15.095: Machine Learning Under a Modern Optimization Lens **Fall 2021**

Homework 3: Due October 15

Hand in: **pdf** upload to Canvas. Please append any Julia code **at the end** of the whole pdf. If you include Jupyter notebooks in your submissions, please truncate any irrelevant output.

Note: this homework covers lectures six and seven, and recitation four. We recommend attempting questions after the relevant content has been covered in class.

3.1 Question 1: Holistic Regression (50 Points)

In this problem, we utilize the **Airfoil Self-Noise Dataset**, a NASA dataset obtained from a series of aerodynamic and acoustic tests of two and three-dimensional airfoil blade sections conducted in an anechoic wind tunnel (available at the UCI Machine Learning Repository).

The original dataset consists of about 1500 data points and the following 5 features:

X_1: Frequency, in Hertz.

X_2: Angle of attack, in degrees.

X_3: Chord length, in meters.

X_4: Free-stream velocity, in meters per second.

X_5: Suction side displacement thickness, in meters.

We have appended the original dataset by considering interaction terms, i.e., terms of the form X_iX_j, $1 \leq i < j \leq 5$, so the resulting data matrix is

$$\boldsymbol{X} := \begin{bmatrix} | & | & | & | & | & | \\ X_1 & \dots & X_5 & X_6 := X_1X_2 & \dots & X_{15} := X_4X_5 \\ | & | & | & | & | & | \end{bmatrix} \in \mathbb{R}^{n \times p}.$$

In the files `airfoil_X_train.csv` and `airfoil_X_test.csv`, we provide you with training and testing data, respectively. The goal is to predict the following target variable (which you will find in the files `airfoil_Y_train.csv` and `airfoil_Y_test.csv`):

Y: Scaled sound pressure level, in decibels.

We have also normalized the data, so that all variables have zero mean and unit variance.

(a) (5 Points) Fit any regularized linear regression model (e.g., Lasso) to the data. Evaluate its out-of-sample predictive performance.

*(Hint: If you choose ℓ_1-regularization, consider using the `GLMNet` Julia package, a **very** efficient implementation of LASSO; documentation here. Alternatively, see Recitation 3.)*

(b) (10 Points) You are given the following two requirements for your final model:

- To enhance interpretability, the model must consists of at most 10 features.
- To avoid collinearity issues, the model must not contain any two features whose pairwise correlation exceeds 0.7.

Develop a manual procedure to perform feature selection and satisfy the above requirements. *(Note: You do not need to be rigorous here, i.e., it suffices to heuristically post-process your model from Part (a).)* Then, refit a linear regression model using the selected features. Evaluate its out-of-sample predictive performance.

(c) (25 Points) Given initial data $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{y} \in \mathbb{R}^n$, obtain the transformed data $\tilde{\boldsymbol{X}} \in \mathbb{R}^{n\times \tilde{p}}$, with $\tilde{p} = 4p$, by replacing in the data matrix each feature $X_j,\ j = 1,\dots,p$, with the following transformations:

$$\mathcal{T}_j \ = \ \{\tilde{X}_{4(j-1)+1} := X_j,\ \tilde{X}_{4(j-1)+2} := X_j^2,\ \tilde{X}_{4(j-1)+3} := \sqrt{|X_j|},\ \tilde{X}_{4(j-1)+4} := \log\left(|X_j|+\epsilon\right)\}.$$

Then, consider the following (simplified) version of the holistic regression framework:

$$\min_{\boldsymbol{\beta},\beta_0,\boldsymbol{z}} \quad \frac{1}{2}\|\mathbf{y} - \tilde{\boldsymbol{X}}\boldsymbol{\beta} - \beta_0\|_2^2 + \Gamma\|\boldsymbol{\beta}\|_1 \tag{3.1}$$

$$\text{s.t.} \quad -Mz_i \le \beta_i \le Mz_i, \qquad i = 1,\cdots,\tilde{p} \tag{3.2}$$

$$\sum_{i=1}^{\tilde{p}} z_i \le k \tag{3.3}$$

$$\sum_{i:\tilde{X}_i\in\mathcal{T}_j} z_i \le 1, \quad j = 1,\cdots,p \tag{3.4}$$

$$z_i + z_j \le 1, \quad \forall i,j \in \mathcal{HC}(\tilde{\boldsymbol{X}}) \tag{3.5}$$

