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.

- False. In the case of logistic regression, the regularization term to achieve robustness is added within the exponential function of the objective and not directly to the objective.

(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).

- False. We need to invert a $n \times n$ matrix in (1) and a $k \times k$ matrix in (2). Given that $k << n$, (2) is more efficient. Given that this was not stressed in the lecture, we also accepted those that answered True and explained that the first version is convex while the second one is not.

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

- False. There is a chance that our entire model is overfitting and thus we do worse than the baseline!

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

- False. The time decreases with n in a certain range according to the phase transition phenomena.

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

- True. Change the data point that corresponds to the median residual!

(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.

- False. We only proved in class that an OCT-H can classify at least as well as a given classification FNN with ReLu, and we never proved the opposite, so this statement is False. However, given the possible confusion surrounding this, we also accepted True answers that explained well the equivalence.

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

- True/False. The stable linear regression problem with a L1 regularization term can reduce to a linear optimization problem but one with L2 cannot, so both answers are correct depending on what you define and discuss as stable linear regression.

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

- True. With sparsity one, that means at every split, OCT-H can only select one variable, and thus is equivalent to OCT.

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

- True. The x-axis of the ROC curve is the false positive rate, which is 1 - specificity.

(k) Lasso leads to sparse models.

- False. Lasso is a form of robustness that sometimes leads to sparse solutions, but is not a sparse method in itself.

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.

   Solution:

   Algorithm is as follows:

   (a) For each $j = 1, \ldots, p$:

   i. Fit a univariate regression using only feature j:

   $$\begin{aligned} \beta_j^\star &= \min_\beta \sum_{i=1}^n (y_i - \beta X_{ij})^2 + \Gamma\beta^2 \\ &= \frac{\boldsymbol{y}^T \boldsymbol{X_j}}{\|\boldsymbol{X_j}\|^2 + \Gamma} \end{aligned}$$

   ii. Plug $\beta_j^\star$ into the loss function and compute the loss associated with feature j:

   $$\begin{aligned} L_j &= \|\boldsymbol{y} - \boldsymbol{X_j}\beta_j^\star\|^2 + \Gamma\beta_j^{\star 2} \\ &= \|\boldsymbol{y}\|^2 - \frac{(\boldsymbol{y}^T \boldsymbol{X_j})^2}{\|\boldsymbol{X_j}\|^2 + \Gamma} \end{aligned}$$

(b) Select feature $j^\star = \text{argmin}_j L_j$ and return $(j^\star, \beta^\star_{j^\star})$.

The complexity of the proposed approach is $\mathcal{O}(np)$: we iterate over p features and, for each of them, compute inner products between n-dimensional vectors.

If $\|\boldsymbol{X_j}\|^2 = 1,\ \forall j$, then it follows that

$$\min_j\ L_j\ =\ \min_j\ \|\boldsymbol{y}\|^2 - \frac{(\boldsymbol{y}^T\boldsymbol{X_j})^2}{\|\boldsymbol{X_j}\|^2+\Gamma}\ =\ \max_j\ \frac{(\boldsymbol{y}^T\boldsymbol{X_j})^2}{1+\Gamma}.$$

Thus, we end up selecting the feature $\boldsymbol{X_j}$ that achieves the maximum absolute (empirical) correlation with the response vector $\boldsymbol{y}$.

Marking scheme: (Points add up to 11.)

- 2 pts for describing a solution method (possibly not using that $k = 1$).
- 1 pt for using the fact that the resulting model is univariate.
- 3 pts for describing the iterative process.
- 2 pts for correctly specifying formulation that is solved in each iteration.
- 1 pt for noting that formulation in each iteration is closed-form solvable.
- 1 pt for giving correct closed-form solution.
- 1 pt for noting that selected feature is the most correlated with the response.

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.

Solution:

The approach does not generalize. We would have to exhaustively enumerate all possible subsets of features of size $s = 1, \ldots, k$, which would yield a total of $\sum_{s=1}^{k} \binom{p}{s}$ subsets. This gives a complexity that grows as $\mathcal{O}(p^k)$, which is not polynomial in p and k.

