Linear Models and SVMs

Optimization for Machine Learning — Exercise #2

Monday 24th April, 2023

Part I: Theory

I.1. Linear model example

Exercise I.1 (Polynomial Curve Fitting)**.** Given a set of points and their targets $\{x_i, t_i\}_{i=1}^N$ so that for $i \in [N]$, $x_i \in \mathbb{R}$ and $t_i \in \mathbb{R}$, the *curve fitting problem* is loosely defined as finding a function $f\colon \mathbb{R} \to \mathbb{R}$ such that $f(x_i) \approx t_i$ for all $i \in [N]$.

In order to find such a function, we restrict ourselves to a set of parametrized functions $\mathcal{F}$: each function can be parametrized with a vector $\boldsymbol{w} \in \mathbb{R}^{d+1}$.

To quantify the problem further, in this exercise, we limit ourselves to *polynomial functions* of degree d for the set $\mathcal{F}$, and can therefore write

$$f(x, \boldsymbol{w}) = w_0 + w_1 x + \ldots + w_d x^d = \sum_{k=0}^{d} w_k x^k \tag{1}$$

Notice how f is *linear* in $\boldsymbol{w}$, the parameter. Such model is called a *linear model*.

With N samples, we defined the loss (or error, or energy) of our parameter as the point-wise square distance between its estimation and the target:

$$E(\boldsymbol{w}) = \frac{1}{2} \sum_{i=1}^{N} \left(f(x_i, \boldsymbol{w}) - t_i\right)^2 \tag{2}$$

1. Is the function E convex in $\boldsymbol{w}$? How to find the optimal parameter $\boldsymbol{w}^\star$ at which the loss is minimum?
2. Compute the gradient $\nabla_{\boldsymbol{w}} E(\boldsymbol{w}) = \begin{pmatrix} \partial_{w_0} E(\boldsymbol{w}) \\ \vdots \\ \partial_{w_d} E(\boldsymbol{w}) \end{pmatrix} \in \mathbb{R}^{d+1}$.
3. Show that the optimal parameter $\boldsymbol{w}^\star$ satisfies the following system of equation:

$$\forall k \in [d+1], \quad \sum_{j=0}^{d} A_{kj} w_j^\star = T_k,$$

where

$$A_{kj} = \sum_{i=1}^{N} (x_i)^{k+j}, \qquad T_k = \sum_{i=1}^{N} (x_i)^k t_i. \tag{3}$$

4. Is such a system of equation solvable? When / not?

Answers

1. Each term in E is convex in $\boldsymbol{w}$ as the composition of a linear function $\boldsymbol{w} \mapsto f(x, \boldsymbol{w})$ and the convex functions $z \mapsto (z-t)^2$. The function E is then convex as the sum of convex functions in $\boldsymbol{w}$.

   Therefore, the optimum $\boldsymbol{w}^\star$ can be found with the first-order optimality condition: $\boldsymbol{w}^\star$ satisfies $\nabla E(\boldsymbol{w}^\star) = 0$.

2. For $k \in [d+1]$, we write $\partial_k := \partial_{w_k}$. The gradient can then be computed as

$$\begin{aligned}
\partial_k E(\boldsymbol{w}) &= \partial_k \left(\frac{1}{2} \sum_{i=1}^{N} (f(x_i, \boldsymbol{w}) - t_i)^2 \right) &(4)\\
&= \sum_{i=1}^{N} (f(x_i, \boldsymbol{w}) - t_i) \partial_k f(x_i, \boldsymbol{w}) &(5)\\
&= \sum_{i=1}^{N} (f(x_i, \boldsymbol{w}) - t_i)(x_i)^k, &(6)\\
&= \sum_{i=1}^{N} \left(\sum_{j=0}^{d} w_j (x_i)^j - t_i \right) (x_i)^k, &(7)
\end{aligned}$$

   and $\nabla E(\boldsymbol{w}) = (\partial_k E(\boldsymbol{w}))_{k \in [d+1]}$.

