Optimization for Data Science

F. Rinaldi[1]

1 Università degli Studi di Padova | Dipartimento Matematica

Padova
2020

Outline

Optimization for Data Science

Min-max Problems

Min-max Problem

$$\min_{x \in X} \max\{f_1(x), \ldots, f_l(x)\}. \tag{1}$$

Assume that X is a polyhedron and that $f_i,\ i = 1, \ldots, l$ are linear functions:

$$f_i(x) = a_i^T x + b_i, \quad i = 1, \ldots, l.$$

- It is easy to verify that the problem is nonlinear.
- Use new variable z to rewrite the problem as follows:

$$\begin{aligned} &\min_{x,z} z \\ &z = \max\{f_1(x), \ldots, f_l(x)\} \\ &x \in X. \end{aligned}$$

- We can hence replace the max function with a set of constraints:

$$\begin{aligned} &\min_{x,z} z \\ &f_i(x) \le z, \quad i = 1, \ldots, l \\ &x \in X \end{aligned}$$

LP Formulation

LP Formulation of Min-max Problem

keeping in mind that all functions are linear, we get:

$$\begin{aligned} &\min_{x,z} z \\ &a_i^T x + b_i \leq z, \quad i = 1, \ldots, l \\ &x \in X. \end{aligned}$$

- In general, we might have a feasible solution that satisfies the following inequality

$$z > \max\{f_i(x), i = 1, \ldots, l\}.$$

- the objective function ensures that the optimal value of z is the same as the value of $\max\{f_i(x), i = 1, \ldots, l\}$.
- We can easily check that in the max-max case the transformation does not hold.

Absolute Value Program

Absolute Value Program

Let us consider the following problem:

$$\begin{aligned}&\min_{x,y} \sum_j c_j |x_j| + \sum_k d_k y_k \\ &(x, y) \in C\end{aligned}$$

assume that C is a polyhedron and that $c_j \geq 0 \; \forall j$.

Comments

- We transform variables the following way:

$$x = x_j^+ - x_j^-, \quad x_j^+ \geq 0, x_j^- \geq 0.$$

- Transformation we give here is not unique:
  - case 1: $x_j \geq 0$. We get $x_j^+ = x_j + \delta = |x_j| + \delta$ e $x_j^- = \delta$,
  - case 2: $x_j < 0$. We get $x_j^+ = \delta$ e $x_j^- = -x_j + \delta = |x_j| + \delta$,

  with $\delta \geq 0$. If $\delta = 0$, one component must be zero and the other one is equal to $|x_j|$.
- Notice that

$$x_j^+ + x_j^- = |x_j| + 2\delta.$$

- Replacing the term $|x_j|$ in the objective function with the sum of x_j^+ and x_j^-, we get (due to $c_j \geq 0$) that for the optimal solution $x_j^+ = 0$ or $x_j^- = 0$ (or, equivalently, $\delta = 0$).

LP Reformulation

LP model

$$
\begin{aligned}
&\min_{x,y} \sum_j c_j(x_j^+ + x_j^-) + \sum_k d_k y_k \\
&(x_j^+ - x_j^-, y) \in C \\
&x_j^+ \geq 0, x_j^+ \geq 0, \quad = 1, \ldots, n
\end{aligned}
$$

Different LP Formulation

- A different formulation can be obtained by simply rewriting the absolute value as follows

$$|x_j| = \max\{x_j, -x_j\}.$$

Min-max Formulation

Replacing the absolute value in the objective function with this new term, we have:

$$\min_{x,y} \sum_j c_j \max\{x_j, -x_j\} + \sum_k d_k y_k$$
$$(x, y) \in C$$

Getting the new LP

- if we introduce new variables z_j, likewise the min-max problem, we can write:

$$\begin{aligned} &\min_{x,y,z} \sum_j c_j z_j + \sum_k d_k y_k \\ &z_j = \max\{x_j, -x_j\} \\ &(x, y) \in C. \end{aligned}$$

LP Model

$$\begin{aligned} &\min_{x,y,z} \sum_j c_j z_j + \sum_k d_k y_k \\ &-z_j \leq x_j \leq z_j \quad j = 1, \ldots, n \\ &(x, y) \in C. \end{aligned}$$

Linear Regression Models

Our Model

Build a mathematical (linear) model related to a specific physical problem, having a finite set of experimental measurements available. Let

$$y = a^T x + b$$

be the model considered, where

- $x \in \mathbb{R}^n$ is the input vector for the model;
- $y \in \mathbb{R}$ is the output of the model;
- $a \in \mathbb{R}^n$ and $b \in R$ are the parameters related to the model.

- We further assume to have a set of input-output samples (*training set*):

$$T = \{(x^1, y^1), \dots, (x^m, y^l)\}.$$

- For each sample, we define the error between real and model output (i.e., the *error term*):

$$E_i = y^i - (a^T x^i + b).$$

Classic Approach

- GOAL: Determine the set of parameters that better represent the phenomenon under analysis, i.e., the one minimizing the errors over the training set $E_i, \; i = 1, \ldots, l$.
- Simplest choice: use least-square error!

Least-square Problem

Consider the square loss function and then solve the well-known *least-square* problem

$$\min_{a,b} \sum_{i=1}^{l} (y^i - a^T x^i - b)^2.$$

REMARK: We will see two alternative ways to model the problem using LP.

