15.095 Machine Learning Under a Modern Optimization Lens

October 21, 2019

Instructions:

1. There are 100 points in this exam.
2. You have 90 minutes to complete the examination.
3. You may use the book by Bertsimas and Dunn, the slides, the recitation slides, your homeworks, the solutions and your notes.
4. Please explain your work carefully.
5. Good luck!

Problem 1 (40 points)

Indicate whether each statement listed below is True or False. In either case, write a couple of sentences explaining your answer. The correct answer is worth 1 point and the remaining points are given for short correct explanations.

1. (3 points) The value of M in a big-M optimization formulation for a MIO problem impacts solve times because of numerical accuracy issues.

   - False. The impact on solve times is mainly because larger values of M provide weaker continuous relaxations, which allow fewer nodes in a branch-and-bound tree to be pruned.

2. (3 points) Warm starts are used in mixed integer optimization problems to improve the incumbent objective value, but do not improve the lower bound (when minimizing).

- False. By installing an incumbent solution, we are able to prune parts of the search tree, which also improves the lower bound.

3. (4 points) The cutting plane method cannot be applied to sparse logistic regression problems because, unlike sparse linear regression, there is no closed form expression like the one we have for linear regression below:

$$\min_{\boldsymbol{s}\in\{0,1\}^p:\ \boldsymbol{e}^\top \boldsymbol{s}\leq k} f(\boldsymbol{s})$$

where $f(\boldsymbol{s}) := \frac{1}{2}\boldsymbol{y}^\top \left(\mathbb{I}_n + \gamma \sum_{j=1}^p s_j \boldsymbol{K}_j\right)^{-1} \boldsymbol{y}$.

- False. We can apply a cutting-plane method to problems where $f(\boldsymbol{s})$ is convex in $\boldsymbol{s}$, even if $f(\boldsymbol{s})$ cannot be evaluated in closed form. Indeed, we saw in a recitation that we can apply the cutting-plane method to sparse logistic regression.

4. (4 points) Consider the following big-M formulation of the sparse regression problem:

$$\begin{aligned} \min_{\boldsymbol{\beta}\in\mathbb{R}^p, \boldsymbol{z}\in\{0,1\}^p} \quad & \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \\ \text{s.t.} \quad & -M_i z_i \leq \beta_i \leq M_i z_i,\ \forall i \in [p], \\ & \boldsymbol{e}^\top \boldsymbol{z} \leq k. \end{aligned}$$

Then, a valid bound on M_i is given by minimizing/maximizing β_i under the constraint

$$\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 \leq \|\boldsymbol{y}\|_2^2.$$

- True. This follows because $\boldsymbol{\beta} = \boldsymbol{0}$ is a valid solution to the best subset selection problem for any sparsity level, and the optimal objective value of the best subset selection problem therefore cannot exceed $\|\boldsymbol{y}\|_2^2$.

5. (3 points) The out-of-sample AUC of a logistic regression model is always smaller than the in-sample AUC.

- False. For instance, if we have underfit the model and we got "unlucky" with our in-sample AUC then the out-of-sample AUC could certainly be higher.

6. (4 points) The confidence interval of the coefficient β_1 corresponding to variable x_1 in a linear regression model is $[-1, 2]$. This means that we can conclude that variable x_1 does not influence the dependent variable and we can safely conclude that $\beta_1 = 0$.

- False. A better interpretation would be that we can't rule out the possibility that $\beta_1 = 0$ (if we were to take this to be the null hypothesis), but we certainly can't claim that $\beta_1 = 0$ based on this argument.

7. (3 points) Given that feedforward neural networks can be simulated by optimal classification trees with hyperplanes (OCT-H) and that empirically OCT-H stabilize their performance after depth 5, it follows that feedforward neural networks can be simulated by OCT-H of depth 5 with excellent results.

   - True, since we can find an OCT-H of depth 5 which is equally powerful as the OCT-H which exactly simulates the feedforward neural network, by construction.