$$z_i \in \{0,1\}, \quad i = 1,\cdots,\tilde{p}$$

The term $\Gamma\|\boldsymbol{\beta}\|_1$ in the objective function (3.1) models robustness. Constraints (3.2) and (3.3) model sparsity, by requiring that at most k out of the $\tilde{p}$ variables are selected in the linear regression model. Constraint (3.4) models nonlinear transformations, i.e., requires that at most one "version" of each variable is selected. Finally, Constraint (3.5) models pairwise collinearity, where $\mathcal{HC}(\boldsymbol{X})$ is the set

$$\mathcal{HC}(\boldsymbol{X}) = \{(i,j) : |\text{Corr}(X_i, X_j)| \ge \rho\}$$

for some predefined correlation ρ cutoff.

Implement holistic regression as modeled in (3.1)-(3.5), apply it to the (transformed) airfoil data, and evaluate its out-of-sample predictive performance. Concerning the model's hyperparameters, set those that are discussed in Part (b) to the same values, and cross-validate the rest.

(d) (10 Points) For the last part, you need to collaborate with (at least) two classmates of yours. Compare your learned models from Part (b) and from Part (c). Quantify the similarity between the models in each part.

What (other) benefits does holistic regression have over the approach from Part (b)?

3.2 Question 2: Robust Classification (50 points + 10 bonus points)

Support vector machines (SVMs) are binary classification models that have been widely utilized in a variety of applications, especially in the early 2000's. *(Note: The paper that coined the term "Support Vector Machine"*

has $\sim 50,000$ *citations on Google Scholar.)* In its original form, SVM searches for a $(p-1)$-dimensional hyperplane (where p is the data dimension) that achieves the maximum separation between the two classes (assuming such a hyperplane exists, i.e., the data are linearly separable). Here, we study a variation of SVM, namely, soft-margin SVM with ℓ_1-regularization, which relaxes the requirement that the data be linearly separable and instead allows for points to be incorrectly classified. Moreover, the ℓ_2-regularization that appears in the original SVM is replaced by ℓ_1-regularization.

For the computational part of this problem, we use the **Congressional Voting Records Dataset**, which includes votes for each of the U.S. House of Representatives Congressmen on 16 issues such as immigration and education spending; the goal is to predict whether a Congressman is Democrat (-1) or Republican (+1). We have preprocessed the data for you in the files `votes_X_train.csv`, `votes_X_test.csv`, `votes_Y_train.csv`, and `votes_Y_test.csv`.

Formally, consider n data points $(\boldsymbol{x}_i, y_i)$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ and $y_i \in \{\pm 1\}$. Further, assume the classes are imbalanced, so the loss associated with each data point is weighed by

$$c_i = \begin{cases} \frac{n}{\sum_{j=1}^n \mathbb{1}(y_j=1)}, & \text{if } y_i = 1, \\ \frac{n}{\sum_{j=1}^n \mathbb{1}(y_j=-1)}, & \text{if } y_i = -1. \end{cases}$$

The goal is to find a vector $\boldsymbol{w} \in \mathbb{R}^p$ which minimizes the so-called hinge loss:

$$\min_{\boldsymbol{w}} \quad C \cdot \sum_{i=1}^n c_i \max\left\{1 - y_i \boldsymbol{x}_i^\top \boldsymbol{w}, 0\right\} + \|\boldsymbol{w}\|_1, \tag{3.6}$$

where the hyperparameter C can be viewed as the inverse of the regularization weight and trades-off between increasing the margin size and ensuring that most data points lie on the correct side of the margin.

(a) (5 Points) Implement a solver for Problem (3.6). Evaluate both its in-sample and out-of-sample classification accuracy on the given dataset.

In presence of corrupted labels, consider the robust problem:

$$\min_{\boldsymbol{w}} \max_{\boldsymbol{\Delta y} \in \mathcal{U}} \quad C \cdot \sum_{i=1}^n c_i \max\left\{1 - y_i(1 - 2\Delta y_i) \boldsymbol{x}_i^\top \boldsymbol{w}, 0\right\} + \|\boldsymbol{w}\|_1, \tag{3.7}$$

with $\mathcal{U} = \{\boldsymbol{\Delta y} \in \{0,1\}^n : \quad \boldsymbol{e}^\top \boldsymbol{\Delta y} \leq \Gamma\}$.