Marking scheme:

- 1 pt for discussing the complexity of some approach (possibly other than the polynomial-time approach we were looking for; e.g., many students discussed the complexity of the cutting planes method, which -in the worst-case- is exponential).
- 1 pt for discussing the complexity of the approach from Part 1 (possibly claiming that approach does extend in polynomial time).
- 1 pt for correctly noting that the approach does not generalize in polynomial time.
- 2 pts for giving the actual complexity (not necessarily rigorously - any answer that correctly captures the scaling behavior of the algorithm is accepted).

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?

Solution:

The answer needs to include the following three parts:

- Definition and offline computation of the set

$$\mathcal{HC} = \{(i,j) : \quad |\text{cor}(\boldsymbol{X}_i, \boldsymbol{X}_j)| \geq \rho_{\max}\},$$

  where $\text{cor}(\boldsymbol{X}_i, \boldsymbol{X}_j)$ denotes the correlation between features i and j. This is exactly as we did for holistic regression.
- Computation of a **feasible** starting point $\boldsymbol{z_0}$ for the **extended** sparse regression formulation (i.e., $\boldsymbol{z_0}$ should satisfy both the sparsity constraint and the low-pairwise-correlation constraint).

  Here are a few valid choices for $\boldsymbol{z_0}$:
  - $\boldsymbol{z_0} = \boldsymbol{0}$, which trivially satisfies both constraints.
  - Let $j^\star$ be the feature selected in Part 1. Set $z_{j^\star} = 1$ and $z_j = 0,\ \forall j \neq j^\star$.
  - A more sophisticated answer (which would also be a warm-start for the extended formulation) is as follows:

    (a) Rank features based on their associated losses from Part 1: $L_{(1)}, L_{(2)}, \ldots, L_{(p)}$.

    (b) Repeat until k features have been selected:

      i. Among the features that have not already been examined, pick the one that achieves the lowest loss.

      ii. Select this feature if its pairwise correlation with all previously selected features is less than $\rho_{\max}$. Mark this feature as "examined" and proceed to the next feature.

  Note that a starting point obtained via the first-order method that we saw in Lecture 3 or via Lasso is not guaranteed to satisfy the correlation constraint (i.e., in such a solution, we may simultaneously select pairs of features whose pairwise correlation exceeds $\rho_{\max}$).
- Describe and correctly extend cutting planes algorithm shown, e.g., in Lecture 3. In particular, the following modifications are needed:
  - The starting point must be computed as described above.
  - The outer problem must include a low pairwise correlation constraint:

$$z_i + z_j \leq 1,\ \forall (i,j) \in \mathcal{HC}.$$

    This affects the MIO problem solved in each iteration of the cutting planes method, which, in iteration $T+1$, reads:

$$\begin{aligned} \min_{\eta \in \mathbb{R}, \boldsymbol{z} \in \{0,1\}^p} \quad & \eta \\ \text{s.t.} \quad & \eta \geq c(\boldsymbol{z}^t) + \nabla c(\boldsymbol{z}^t)(\boldsymbol{z} - \boldsymbol{z}^t), \quad \forall t = 1, \ldots, T, \\ & \sum_j z_j \leq k, \\ & z_i + z_j \leq 1, \quad \forall (i,j) \in \mathcal{HC}. \end{aligned}$$

Marking scheme: (Points add up to 11.)

- 2 pts for proposing a solution.
- 2 pts for giving a correct cutting planes algorithm for the sparse regression problem.
- 2 pts for defining the set $\mathcal{HC}$ correctly.
- 2 pts for incorporating the low pairwise correlation constraint into the formulation.
- 2 pts for giving any feasible initial point (you do have to give such a point - just stating that the initial point has to be feasible would not be enough).
- 1 pt for giving a "good" initial point (e.g., bullets 2 and 3 in the proposed solution above).

