15.095, Practice Exam

October 2018

Instructions:

1. You have 90 minutes to complete the examination.
2. This exam is closed book and closed notes.
3. Please explain your work carefully for the short answer and proof questions.
4. Good luck!

Problem 1

When a question is true or false please give 2–3 sentences of explanation.

(a) Lasso always produces sparse solutions. True or False?

(b) Lasso always produce robust solutions. True or False?

(c) The random forests method has no tuning parameters. True or False?

(d) Optimal Trees can greatly improve by applying the Random Forests methodology to create a forest of Optimal Trees. True or False?

(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 of False?

(f) The training AUC in logistic regression is always at least 1/2. True or False?

(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 sest. True or false?

(h) What is an intuitive explanation of R^2?

(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.

Problem 2

George, a member of the Analytics department, has developed an optimal tree model to predict demand Y for a new clothing product. The model is based on two covariates—s, a measure of previous sales of shirts, and p, a measure of previous sales of pants—and is as follows:

- If $s \leq 5$ and $p \leq 10$, the model predicts that $Y = 80$. We call this left A.
- If $s \leq 5$ and $p > 10$, the model predicts that $Y = 100$. We call this leaf B.
- If $s > 5$, the model predicts that $Y = 60$. We call this leaf C.

There are 10 points (s_i, p_i, y_i) in the training set that are distributed as follows:

- $y_1 = 70$, $y_3 = 90$, and $y_8 = 80$ go to leaf A.
- $y_2 = 120$, $y_r = 80$, $y_5 = 100$, and $y_6 = 100$ go to leaf B.
- $y_7 = 50$, $y_9 = 70$, and $y_{10} = 60$ go to leaf C.

Suppose that George would like to use this tree to decide how much to order when $s = 4$ and $p = 8$. In this case, how much should George order if each unit of item sold brings \$10,000 of revenue and costs \$6,000 to order?

Problem 3 (20 points)
Given data $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, $y_i \in \mathcal{R}$ and $\mathbf{x}_i \in \mathcal{R}^d$, we want to partition $\mathcal{R}^d$ into k polyhedral sets P_j. With every partition we associate $\boldsymbol{\beta}_j \in \mathcal{R}^d$, such that we minimize the least absolute deviation error. For $k = 1$ this is the classical regression problem with absolute deviation criterion:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} |y_i - \mathbf{x}_i^T \boldsymbol{\beta}|.$$

Given a partition $(P_1, \ldots, P_k)$, we let $F_j = \{i : \ \mathbf{x}_i \in P_j\}$. The overall objective is:

$$\min_{P_1,\ldots,P_k} \min_{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_k} \sum_{j=1}^{k} \sum_{i \in F_j} |y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j|.$$

(a) (10 points) Suppose $k = 2$, $P_1 = \{\mathbf{x} : \mathbf{a^T x} \leq b\}$ and $P_2 = \{\mathbf{x} : \mathbf{a^T x} \geq b\}$ with $\mathbf{a} \in [-1, 1]^d$ and $b \in [-1, 1]$. Formulate the problem as a linear mixed integer optimization problem. The formulation should identify both the vector $\mathbf{a}$, the scalar b and the regression coefficients $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$.

(b) (10 points) Consider the two hyperplanes $\mathbf{a}_i^T \mathbf{x} = b_i$, $i = 1, 2$ with $\mathbf{a}_i \in [-1, 1]^d$ and $b_i \in [-1, 1]$ They define four polyhedral sets:

$$P_1 = \{\mathbf{x} : \ \mathbf{a}_1^T \mathbf{x} \leq b_1, \ \mathbf{a}_2^T \mathbf{x} \leq b_2\},$$

$$P_2 = \{\mathbf{x} : \ \mathbf{a}_1^T \mathbf{x} \leq b_1, \ \mathbf{a}_2^T \mathbf{x} \geq b_2\},$$

$$P_3 = \{\mathbf{x} : \ \mathbf{a}_1^T \mathbf{x} \geq b_1, \ \mathbf{a}_2^T \mathbf{x} \leq b_2\},$$

$$P_4 = \{\mathbf{x} : \ \mathbf{a}_1^T \mathbf{x} \geq b_1, \ \mathbf{a}_2^T \mathbf{x} \geq b_2\}.$$

Formulate the problem as a linear mixed integer optimization problem. The formulation should identify both the vectors $\mathbf{a}_1, \ \mathbf{a}_2$, scalars $b_1, \ b_2$ as well as the regression coefficients $\boldsymbol{\beta}_i$, $i = 1, \ldots, 4$.