15.094J/1.142J Robust Modeling, Optimization and Computation

Problem Set 2

Due March 8, in class

Exercise 2.1

Consider the RO problem

$$\begin{aligned} &\max\ \mathbf{c}^T\mathbf{x} \\ &\text{s.t. } \mathbf{a}^T\mathbf{x} \le \mathbf{b},\ \forall \mathbf{a} \in \mathcal{U}, \end{aligned} \tag{1}$$

where

$$\mathcal{U} = \{\mathbf{a} \mid \mathbf{a} = \hat{\mathbf{a}} + \mathbf{\Delta u},\ \|\mathbf{u}\| \le 1\} \tag{2}$$

for a given matrix (of appropriate dimensions) $\mathbf{\Delta}$ and norm $\|\cdot\|$.

Write down (with proof) the robust counterpart of problem (1) when the norm used to define (2) is $\ell_1 \cap \ell_\infty$, defined by

$$\|\mathbf{u}\|_{\ell_1 \cap \ell_\infty} = \max\left\{\frac{1}{\Gamma}\|\mathbf{u}\|_1, \|\mathbf{u}\|_\infty\right\}$$

for a fixed positive constant Γ.

Exercise 2.2

Consider the robust combinatorial optimization problem

$$\begin{aligned} &\min_{\mathbf{x}}\ \mathbf{c}^T\mathbf{x} + \max_{\substack{S,T: \\ S\subseteq N,\ |S|\le\Gamma_1 \\ T\subseteq M,\ |T|\le\Gamma_2}} \left(\sum_{j\in S} d_j x_j + \sum_{k\in T} f_k x_k\right) \\ &\text{s.t. } \mathbf{x} \in X \subseteq \{0,1\}^{2n} \end{aligned} \tag{3}$$

where $N = \{1,\ldots,N\}$ and $M = \{n+1,\ldots,2n\}$. Assume that $d_1 \ge d_2 \ge \ldots \ge d_n \ge 0$ and $f_{n+1} \ge f_{n+2} \ge \ldots \ge f_{2n} \ge 0$, Γ_1, Γ_2 both positive integers, and X is a subset of $\{0,1\}^{2n}$.

Essentially, what we are modeling here is that at most Γ_1 of $\{c_1,\ldots,c_n\}$ and Γ_2 of $\{c_{n+1},\ldots,c_{2n}\}$ can vary from their nominal values.

(a) Using ideas from Lecture 5, find the resulting robust counterpart of (3).

(b) Suppose we have a specialized fast subroutine for solving problems of the form

$$\min_{\mathbf{x}}\ \bar{\mathbf{c}}^T\mathbf{x} \tag{4}$$

$$\text{s.t. } \mathbf{x} \in X \subseteq \{0,1\}^{2n} \tag{5}$$

Propose an algorithm which solves problem (3) using the above subroutine.

Exercise 2.3

In this question we deal with the Framingham Heart Study dataset which is provided in the file“Framingham.csv”. This dataset comes from a cardiovascular study on residents of Framingham, MA, and contains data about the study participants such as sex, age, education, smoking habits, and medical measurements. We denote the vector of these measurements for each patient i as $\mathbf{x_i}$. The final column of the dataset, denoted $y_i \in \{-1, 1\}$ indicates whether the patient i is at risk for coronary heart disease in the next ten years, as determined by a doctor (-1 means no and $+1$ means yes). Using the data that has been collected, we want to develop a model for predicting this outcome without the need of a doctor.

To do this, we will use a variant of the Support Vector Machine (SVM), a common machine learning model. The SVM learns a vector $\mathbf{w}$ and scalar b, and then uses these to predict the value of y using $y = sign(\mathbf{w}^T\mathbf{x} - b)$. To learn $\mathbf{w}$ and b, we solve the following optimization problem:

$$\min_{\mathbf{w},b} \sum_{i=1,\dots,n} c_i \max\{1 - y_i(\mathbf{w}^T\mathbf{x}_i - b), 0\},$$

where we define $c_i = 1$ if $y_i = -1$ and $c_i = 5$ if $y_i = +1$ in order to rebalance the dataset. This can be rewritten as a linear optimization problem:

$$\begin{aligned} \min_{\mathbf{w},b,\mathbf{z}} \quad & \sum_{i=1,\dots,n} c_i z_i \\ \text{s.t.} \quad & z_i \geq 1 - y_i(\mathbf{w}^T\mathbf{x}_i - b) \quad && i = 1,\dots,n \\ & z_i \geq 0 && i = 1,\dots,n \end{aligned}$$

To provide some intuition for this model, note that $y_i(\mathbf{w}^T\mathbf{x}_i - b) > 0$ only if y_i and the SVM prediction for $\mathbf{x}_i$ are the same, meaning point i is correctly classified. In this way, we can interpret $y_i(\mathbf{w}^T\mathbf{x}_i - b)$ as how “correctly” we have classified point i. The objective penalizes points with $y_i(\mathbf{w}^T\mathbf{x}_i - b) < 1$, which are those that are incorrectly classified or not classified “correctly enough”, with the penalty increasing according to the degree of misclassification.

Unfortunately, medical data in particular tends to be prone to errors and uncertainty in the dataset, meaning that a predictive model created without accounting for uncertainty is likely to perform poorly on new data. Your task is to use the provided data to develop a robust formulation of this SVM model that can predict whether the patient is at risk for coronary heart disease in the next ten years. You should implement this formulation, test it out using the provided data, and discuss the results. This is intended to be an open-ended question; there are many ways that you could make use of the data to robustify the problem, and you might like to compare some alternatives.