Problem 3, 25 points

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

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

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

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

$$\begin{aligned}
\min_{t_i, s_i, z_{ij}, \mathbf{a}_j} \quad & \sum_{i=1}^{n} t_i \\
\text{s.t.} \quad & t_i \geq y_i - s_i, \quad \forall i \in [n], \\
& t_i \geq -(y_i - s_i), \quad \forall i \in [n], \\
& s_i \geq \mathbf{a}_j^T \mathbf{x}_i, \quad \forall i \in [n], j \in [k], \\
& s_i \leq \mathbf{a}_j^T \mathbf{x}_i + M(1 - z_{ij}), \quad \forall i \in [n], j \in [k], \\
& \sum_{j=1}^{k} z_{ij} = 1, \quad \forall i \in [n], \\
& z_{ij} \in \{0, 1\}, \quad \forall i \in [n], j \in [k].
\end{aligned}$$

   General grading guidelines:

   - Linearizing the absolute values (6 points).
   - Introducing binary variables (2 points).
   - Correct modeling of the max operator (5 points).

   Common mistake:

   - $s_i \geq \mathbf{a}_j^T \mathbf{x}_i$ is NOT SUFFICIENT to model the max operator since we are not minimizing the maximum.

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

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

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

Solution:

We introduce binary variables v_j s.t. $v_j = 1$ if the linear function $\mathbf{a}_j$ is selected, and $\sum_{j=1}^{k_{\max}} v_j = k$. The complete mixed-integer linear optimization formulation is given below

$$
\begin{aligned}
\min_{t_i, s_i, z_{ij}, \mathbf{a}_j, k} \quad & \sum_{i=1}^{n} t_i + \lambda k \\
\text{s.t.} \quad & t_i \geq y_i - s_i, \quad \forall i \in [n], \\
& t_i \geq -(y_i - s_i), \quad \forall i \in [n], \\
& s_i \geq \mathbf{a}_j^T \mathbf{x}_i - M(1 - v_j), \quad \forall i \in [n], j \in [k_{\max}], \\
& s_i \leq \mathbf{a}_j^T \mathbf{x}_i + M(1 - z_{ij}), \quad \forall i \in [n], j \in [k_{\max}], \\
& \sum_{j=1}^{k_{\max}} z_{ij} = 1, \quad \forall i \in [n], \\
& z_{ij} \in \{0, 1\}, \quad \forall i \in [n], j \in [k_{\max}], \\
& -M v_j \leq a_{ji} \leq M v_j, \quad \forall i \in [p], j \in [k_{\max}], \\
& v_j \in \{0, 1\}, \quad \forall j \in [k_{\max}], \\
& z_{ij} \leq v_j, \quad \forall j \in [k_{\max}], \\
& k = \sum_{j=1}^{k_{\max}} v_j,
\end{aligned}
$$

where a_{ji} is the i-th element of $\mathbf{a}_j$.

General grading guidelines:

- Linearizing the absolute values (3 points).
- Introducing binary variables (2 points).
- Correct modeling of the max operator (5 points).
- Correct modeling of the variable k (2 points).

Common mistake:

- Using nonlinear formulations which include product of two variables.

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?

$\max_{\mathbf{z}} \frac{1}{T} \sum_{t=1}^{T} \frac{1}{L_T} \sum_{j \in L_t} C \min(z, y_j) - D \max(z - y_j, 0)$

where

- $z \leq 8$
- $z \geq 0$
- z is the amount of the drug prescribed to the patient in mg
- C and D are the constants
- L_t is the leaf node the new point belongs to in the tth tree

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.

   For the first guideline, we add the following variable and constraints – , $Mv \geq \log(x_1) * x_2^2 - 5$, $z \leq M * (1 - v) + 6, v \in \{0, 1\}$

   For the second guideline, we add the constraint $z \geq 2 - M * (1 - x_2)$

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.

$$1 - \frac{2.5 - 2.1}{3.7 - 2.1} = 0.75$$

This is a pretty good score, so the model is making useful prescriptions.