3. From 1, the minimizer $\boldsymbol{w}^\star$ is found by solving the equation $\nabla E(\boldsymbol{w}^\star) = 0$, i.e., for all $k \in [d+1]$,

$$\begin{aligned}
\partial_k E(\boldsymbol{w}^\star) \stackrel{!}{=} 0 \implies & \sum_{i=1}^{N} \left(\sum_{j=0}^{d} w_j^\star (x_i)^j (x_i)^k - t_i (x_i)^k \right) = 0 &(8)\\
\implies & \sum_{j=0}^{d} \underbrace{\sum_{i=1}^{N} (x_i)^{j+k}}_{A_{kj}} w_j^\star = \underbrace{\sum_{i=1}^{N} (x_i)^k t_i}_{T_k}. &(9)
\end{aligned}$$

4. This linear system can be written as $A\boldsymbol{w} = T$, where $A = (A_{kj})_{k,j \in [d+1] \times [d+1]} \in \mathbb{R}^{(d+1) \times (d+1)}$, and $T = (T_k)_{k \in [d+1]}$. Therefore, there are several cases possible.

   - First case: $\operatorname{rank} A = d + 1$ (i.e. A is invertible). In this case, there is a unique solution $\boldsymbol{w}^\star = A^{-1} T$.

- Second case: $\operatorname{rank} A < d+1$. In this case,
  - if $T \in \operatorname{span} A$, there is an infinite number of solutions $\boldsymbol{w}^\star = \boldsymbol{w}_\mathrm{p} + \boldsymbol{w}_\mathrm{k}$, with $\boldsymbol{w}_\mathrm{p}$ any particular solution in $(\ker A)^\perp$ and $\boldsymbol{w}_\mathrm{k} \in \ker A$.
  - if $T \notin \operatorname{span} A$, there are 0 solutions.

  Remark: Even when A is invertible, it might be ill-conditioned (meaning its inverse is unstable to compute).

It is usual to add a *regularizer* to the objective, penalizing "complex" models. This also can help selecting a model when several models are solutions to the optimization problem.

One of the most common regularizer is the parameter squared-norm: with a *penalizer weight* $\lambda \in \mathbb{R}_+$, the Equation (2) is modified to give

$$E_\lambda(\boldsymbol{w}) = E(\boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 = \frac{1}{2}\sum_{i=1}^{N}\left(f(x_i, \boldsymbol{w}) - t_i\right)^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|^2. \tag{10}$$

5. What is the role of λ?

6. Is E_λ convex?

7. Show that each component of the optimal weight $\boldsymbol{w}_i^\star$ is now found by solving

$$\forall k \in [d+1], \quad \sum_{j=0}^{d} A_{kj} w_j + \lambda w_k = T_k,$$

   with A_{kj} and T_k defined as in Equation (3).

Answers

5. λ is the regularizer weight, and make the error favour weights with small ℓ_2 norm. The higher the λ, the stronger weights with high ℓ_2 norm are avoided.

6. E_λ is still convex (as long as $\lambda \geqslant 0$), as a sum of convex functions in $\boldsymbol{w}$.

7. The first order optimality condition still applies, and we find, similarly to 3,

$$\begin{aligned}
\partial_k E_\lambda(\boldsymbol{w}^\star) \overset{!}{=} 0 \implies & \sum_{i=1}^{N}\sum_{j=0}^{d} w_j (x_i)^{j+k} - \sum_{i=1}^{N} t_i (x_i)^k + \frac{\lambda}{2}\partial_k\|\boldsymbol{w}\|^2 = 0 \\
\implies & \sum_{j=0}^{d} \underbrace{\sum_{i=1}^{N} (x_i)^{j+k}}_{A_{kj}} w_j + \lambda w_k = \underbrace{\sum_{i=1}^{N} t_i (x_i)^k}_{T_k}.
\end{aligned}$$

   Remark: Now, the system can be written $(A + \lambda I_{d+1})\boldsymbol{w}^\star = T$. Even if A is singular, $(A + \lambda I_{d+1})$ will be invertible, making the problem well-posed.

