15.095, Practice Exam Answer

November 2, 2018

1 Problem 1

(a) Lasso always produces sparse solutions. **False**. *Lasso produces partially sparse solutions, but it does not recover the true sparse solution.*

(b) Lasso always produce robust solutions. **True**. *Lasso is equivalent to a robust optimization problem.*

(c) The random forests method has no tuning parameters. **False**. *Random forest includes tuning parameters such as minbucket and number of trees.*

(d) Optimal Trees can greatly improve by applying the Random Forests methodology to create a forest of Optimal Trees. **False**. *The locally optimal trees found by the local search algorithm are too similar to each other, so it is not useful to average the predictions from multiple optimal trees.*

(e) The problem $\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_1 + \rho\|\boldsymbol{\beta}\|_1$ can be written as a linear optimization problem. **True**. *We can linearize the 1-norm penalty terms in the objective function by adding in auxiliary variables and linear constraints.*

(f) The training AUC in logistic regression is always at least 1/2. **True**. *A baseline model achieves training AUC = 0.5, and logistic regression finds a model with as good or better performance on the training set.*

(g) Consider a regression problem where our input data is one-dimensional. When normalizing the data, first we subtract the mean of this vector, then we divide by the standard deviation of this vector. In the end, we split them into training, validation, and test sets. **False**. *We normalize all data in the same way by using the mean and standard deviation of the training data.*

(h) R^2 measures how accurate our model predictions are in comparison to a baseline model which predicts the mean of y from the training set.

(i) We want to use mixed integer optimization to ensure that the coefficients in linear regression are significant. Please write down the constraints that achieve this. **See lecture 4 page 10**.

2 Problem 2

Assume George likes to order z. First, based on the decision tree and the value of s and p, we know the $s = 4, p = 8$ belongs to leaf A. Let $L_A = \{1, 3, 8\}$ denote the indices in leaf A. Now we use a prescriptive method to decide z. The optimization problem we want to solve is

$$\begin{aligned}
&\max_z \frac{1}{|L_A|} \sum_{j \in L_A} \Big(10 \min(y_j, z) - 6z\Big) \\
&= \max_z \frac{1}{3}(10\min(70, z) - 6z) + \frac{1}{3}(10\min(80, z) - 6z) + \frac{1}{3}(10\min(90, z) - 6z) \\
&= \max_z \frac{10}{3}\Big(\min(70, z) + \min(80, z) + \min(90, z)\Big) - 6z.
\end{aligned}$$

Case 1: $z \leq 70$

$$\frac{10}{3}(3z) - 6z = 4z \leq 4(70) = 280.$$

Case 2: $70 < z \leq 80$

$$\frac{10}{3}(70 + 2z) - 6z = \frac{700}{3} + \frac{2}{3}z \leq \frac{700}{3} + \frac{2}{3}(80) = \frac{860}{3} \approx 287.$$

Case 3: $80 < z \leq 90$

$$\frac{10}{3}(70 + 80 + z) - 6z = 500 - \frac{8}{3}z \leq 500 - \frac{8}{3}(80) = \frac{860}{3} \approx 287.$$

Case 4: $z \geq 90$

$$\frac{10}{3}(70 + 80 + 90) - 6z = 800 - 6z \leq 800 - 6(90) = 260.$$

So we see that the optimal solution is $z = 80$, with optimal cost $\approx \$287,000$.

3 Problem 3

This problem asks you to model an easy version of regression trees using MIO. The problem can be formulated as follows.

Part a:

$$\begin{aligned}
\min_{\boldsymbol{\beta},\mathbf{a},b} \quad & \sum_{i=1}^{n} L_i & & \\
\text{subject to} \quad & L_i \geq f_i - y_i & & \forall i, \\
& L_i \geq y_i - f_i & & \forall i, \\
& f_i - \boldsymbol{\beta}_t^T \mathbf{x}_i \leq M(1 - z_{it}) & & \forall i, t, \\
& f_i - \boldsymbol{\beta}_t^T \mathbf{x}_i \geq -M(1 - z_{it}) & & \forall i, t, \\
& \mathbf{a}^T x_i \leq b + M(1 - z_{it}) & & \forall i, t \\
& \mathbf{a}^T x_i \geq b - M(1 - z_{it}) & & \forall i, t \\
& \sum_{t=1}^{2} z_{it} = 1 & & \forall i.
\end{aligned} \tag{1}$$

Part b:

$$
\begin{aligned}
\min_{\beta,\mathbf{a},b} \quad & \sum_{i=1}^{n} L_i \\
\text{subject to} \quad & L_i \geq f_i - y_i, && i \in [n] \\
& L_i \geq y_i - f_i, && i \in [n] \\
& f_i - \beta_t^T x_i \geq -M(1 - z_{it}), && i \in [n], t \in \Gamma_L \\
& f_i - \beta_t^T x_i \leq M(1 - z_{it}), && i \in [n], t \in \Gamma_L \\
& \mathbf{a}_m^T \mathbf{x}_i + \epsilon \leq b_m + M(1 - z_{it}), && t \in \Gamma_L, i \in [n], m \in L(t) \\
& \mathbf{a}_m^T \mathbf{x}_i \geq b_m - M(1 - z_{it}), && t \in \Gamma_L, i \in [n], m \in R(t) \\
& \sum_{t \in \Gamma_L} z_{it} = 1. && i \in [n]
\end{aligned}
\tag{2}
$$

where Γ_L is the set of leaf nodes, $L(t), R(t)$ represents the left branch and right branch ancestor of leaf node t.

Read Chapter 11 (page 222-225) in the book to get more details about this problem.