Min-max Formulation

Min-max Formulation

$$\min_{a,b} \max_{i} |y^i - a^T x^i - b|.$$

IDEA: minimize the possible loss for a worst case (*maximum loss*) scenario, that is minimizing the biggest error over the training set.

- Taking into account the transformations we described, we can write:

$$\begin{aligned} &\min_{a,b,z} z \\ &|y^i - a^T x^i - b| \leq z \quad \forall\, i = 1, \ldots, l \end{aligned}$$

- Using same tricks as before we get an LP:

$$\begin{aligned} &\min_{a,b,z} z \\ &-z \leq y^i - a^T x^i - b \leq z \quad \forall\, i = 1, \ldots, l. \end{aligned}$$

Absolute Value Formulation

Absoute Value Formulation

$$\min_{a,b} \sum_{i=1}^{l} |y^i - a^T x^i - b|.$$

This is the well-known Least Absolute Deviation (LAD) model (also known as least absolute residual, least absolute error or least absolute value model).

The reasons why we choose to use ℓ_1-norm formulation are the following:

1. the model we get is very easy to solve (since it is equivalent to a linear programming problem);
2. ℓ_1-norm is less sensitive to outliers (i.e., usually occurring when the underlying data distributions have pronounced tails).

LP Transformations

- Using the transformations we described, we get:

$$\min_{a,b,z} \sum_{i=1}^{l} z_i$$
$$|y^i - a^T x^i - b| \leq z_i \quad \forall\, i = 1, \ldots, l$$

- that is

$$\min_{a,b,z} \sum_{i=1}^{l} z_i$$
$$-z_i \leq y^i - a^T x^i - b \leq z_i \quad \forall\, i = 1, \ldots, l.$$

- Using the alternative transformation, we can write:

$$\min_{a,b,v,u} \sum_{i=1}^{l} v_i + u_i$$
$$v_i - u_i = y^i - a^T x^i - b \;\forall\, i = 1, \ldots, l$$
$$v \geq 0, \;\; u \geq 0.$$

Compressive Sensing

GOAL

Reconstructing a given input signal by means of a linear combination of elementary signals.

- These elementary signals do usually belong to a large, linearly dependent collection (Dictionary).
- A preference for linear combinations involving only a few elementary signals is obtained by penalizing non-zero coefficients.
- A well-known “penalty function” is the number of elementary signals used in the approximation.
- Obviously the choice we make about the specified collection, the linear model and the sparsity criterion must be justified by the domain of the problem we deal with.

Overcomplete Dictionaries

- Consider a real-valued, finite-length, one-dimensional, discrete-time input signal b, which we view as an $m \times 1$ column vector in R^m with elements b_i $i = 1, \ldots, m$.
- A *dictionary*
$$D = \{A_j \in \mathbb{R}^m \;:\; j = 1, \ldots, n\}$$
of elementary discrete-time signals, usually called **atoms**.
- Represent our signal as a linear combination of the atoms in this dictionary:
$$b = \sum_{j=1}^{n} x_j A_j \;.$$
- In many applications the dictionary we deal with is *overcomplete*, which means $m < n$.
- In this case, the atoms form a linear dependent set...there exists an infinite number of approximations for a given input signal.
- We are basically interested in representations having as few nonzero coefficients x_j as possible!

Enforcing Sparsity

Sparse Optimization Problem

Function $P(x)$ measuring the *sparsity* of a solution x is needed. The optimization problem we want to solve is

$$\begin{aligned} &\min_{x \in \mathbb{R}^n} P(x) \\ &Ax = b \end{aligned} \tag{2}$$

with A an $\mathbb{R}^{m \times n}$ matrix having as columns A_j the elementary signals of the dictionary D.

- A good measure of sparsity is the number of nonzero elements of the vector x.
- We can use the *support function* (i.e., function measuring the number of non-zero component in x)!

Minimum Weight Solution to Linear Equations

New Model

Set $P(x) = |\text{supp}(x)| = \text{card}\{i : x_i \neq 0\}$, thus getting:

$$\begin{aligned} &\min_{x\in\mathbb{R}^n} |\text{supp}(x)| \\ &Ax = b\ . \end{aligned} \tag{3}$$

- This is a classical problem and was referred to as *minimum weight solution to linear equations*.
- There is no polynomial time algorithm that computes an approximate solution for it.
- It is a hard problem!!!
- Support function is usually called ℓ_0 norm in Machine Learning.

How to Handle the Problem?

- IDEA: Replace the objective function with a relaxed version that can be handled efficiently.
- ℓ_1 norm represents the best convex approximant of the ℓ_0 norm.
- Using the ℓ_1 norm in place of the ℓ_0 norm is a natural strategy to obtain a convex problem we can easily handle.
- This is the well-known *Basis Pursuit Method* proprosed by Chen, Donoho and Saunders.

Classic Basis Pursuit (BP) Problem

$$\begin{aligned} &\min_{x \in \mathbb{R}^n} \|x\|_1 \\ &Ax = b\ , \end{aligned} \tag{4}$$

with $\|x\|_1 = \sum_{i=1}^n |x_i|$.

Getting an LP

LP Formulation for BP

BP can be expressed as the following linear programming problem

$$\min_{x,y} \sum_{i=1}^{n} y_i$$
$$Ax = b,$$
$$-y \leq x \leq y$$

- Solved efficiently using modern methods.
- In some cases BP equivalent to the original ℓ_0 norm problem.