Matrix expression It is sometimes preferable to deal with vector and matrices, rather than scalar expressions. When the model is linear in $\boldsymbol{w}$, it is possible to express it as a *linear product* between a matrix and a vector. The expression in Equation (1) can be thought as a dot product between w and the vector of powers of x, that define as $\boldsymbol{\phi}(x) := \begin{pmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^d \end{pmatrix}$, so that

$$f(x, \boldsymbol{w}) = \boldsymbol{w}^\top \boldsymbol{\phi}(x).$$

Stacking all the N examples in a matrix, and denoting $\boldsymbol{\phi}_i := \boldsymbol{\phi}(x_i)$, we define

$$\Phi := \begin{pmatrix} | & | & & | \\ \boldsymbol{\phi}_1 & \boldsymbol{\phi}_2 & \cdots & \boldsymbol{\phi}_N \\ | & | & & | \end{pmatrix} \in \mathbb{R}^{(d+1)\times N}$$

and can therefore compute the model *on the whole dataset* in one expression: $\boldsymbol{y}(\boldsymbol{w}) = \Phi^\top \boldsymbol{w} \in \mathbb{R}^N$. Each entry i of $\boldsymbol{y}$ corresponds to a different sample x_i. Then, stacking the targets into a vector $\boldsymbol{t} \in \mathbb{R}^N$, the error function (2) can equivalently written as

$$E(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{y}(\boldsymbol{w}) - \boldsymbol{t}\|^2,$$

and the regularized error as

$$E_\lambda(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{y}(\boldsymbol{w}) - \boldsymbol{t}\|^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|^2$$

8. Show that $\nabla E_\lambda(\boldsymbol{w}) = \Phi(\Phi^\top \boldsymbol{w} - \boldsymbol{t}) + \lambda \boldsymbol{w}$, so that $\boldsymbol{w}^\star$ solves the linear equation

$$(\Phi\Phi^\top + \lambda I_{d+1})\boldsymbol{w}^\star = \Phi \boldsymbol{t}.$$

Answer

8. Since the unregularized error E is obtained with $\lambda = 0$, we can focus on the regularized objective E_λ, for $\lambda \geqslant 0$.

   We compute the gradient using the product rule from Exercise Sheet #1:

$$\begin{aligned}\nabla_{\boldsymbol{w}} E_\lambda(\boldsymbol{w}) &= \nabla_{\boldsymbol{w}}\left(\frac{1}{2}\|\boldsymbol{y}(\boldsymbol{w}) - \boldsymbol{t}\|^2\right) + \nabla_{\boldsymbol{w}}\left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2\right) \\ &= J_{\boldsymbol{y}}(\boldsymbol{w})^\top(\boldsymbol{y}(\boldsymbol{w}) - \boldsymbol{t}) + \lambda \boldsymbol{w} \qquad (\boldsymbol{y}(\boldsymbol{w}) = \Phi^\top \boldsymbol{w} \implies J_{\boldsymbol{y}}(\boldsymbol{w}) = \Phi^\top) \\ &= \Phi(\Phi^\top \boldsymbol{w} - \boldsymbol{t}) + \lambda \boldsymbol{w}.\end{aligned}$$

   Therefore, the first-order optimality condition can be written

$$\nabla E_\lambda(\boldsymbol{w}^\star) = 0 \implies (\Phi\Phi^\top + \lambda I_{d+1})\boldsymbol{w}^\star = \Phi \boldsymbol{t}.$$

I.2. Subgradients

When a convex loss function $E\colon \mathbb{R}^d \to \mathbb{R}$ is not differentiable, its *subgradient* can be used. It is defined as the set, for $x \in \mathbb{R}^d$,

$$\partial E(x) = \{g \in \mathbb{R}^d \mid \forall y \in \mathbb{R}^d,\ E(y) \geqslant E(x) + \langle g, y - x\rangle\}.$$

If E is differentiable at x, then $\partial E(x) = \{\nabla E(x)\}$.

