Tutorial 2: Optimization problems

Optimization for machine learning, M2 MIAGE ID Apprentissage

September 25, 2023

Exercise 1: Affine regression

We consider a dataset under the form of a feature matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ and a vector of labels $\boldsymbol{y} \in \mathbb{R}^n$. We seek an affine relationship between the features and the labels, hence we consider the following affine regression:

$$\underset{\substack{\boldsymbol{w} \in \mathbb{R}^d \\ z \in \mathbb{R}}}{\text{minimize}} \; f\left(\begin{bmatrix} \boldsymbol{w} \\ z \end{bmatrix}\right) := \frac{1}{2} \|\boldsymbol{X}\boldsymbol{w} + z\boldsymbol{e} - \boldsymbol{y}\|^2, \quad \boldsymbol{e} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}. \tag{1}$$

a) The function f is continuously differentiable, and its gradient is given by

$$\nabla f\left(\begin{bmatrix} \boldsymbol{w} \\ z \end{bmatrix}\right) = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} \begin{bmatrix} \boldsymbol{w} \\ z \end{bmatrix} - \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{y},$$

where $\boldsymbol{Y} = [\boldsymbol{X} \; \boldsymbol{e}] \in \mathbb{R}^{n \times (d+1)}$. Using this expression, justify that the first-order optimality conditions for this problem correspond to a linear system of equations.

b) Suppose that there exists $\boldsymbol{w}^* \in \mathbb{R}^d$ such that $\boldsymbol{X}\boldsymbol{w}^* = \boldsymbol{y}$. Show that $(\boldsymbol{w}^*, z^* = 0)$ is a solution of the affine regression problem.

Exercise 2: Stratified model

Consider a dataset divided in two groups:

$$\boldsymbol{X}_1 \in \mathbb{R}^{n_1 \times d}, \boldsymbol{y}_1 \in \mathbb{R}^{n_1}, \quad \boldsymbol{X}_2 \in \mathbb{R}^{n_2 \times d}, \boldsymbol{y}_2 \in \mathbb{R}^{n_2}.$$

Such a division is typically the result of a striking difference between the examples (for instance, medical data for two age categories).

For each group, we seek a linear model that best fits the data, i.e. a vector $\boldsymbol{w}_1 \in \mathbb{R}^d$ such that $\boldsymbol{X}_1\boldsymbol{w}_1 \approx \boldsymbol{y}_1$ and a vector $\boldsymbol{w}_2 \in \mathbb{R}^d$ such that $\boldsymbol{X}_2\boldsymbol{w}_2 \approx \boldsymbol{y}_2$. Because of the similar nature of these data samples, we would like the models to be as close as possible, i.e. $\boldsymbol{w}_1 \approx \boldsymbol{w}_2$. All these modeling assumptions lead to the following problem:

$$\underset{\boldsymbol{w}_1 \in \mathbb{R}^d, \boldsymbol{w}_2 \in \mathbb{R}^d}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{X}_1\boldsymbol{w}_1 - \boldsymbol{y}_1\|_2^2 + \frac{1}{2}\|\boldsymbol{X}_2\boldsymbol{w}_2 - \boldsymbol{y}_2\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2^2, \tag{2}$$

where $\lambda \geq 0$.

a) Suppose first that $\lambda = 0$, and that there exists $\boldsymbol{w}_1^* \in \mathbb{R}^d$ and $\boldsymbol{w}_2^* \in \mathbb{R}^d$ such that $\boldsymbol{X}_1\boldsymbol{w}_1^* = \boldsymbol{y}_1$ and $\boldsymbol{X}_2\boldsymbol{w}_2^* = \boldsymbol{y}_2$. Justify that the pair $(\boldsymbol{w}_1^*, \boldsymbol{w}_2^*)$ forms a solution of problem (2).

b) Suppose now that $\lambda > 0$.
   i) Suppose first that $\boldsymbol{w}_1^* = \boldsymbol{w}_2^*$. Justify that the result of question a) remains valid.
   ii) Suppose now that $\boldsymbol{w}_1^* \neq \boldsymbol{w}_2^*$, i.e. $\|\boldsymbol{w}_1^* - \boldsymbol{w}_2^*\| > 0$. Suppose further that there exists a vector $\bar{\boldsymbol{w}} \in \mathbb{R}^d$ such that

$$\|\boldsymbol{X}_1\bar{\boldsymbol{w}} - \boldsymbol{y}_1\|^2 < \frac{\lambda}{2}\|\boldsymbol{w}_1^* - \boldsymbol{w}_2^*\|^2, \quad \|\boldsymbol{X}_2\bar{\boldsymbol{w}} - \boldsymbol{y}_2\|^2 < \frac{\lambda}{2}\|\boldsymbol{w}_1^* - \boldsymbol{w}_2^*\|^2.$$

   Show that the existence of $\bar{\boldsymbol{w}}$ implies that the pair $(\boldsymbol{w}_1^*, \boldsymbol{w}_2^*)$ cannot be a solution of the problem.