8. (4 points) Consider the robust regression

$$\min \max_{\|\boldsymbol{\delta x}_i\|_2 \leq \rho} \|\boldsymbol{y} - (\boldsymbol{X} + \boldsymbol{\Delta X})\boldsymbol{\beta}\|_2,$$

   where $\boldsymbol{x}_i$ is the ith row of matrix $\boldsymbol{X}$ and $\boldsymbol{\delta x}_i$ is the uncertainty in the ith row. The equivalent regularized problem is lasso.

   - False. Lasso allows feature-wise uncertainty, so the formulation should be by column and not row.

9. (4 points) Lasso is a sparse regression method.

   - False. Lasso is a robust regression method. It does induce some sparsity by construction but it is not a sparse method by definition.

10. (4 points) Random Forests and OCT both give the relative significance of variables. Therefore, they are equally interpretable.

    - False. OCT is more interpretable as it only has one tree and one can clearly see the decision process. In contrast, random forests involve many trees and are less interpretable.

11. (4 points) OCT is a nonlinear prediction method, whereas logistic regression is a linear prediction method.

    - False. Both methods are nonlienar prediction methods. The splits of OCT are nonlinear, and logistic regression is nonlinear due to its sigmoid function.
    - (Alternative Answer) True. Logistic regression is a linear classifier (its boundaries are always linear) while OCT is nonlinear.

Problem 2 (10 points)

Consider the following data: $(x_{i1}, x_{i2}, y_i),\ i \in N$. We are interested in

$$\min_z E_Y[f(y, z) |\ \boldsymbol{x} = (x_{01}, x_{02})],$$

where $f(y, z) = (\min(y, z))^2$. Show how to solve the problem using OCT as the underlying predictive method.

Note: Following is a sample solution. Other thoughtful constructions that utilize OCT effectively would also be accepted.

Let us use the data to construct an OCT in which the features are (x_{i1}, x_{i2}) and the outcome variable is y_i. Then let A be the leaf that (x_{01}, x_{02}) falls under. Then we can let the $y_i \in A$ be the sample points in which we are taking the expectation over, and approximate $E_Y[f(y, z) \mid \boldsymbol{x} = (x_{01}, x_{02})]$ with:

$$\frac{1}{|A|} \sum_{(x_{i1}, x_{i2}) \in A} f(y_i, z)$$

So the optimization problem now becomes:

$$\min_z \frac{1}{|A|} \sum_{(x_{i1}, x_{i2}) \in A} (\min(y_i, z))^2$$

Problem 3 (25 points)

In this question, we place the problem of compressed sensing, which is perhaps the most important problem in the field of signal processing, under a modern optimization lens. Assume that we are given a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times m}$ and a vector $\boldsymbol{b} \in \mathbb{R}^m$, such that the system $\boldsymbol{Ax} = \boldsymbol{b}$ is over-determined, i.e., admits many feasible solutions $\boldsymbol{x}$. Our task is to find the sparsest vector $\boldsymbol{x} \in \mathbb{R}^n$ such that $\boldsymbol{Ax} = \boldsymbol{b}$. Formally, we consider the problem:

$$\min_{\boldsymbol{x}} \quad \|\boldsymbol{x}\|_0 \text{ s.t. } \boldsymbol{Ax} = \boldsymbol{b}. \tag{1}$$

(a) (5 points) Please write down a big-M formulation of this problem, which could be solved directly by CPLEX or Gurobi.

- $$\min_{\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{z} \in \{0,1\}^n} \quad \sum_i z_i \text{ s.t. } \boldsymbol{Ax} = \boldsymbol{b},\ -Mz_i \le x_i \le Mz_i,\ \forall i \in [n].$$

(b) (5 points) Due to the NP-hardness of part (a)'s formulation, the signal processing community instead solves the following surrogate problem:

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \|\boldsymbol{x}\|_1 \text{ s.t. } \boldsymbol{Ax} = \boldsymbol{b}.$$

Please rewrite this problem as a linear optimization problem.

- $$\min_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{z}\in\mathbb{R}^n} \quad \sum_i z_i \text{ s.t. } \boldsymbol{A}\boldsymbol{x}=\boldsymbol{b},\ -z_i\le x_i\le z_i,\ \forall i\in[n].$$

(c) (5 points) Suppose that after we solve part (b)'s formulation we obtain a solution $\boldsymbol{x}^\star$. Discuss how we can use this solution to provide an upper bound on part (a)'s optimal objective, via a warm-start.

