15.095: Machine Learning Under a Modern Optimization Lens **Fall 2020**

Midterm Examination, October, 21, 2020

Problem 1, 30 points, 3 points per question

Indicate whether each statement listed below is True or False. In either case, write one sentence explaining your answer.

(a) To achieve robustness in any machine learning problem (linear regression, logistic regression, etc), we only need to add a regularization term on the coefficients β, e.g. $\|\beta\|_1, \|\beta\|_2$ to the objective function directly.

(b) It is known that the objective function $c(\mathbf{s})$ in the dual formulation of sparse linear regression has two equivalent forms, namely

$$c(\mathbf{s}) = \frac{1}{2}\mathbf{y}^T \left(\mathbf{I}_n + \gamma \sum_i s_i \mathbf{X}_i \mathbf{X}_i^T \right)^{-1} \mathbf{y} \tag{1}$$

and

$$c(\mathbf{s}) = \frac{1}{2}\mathbf{y}^T \left(\mathbf{I}_n - \mathbf{X}_s \left(\frac{\mathbf{I}_k}{\gamma} + \mathbf{X}_s^T \mathbf{X}_s \right)^{-1} \mathbf{X}_s^T \right) \mathbf{y} \tag{2}$$

where n is the sample size and k is the sparsity. Then for the cutting-plane method, (1) is always more powerful than (2).

(c) The out-of-sample R^2 error for a linear model is always non-negative.

(d) The time necessary to solve the sparse regression problem using the cutting-plane method always increases with the sample size n.

(e) Changing one data point in the data may change the median regression estimate.

(f) Feedforward neural networks with activation function $\phi(x) = \max(x, 0)$ and optimal classification trees with hyperplanes are equivalent in terms of modeling power.

(g) Stable linear regression reduces to a linear optimization problem.

(h) Optimal classification trees with hyperplanes (OCT-Hs) with sparsity one are equivalent to optimal classification trees (OCTs).

(i) In classification problems, as we traverse from left to right the ROC curve, specificity decreases.

(k) Lasso leads to sparse models.

Problem 2, 25 Points

In this question, we study some aspects of the sparse linear regression problem. As usual, we are given data $(\boldsymbol{x_i}, y_i),\ \boldsymbol{x_i} \in \mathbb{R}^p,\ y_i \in \mathbb{R},\ i = 1, \ldots, n$, and our goal is to estimate

$$\boldsymbol{\beta^\star} \quad = \quad \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \quad \|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + \Gamma\|\boldsymbol{\beta}\|^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_0 \leq k. \tag{3}$$

1. (10 Points) Assume $k = 1$. Describe a polynomial in n, p, k algorithm that solves the sparse regression problem. Assuming that the features have been scaled so that $\|\boldsymbol{X_j}\|^2 = 1,\ \forall j$, interpret the solution in terms of the problem data.

2. (5 Points) Does the polynomial-time approach from Part 1 extend for $k > 1$? If the answer is yes, elaborate. If the answer is no, give the complexity of the approach in terms of n, p, k.

3. (10 Points) By introducing binary variables $\boldsymbol{z} \in \{0, 1\}^p$ encoding the support of $\boldsymbol{\beta}$, i.e., $\beta_j \neq 0 \Rightarrow z_j = 1,\ j = 1, \ldots, p$, we were able to reformulate the sparse regression problem as

$$\min_{\boldsymbol{z} \in \{0,1\}^p} \quad c(\boldsymbol{z}) \quad \text{s.t.} \quad \boldsymbol{e}^T\boldsymbol{z} \leq k. \tag{4}$$

   You are given functions $c : \{0, 1\}^p \to \mathbb{R}$ and $\nabla_{\boldsymbol{z}} c : \{0, 1\}^p \to \mathbb{R}^p$, which, on input $\boldsymbol{z} \in \{0, 1\}^p$, compute the associated loss $c(\boldsymbol{z})$ and its gradient with respect to $\boldsymbol{z}$, respectively. (*Note: You can use those functions as black-boxes in your answer.*)

   Describe a **complete cutting planes algorithm** for an **extended sparse regression** problem, whereby features whose pairwise correlation exceeds $\rho_{\max}$ are not allowed to be selected simultaneously. How would you obtain a initial point $\boldsymbol{z_0}$ for the extended cutting planes method?

Problem 3, 25 points

In this question, we study the piece-wise linear regression problem

$$\min_{\mathbf{a}_j} \quad \sum_{i=1}^{n} |y_i - \max_{j=1,\ldots,k} \{\mathbf{a}_j^T \mathbf{x}_i\}| \tag{5}$$

given input data $\mathbf{x}_i \in \mathbb{R}^p, y_i \in \mathbb{R}, i = 1, \ldots, n$.

1. (13 points) Given k, formulate Problem (5) as a **mixed-integer linear** optimization problem.

2. (12 points) Suppose now we treat k as a decision variable and control the complexity of the model by adding a regularization term to (5). That is we solve the following problem

$$\begin{aligned} \min_{\mathbf{a}_j, k} \quad & \sum_{i=1}^{n} |y_i - \max_{j=1,\ldots,k} \{\mathbf{a}_j^T \mathbf{x}_i\}| + \lambda k \\ \text{s.t.} \quad & 1 \leq k \leq k_{\max}, \end{aligned} \tag{6}$$

   for given $\lambda > 0$ and integer $k_{\max}$. Notice that (6) is nonlinear in k. Formulate (6) as a **mixed-integer linear** optimization problem.

Problem 4, 20 Points

People who are sick with Disease X are treated with Drug Y, with dosages measured in mg. Each individual patient i has some threshold t_i where for every mg of the drug they receive up to that amount, their life expectancy increases by C days due to benefits from the drug, but for every mg above the threshold their life expectancy decreases by D days due to toxicity from the drug. They never receive more than 8 mg of the drug.

1. (10 Points) We are given historical data $(X_i; y_i)$ from patients $i = 1, \ldots, n$, where X_i is data for patient i and y_i is the empirically measured threshold for the patient. Features included in X_i include

   - x_1, a lab test value (a continuous number)
   - x_2, a different lab test value (a continuous number)
   - x_3, which is 1 if the patient has a prior condition and 0 otherwise (a binary variable)
   - x_4, the patient's age group (a categorical variable)

   Please formulate the problem of maximizing the life expectancy of a new patient $\bar{X}_{n+1}$ using a prediction/prescription framework with random forests. What are the variables for this problem?

2. (5 points) Talking to doctors, you are given two practical guidelines that your model must follow –

   - If $log(x_1) * x_2^2 \geq 5$, the patient cannot receive more than 6 mg of the drug
   - If the patient has a prior condition, they must receive at least 2 mg of the drug

   Incorporate these guidelines into your model, making sure to only use linear constraints in the formulation.

3. (5 points) You run the model and get that the estimated expected costs due to the prescription are 2.5, the estimated costs in the deterministic perfect-foresight counterpart are 2.1, and the estimated expected costs based only on $\mathbf{y}$ values are 3.7. Use these quantities to find an overall estimation of the quality of the model, and discuss what it means.