Exercise 3: Chebyshev approximation

Let $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ be n vectors in $\mathbb{R}^d$, and $\boldsymbol{y} \in \mathbb{R}^n$. We seek a linear model that explains every coefficient y_i from the vector $\boldsymbol{x}_i$, by solving the following optimization problem:

$$\underset{\boldsymbol{w} \in \mathbb{R}^d}{\text{minimize}} \ \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_\infty = \max_{i=1,\dots,n} \left|\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} - y_i\right|, \tag{3}$$

where $\boldsymbol{X} = [\boldsymbol{x}_i^{\mathrm{T}}]_i \in \mathbb{R}^{n \times d}$.

a) Let $\boldsymbol{w}^* \in \mathbb{R}^n$ be a solution of this optimization problem. Does it imply that the optimal value of the problem is 0?

b) It can be shown that problem (3) is equivalent to the problem

$$\begin{array}{lll} \text{minimize}_{\substack{\boldsymbol{w} \in \mathbb{R}^d \\ t \in \mathbb{R}}} & t & \\ \text{subject to} & -t - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} + y_i & \leq 0, \quad i = 1, \dots, n \\ & -t + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} - y_i & \leq 0, \quad i = 1, \dots, n \\ & t & \geq 0, \end{array} \tag{4}$$

in the sense that solving (4) gives immediately a solution and the optimal value for problem (3). Our goal is to prove this claim.
   i) Consider a solution $(t^*, \boldsymbol{w}^*)$ of problem (4). Show that $t^* = \|\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{y}\|_\infty$.
   ii) Consider now any pair $(t, \boldsymbol{w})$ satisfying the constraints of problem (4). Show that $t^* \leq \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_\infty \leq t$.
   iii) Conclude that $\boldsymbol{w}^*$ is a solution of problem (3), and that t^* is the optimal value.

c) Problem (4) is a linear program. Why is this formulation more interesting than that of the original problem?

Solutions

Solutions for Exercise 1

Goal: Generalize the results for linear least squares (and review these results at the same time).

a) The first-order necessary optimality condition for this problem can be stated as follows. If $(\boldsymbol{w}^*, z^*) \in \mathbb{R}^d \times \mathbb{R}$ is a local minimum of the problem, then

$$\nabla f \left(\begin{bmatrix} \boldsymbol{w}^* \\ z^* \end{bmatrix} \right) = \mathbf{0}.$$

Using the formula for the gradient, we obtain

$$\boldsymbol{Y}^{\mathrm{T}} \boldsymbol{Y} \begin{bmatrix} \boldsymbol{w} \\ z \end{bmatrix} - \boldsymbol{Y}^{\mathrm{T}} \boldsymbol{y} = \mathbf{0} \quad \Leftrightarrow \quad \boldsymbol{Y}^{\mathrm{T}} \boldsymbol{Y} \begin{bmatrix} \boldsymbol{w} \\ z \end{bmatrix} = \boldsymbol{Y}^{\mathrm{T}} \boldsymbol{y},$$

which is indeed a linear system of equations.

b) Evaluating the objective at $\begin{bmatrix} \boldsymbol{w}^* \\ z^* = 0 \end{bmatrix}$ yields

$$\begin{aligned} f \left(\begin{bmatrix} \boldsymbol{w}^* \\ z^* \end{bmatrix} \right) &= \frac{1}{2} \left\| \boldsymbol{X} \boldsymbol{w}^* + z^* \boldsymbol{e} - \boldsymbol{y} \right\|^2 \\ &= \frac{1}{2} \left\| \boldsymbol{X} \boldsymbol{w}^* - \boldsymbol{y} \right\|^2 \\ &= 0. \end{aligned}$$

As a result, the vector $\begin{bmatrix} \boldsymbol{w}^* \\ 0 \end{bmatrix}$ corresponds to a zero objective value. Since

$$f \left(\begin{bmatrix} \boldsymbol{w} \\ z \end{bmatrix} \right) = \frac{1}{2} \left\| \boldsymbol{X} \boldsymbol{w} + z \boldsymbol{e} - \boldsymbol{y} \right\|^2 \geq 0,$$

for any $\boldsymbol{w} \in \mathbb{R}^d$ et $z \in \mathbb{R}$, the vector $\begin{bmatrix} \boldsymbol{w}^* \\ 0 \end{bmatrix}$ is a global minimum of the problem.