(b) (5 Points) Give an interpretation (in one line) of the uncertainty set $\mathcal{U}$.

Parts (c), (d), and (e) guide you through the derivation of the robust counterpart for Problem (3.7). The proof is similar to the one we saw in Recitation 4 for robust-in-label-uncertainty logistic regression. However, if you get stuck, you can use the resulting formulation (Problem (3.11)) and implement the computational parts (f) and (g).

(c) (10 Points) Show that the problem

$$\max_{\boldsymbol{\Delta y} \in \mathcal{U}} \quad \sum_{i=1}^n c_i \max\left\{1 - y_i(1 - 2\Delta y_i) \boldsymbol{x}_i^\top \boldsymbol{w}, 0\right\} \tag{3.8}$$

is equivalent to

$$\max_{\boldsymbol{\Delta y} \in \mathcal{U}} \sum_{i=1}^n c_i \left[(\phi_i - \xi_i)\Delta y_i + \xi_i\right] \tag{3.9}$$

with

$$\phi_i := \max(1 + y_i \boldsymbol{x}_i^\top \boldsymbol{w}, 0) \quad \text{and} \quad \xi_i := \max(1 - y_i \boldsymbol{x}_i^\top \boldsymbol{w}, 0). \tag{3.10}$$

(Caution: Make sure you understand what the variables are in this problem.)

Explain why problems (3.8) and (3.9) can be optimized over $\boldsymbol{\Delta y} \in \tilde{\mathcal{U}} = \{\boldsymbol{\Delta y} \in [0,1]^n : \quad \boldsymbol{e}^\top \boldsymbol{\Delta y} \leq \Gamma\}$ instead of $\mathcal{U}$.

(Hint: Think of why we were able to do a similar trick in stable regression.)

(d) (10 Points) Encode the constraints in equation (3.10) as integer linear constraints.

(Hint: It might be a good idea to introduce a binary variable for each constraint. Then, use big-M.)

(e) (10 Points) Show that problem (3.7) can be reformulated as the following **minimization** mixed-integer **linear** optimization problem:

$$\begin{aligned}
\min \quad & \sum_{j=1}^{p} \omega_j \; + \; C\Gamma q \; + \; C \sum_{i=1}^{n} (r_i + c_i \xi_i) \\
\text{s.t.} \quad & q + r_i \geq c_i(\phi_i - \xi_i), \quad i = 1, \ldots, n, \\
& \phi_i \geq 1 + y_i \mathbf{x_i}^\top \mathbf{w}, \quad i = 1, \ldots, n, \\
& \phi_i \leq 1 + y_i \mathbf{x_i}^\top \mathbf{w} + M(1 - t_i), \quad i = 1, \ldots, n, \\
& \phi_i \geq 0, \quad i = 1, \ldots, n, \\
& \phi_i \leq M t_i, \quad i = 1, \ldots, n, \\
& \xi_i \geq 1 - y_i \mathbf{x_i}^\top \mathbf{w}, \quad i = 1, \ldots, n, \\
& \xi_i \leq 1 - y_i \mathbf{x_i}^\top \mathbf{w} + M(1 - s_i), \quad i = 1, \ldots, n, \\
& \xi_i \geq 0, \quad i = 1, \ldots, n, \\
& \xi_i \leq M s_i, \quad i = 1, \ldots, n, \\
& \omega_j \geq w_j \quad j = 1, \ldots, p, \\
& \omega_j \geq -w_j \quad j = 1, \ldots, p, \\
& q \geq 0, \\
& r_i \geq 0, \quad i = 1, \ldots, n, \\
& t_i \in \{0, 1\}, \quad i = 1, \ldots, n, \\
& s_i \in \{0, 1\}, \quad i = 1, \ldots, n.
\end{aligned} \tag{3.11}$$

(Hint: Convert problem (3.9) into a minimization problem.)

(f) (10 Points) Implement a solver for Problem (3.11). Evaluate both its in-sample and out-of-sample classification accuracy.

(g) (10 Bonus Points) Problem (3.11) is a mixed-integer linear optimization problem. In practice, you may be interested to obtain a solution fast even if that requires sacrificing the solution's optimality guarantees. For this purpose, implement a solver for the linear relaxation of Problem (3.11) (whereby all binary variables are allowed to take values in the $[0, 1]$ interval). Evaluate both its in-sample and out-of-sample classification accuracy.