does boosting differ from bagging for regression trees?

5. Boosting: For the true and false, explain your answer.
(a) What are the three tuning parameters for boosting? Explain what each one does.
(b) T/F: The number of splits in each tree, d, controls the interaction depth.
(c) T/F: Boosting fits a tree sequentially to the residuals of the previous trees.
(d) T/F: Smaller individual trees don't work well for boosting.
(e) T/F: Using all stumps, $d = 1$, leads to an additive model.

6. Suppose we wish to predict at a new point $\boldsymbol{x}_{\text{new}}$. Let $\hat{f}^b(\boldsymbol{x}_{\text{new}})$ be the predictor for the b-th individual tree in a random forest (RF). The RF predictor can then be written as:

$$\hat{f}(\boldsymbol{x}_{\text{new}}) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(\boldsymbol{x}_{\text{new}})$$

(a) Suppose the predictor has constant variance over all trees, i.e., $\text{Var}\{\hat{f}^b(\boldsymbol{x}_{\text{new}})\} = \sigma^2$ for $b = 1, \cdots, B$. If the pairwise correlation between any two treed predictions is $\rho \geq 0$, show that the RF predictor has variance:

$$\text{Var}\{\hat{f}(\boldsymbol{x}_{\text{new}})\} = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2.$$

(b) Suppose $\rho > 0$. What happens to the RF variance $\text{Var}\{\hat{f}(\boldsymbol{x}_{\text{new}})\}$ as the number of trees B grows large (i.e., $B \to \infty$)?