Solutions for Exercise 2

Goal: Manipulate the notion of solution and easy proofs on nonnegative functions.

a) When $\lambda = 0$, the problem becomes

$$\underset{\boldsymbol{w}_1 \in \mathbb{R}^d, \boldsymbol{w}_2 \in \mathbb{R}^d}{\text{minimize}} \; f(\boldsymbol{w}_1, \boldsymbol{w}_2), \quad \text{where} \quad f(\boldsymbol{w}_1, \boldsymbol{w}_2) = \frac{1}{2} \| \boldsymbol{X}_1 \boldsymbol{w}_1 - \boldsymbol{y}_1 \|^2 + \frac{1}{2} \| \boldsymbol{X}_2 \boldsymbol{w}_2 - \boldsymbol{y}_2 \|^2.$$

Evaluating f at $(\boldsymbol{w}_1^*, \boldsymbol{w}_2^*)$ gives

$$f(\boldsymbol{w}_1^*, \boldsymbol{w}_2^*) = \frac{1}{2} \| \boldsymbol{X}_1 \boldsymbol{w}_1^* - \boldsymbol{y}_1 \|^2 + \frac{1}{2} \| \boldsymbol{X}_2 \boldsymbol{w}_2^* - \boldsymbol{y}_2 \|^2 = 0 + 0 = 0,$$

where we used the assumption that $\boldsymbol{X}_1\boldsymbol{w}_1^* = \boldsymbol{y}_1$ and $\boldsymbol{X}_2\boldsymbol{w}_2^* = \boldsymbol{y}_2$. In addition, the function f is nonnegative, thus

$$\forall(\boldsymbol{w}_1, \boldsymbol{w}_2) \in (\mathbb{R}^d)^2, \qquad f(\boldsymbol{w}_1, \boldsymbol{w}_2) \geq 0 = f(\boldsymbol{w}_1^*, \boldsymbol{w}_2^*).$$

This property shows that the pair $(\boldsymbol{w}_1^*, \boldsymbol{w}_2^*)$ is a global minimum of f or, equivalently, a solution to problem (2).

b) (Case $\lambda > 0$).

i) For convenience of notation, define

$$f_\lambda(\boldsymbol{w}_1, \boldsymbol{w}_2) = \frac{1}{2}\|\boldsymbol{X}_1\boldsymbol{w}_1 - \boldsymbol{y}_1\|^2 + \frac{1}{2}\|\boldsymbol{X}_2\boldsymbol{w}_2 - \boldsymbol{y}_2\|^2 + \frac{\lambda}{2}\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|^2.$$

Similarly to question a), this function is a sum of squares, and thus it is nonnegative for any pair $(\boldsymbol{w}_1, \boldsymbol{w}_2)$. Moreover,

$$f_\lambda(\boldsymbol{w}_1^*, \boldsymbol{w}_2^*) = f(\boldsymbol{w}_1^*, \boldsymbol{w}_2^*) + \frac{\lambda}{2}\|\boldsymbol{w}_1^* - \boldsymbol{w}_2^*\|^2 = f(\boldsymbol{w}_1^*, \boldsymbol{w}_2^*) = 0,$$

where we used the assumption that $\boldsymbol{w}_1^* = \boldsymbol{w}_2^*$ made in this question together with the result of question a). Here again, we have shown that $(\boldsymbol{w}_1^*, \boldsymbol{w}_2^*)$ is a solution of the problem (i.e. a global minimum of f_λ) by observing that

$$\forall(\boldsymbol{w}_1, \boldsymbol{w}_2) \in (\mathbb{R}^d)^2, \qquad f_\lambda(\boldsymbol{w}_1, \boldsymbol{w}_2) \geq 0 = f_\lambda(\boldsymbol{w}_1^*, \boldsymbol{w}_2^*)$$

ii) *In this question, we no longer assume that $\boldsymbol{w}_1^*$ and $\boldsymbol{w}_2^*$ are identical.* Evaluating f_λ at the pair $(\bar{\boldsymbol{w}}, \bar{\boldsymbol{w}})$ gives