- Need to mention that you set $z_i=0$ if $x_i^\star=0$ and $z_i=1$ otherwise, in addition to setting $\boldsymbol{x}=\boldsymbol{x}^\star$.

(d) (10 points) Propose an algorithm for Problem (1) based on what we have learned on sparse regression. Examples of valid proposals (up to 5 marks per component) include:

- Impose a ridge regularizer and apply the dual approach.
- Relax the equality constraint to $-\epsilon\boldsymbol{e}\le\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\le\epsilon\boldsymbol{e}$.

Note that we only awarded marks for proposals which weren't mentioned in previous parts of the question.

Problem 4 (25 points)

In the classical ridge regression problem we solve

$$\min\ \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|^2, \tag{2}$$

$\boldsymbol{X}\in\mathbb{R}^{n\times p}$ and outcome variable $\boldsymbol{y}\in\mathbb{R}^n$.

We are interested in problems with data $(\boldsymbol{x}_i(t),\ y_i(t))$, $i\in[n]$, $t\in[T]$ and would like to allow the coefficients β_j, $j\in[p]$ to vary slowly with time, that is to be functions $\beta_j(t)$, $t\in[T]$ such that $|\beta_j(t)-\beta_j(t-1)|\le\delta$, $j\in[p]$, $t\in[T]$ for some $\delta>0$.

In addition to $\beta_j(t)$ being slowly varying, we would like the support of $\boldsymbol{\beta}(t)=(\beta_j(t))_{j\in[p]}$, $t\in[T]$ to be slowly varying as well. Propose an extension of (2) to model these slowly varying conditions as a tractable optimization problem, involving both continuous and discrete decision variables.

The marking scheme is as followed:

- Good Attempt: 10/25
- Wrote down linear constraints modeling $|\beta_j(t)-\beta_j(t-1)|\le\delta$: 15/25

- Formulated the sparsity condition as binary variables $z_j(t)$ with constraints on $\beta_j(t)$: 18/25
- Fully correct: 25/25

Note: The following answer is a model answer and not the only correct answer.

There are two types of constraints we need to model here. The first one is that the coefficients are slowly varying, which gives:

$$\beta_j(t) - \beta_j(t-1) \leq \delta \qquad \beta_j(t-1) - \beta_j(t) \leq \delta$$

The second type of constraints is the "slowly varying support". To do so, we let $z_j(t)$ be the binary indicator variable for whether $\beta_j(t)$ is in the support. Then we first have:

$$\beta_j(t) \leq M z_j(t) \qquad \beta_j(t) \geq -M z_j(t)$$

The support at time t is then $\sum_j z_j(t)$. Thus a slowing varying support means $\sum_j |z_j(t) - z_j(t-1)| \leq k$, which can be reformulated as:

$$\sum_j l_j(t) \leq k \qquad z_j(t) - z_j(t-1) \leq l_j(t) \qquad z_j(t-1) - z_j(t) \leq l_j(t)$$

Therefore, in total, we have:

$$\begin{aligned}
\min \quad & \sum_{t\in[T]} \|\boldsymbol{y}_t - \boldsymbol{X}_t\boldsymbol{\beta}(t)\|_2^2 + \lambda\|\boldsymbol{\beta}(t)\|^2 \\
& \beta_j(t) - \beta_j(t-1) \leq \delta \quad \forall t \in [T] \quad \forall j \in [p] \\
& \beta_j(t-1) - \beta_j(t) \leq \delta \quad \forall t \in [T] \quad \forall j \in [p] \\
& \beta_j(t) \geq -M z_j(t) \quad \forall t \in [T] \quad \forall j \in [p] \\
& \beta_j(t) \leq M z_j(t) \quad \forall t \in [T] \quad \forall j \in [p] \\
& \sum_j l_j(t) \leq k \quad \forall t \in [T] \\
& z_j(t) - z_j(t-1) \leq l_j(t) \quad \forall t \in [T] \quad \forall j \in [p] \\
& z_j(t-1) - z_j(t) \leq l_j(t) \quad \forall t \in [T] \quad \forall j \in [p]
\end{aligned}$$