$$\begin{aligned} f_\lambda(\bar{\boldsymbol{w}}, \bar{\boldsymbol{w}}) &= \frac{1}{2}\|\boldsymbol{X}_1\bar{\boldsymbol{w}} - \boldsymbol{y}_1\|^2 + \frac{1}{2}\|\boldsymbol{X}_2\bar{\boldsymbol{w}} - \boldsymbol{y}_2\|^2 + \frac{\lambda}{2}\|\bar{\boldsymbol{w}} - \bar{\boldsymbol{w}}\|^2 \\ &= \frac{1}{2}\|\boldsymbol{X}_1\bar{\boldsymbol{w}} - \boldsymbol{y}_1\|^2 + \frac{1}{2}\|\boldsymbol{X}_2\bar{\boldsymbol{w}} - \boldsymbol{y}_2\|^2 \\ &\leq \frac{\lambda}{4}\|\boldsymbol{w}_1^* - \boldsymbol{w}_2^*\|^2 + \frac{\lambda}{4}\|\boldsymbol{w}_1^* - \boldsymbol{w}_2^*\|^2 \\ &= \frac{\lambda}{2}\|\boldsymbol{w}_1^* - \boldsymbol{w}_2^*\| = f_\lambda(\boldsymbol{w}_1^*, \boldsymbol{w}_2^*), \end{aligned}$$

where the inequality follows from the assumption on $\bar{\boldsymbol{w}}$, and the last equality comes from the assumption that $\boldsymbol{X}_1\boldsymbol{w}_1^* = \boldsymbol{y}_1$ and $\boldsymbol{X}_2\boldsymbol{w}_2^* = \boldsymbol{y}_2$.
Overall, we have shown that there exists a pair of vectors (here $(\bar{\boldsymbol{w}}, \bar{\boldsymbol{w}})$ that yields a strictly better function value than $(\boldsymbol{w}_1^*, \boldsymbol{w}_2^*)$, from which we can conclude that the latter is not a global minimum of the problem.
Note: The result of this question illustrates that, although both $\boldsymbol{w}_1^*$ and $\boldsymbol{w}_2^*$ perfectly interpolate their own datasets, they do not necessarily form a good model for the entire dataset in the sense of problem (2).

Solutions for Exercise 3

Goal: Go deeper into the linear programming formulation of certain data fitting problems.

a) If $\boldsymbol{w}^*$ is a solution of this optimization problem, it means that $\|\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{y}\|_\infty \leq \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_\infty$ for any $\boldsymbol{w} \in \mathbb{R}^d$. But this does not imply that $\|\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{y}\|_\infty = 0$: this is only true when the linear system $\boldsymbol{X}\boldsymbol{w} = \boldsymbol{y}$ has a solution.

b) (Study of the problem (4))

i) Suppose that $(\boldsymbol{w}^*, t^*)$ is a solution of problem (4). Then, t^* is the smallest nonnegative value such that

$$-t^* \leq \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}^* - y_i \leq t^* \ \forall i \ \Leftrightarrow \ t^* \geq \max_{1\leq i\leq d} |\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}^* - y_i| = \|\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{y}\|_\infty.$$

Since $\|\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{y}\|_\infty$ is nonnegative, it follows that $t^* = \|\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{y}\|_\infty$.

ii) Since $(t, \boldsymbol{w})$ satisfies the constraints of the problem, we have

$$-t \leq \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} - y_i \leq t \ \forall i \ \Leftrightarrow \ t \geq \max_{1\leq i\leq d} |\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} - y_i| = \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_\infty.$$

Now, since $(t^*, \boldsymbol{w}^*)$ is a solution of the problem, we have in particular that $t^* \leq t$. Combining this with the previous inequality as well as the result of question i), we obtain:

$$t^* \leq \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_\infty \leq t.$$

iii) In the previous question, we showed that $\|\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{y}\|_\infty = t^* \leq \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_\infty$ for every $\boldsymbol{w}$, from which we conclude that $\boldsymbol{w}^*$ is a solution of (3). Since $t^* = \|\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{y}\|_\infty$, then t^* is equal to the minimum value of the problem, i.e.

$$\boldsymbol{w}^* \in \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^d} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_\infty \quad \text{and} \quad t^* = \min_{\boldsymbol{w}\in\mathbb{R}^d} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_\infty.$$

NB: Students are expected to follow the idea behind the proof rather than its rigorous unfolding. The amount of math calculations is the most that will be required throughout the course.

c) Linear programs are convex optimization problems that can be solved very efficiently using existing solvers.