For instance, for $E\colon \mathbb{R} \to \mathbb{R}$, $x \mapsto E(x) = |x|$, E is differentiable at any $x \neq 0$, with gradient -1 on $(-\infty, 0)$ and 1 on $(0, +\infty)$.

At $x = 0$, we compute, for any $y \in \mathbb{R}$ and $g \in \mathbb{R}$:

$$\begin{aligned} E(y) \geqslant E(0) + \langle g, y - 0\rangle &\iff |y| \geqslant \langle g, y\rangle \\ &\iff |y| \geqslant gy \end{aligned}$$

This condition has to be true for *any* $y \in \mathbb{R}$. This is only true when $g \in [-1, 1]$. Therefore,

$$\partial E(x) = \begin{cases} \{-1\} & \text{if } x < 0 \\ [-1, 1] & \text{if } x = 0 \\ \{1\} & \text{if } x > 0 \end{cases}$$

Geometrically, this can be interpreted as having, for the absolute value at the origin, any lines with slope between -1 and 1 lower-bounding the graph of the function.

Exercise I.2. Let $E\colon \mathbb{R} \to \mathbb{R}$, $x \mapsto E(x) = \max(0, 1 - x)$.

1. Where is E differentiable?
2. Show that $\partial E(x) = \begin{cases} \{-1\} & \text{if } x < 1 \\ [-1, 0] & \text{if } x = 1 \\ \{0\} & \text{if } x > 1 \end{cases}$

Answer

1. E is differentiable on $(-\infty, 1)$ and on $(1, +\infty)$, but not at 1.
2. Since E is differentiable on $(-\infty, 1)$, $\partial E(x) = \{\nabla E(x)\} = \{-1\}$ for $x \in (-\infty, 1)$. Likewise, $\partial E(x) = \{0\}$ for $x \in (1, +\infty)$.

   For $x = 1$, we are looking for $g \in \mathbb{R}$ such that, for all $y \in \mathbb{R}$,

   $$E(y) \geqslant E(1) + g \cdot (y - 1).$$

   Assuming such g exists, it necessarily satisfies, for all $y \in \mathbb{R}$,

   $$\max(0, 1 - y) \geqslant g \cdot (y - 1) \tag{1}$$

- For $y = 1$, g can be arbitrary (since we get $0 = g \cdot 0$ in this case, which is true for all g).
- For $y > 1$, the necessary condition become

$$0 \geqslant g(y-1) \implies 0 \geqslant g \quad \text{since } y-1>0$$

- For $y < 1$, the necessary condition becomes

$$1 - y \geqslant g(y-1) \implies -1 \leqslant g$$

Since the condition (1) has to be true for *all* $y \in \mathbb{R}$, g has to satisfy $-1 \leqslant g \leqslant 0$.

One verifies that taking $g \in [-1, 0]$ always satisfies (1).

Therefore, $\partial E(x) = \begin{cases} \{-1\} & \text{if } x < 1, \\ [-1, 0] & \text{if } x = 1, \\ \{0\} & \text{if } x > 1. \end{cases}$

Part II: Programming

Exercise II.1 (Model fitting)**.** This exercise implements some results found in Exercise I.1.

1. **Generation of the target.** In this toy example, we generate the N points ourselves. The true target t_i will be sinusoidal, with some noise, i.e. $t_i = \sin(2\pi x_i) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. The different scales (for σ,x_i) are given as $\sigma = 0.1$, and $x_i \sim \mathcal{U}([0, 1])$, uniform distribution on the segment $[0, 1]$.

   The generation of the data is performed by the function `gen_sin_data` in the file `ex02/utils.py`.

2. Implement the parametrization function (1) as `f(x,w)`, where the dimensions d is implied by the size of `w`.

3. Implement the error function E defined in (2), and its gradient $\nabla E(w)$.

4. Find $\boldsymbol{w}^\star$, either by
   a) gradient descent; or
   b) solving the linear system of equations (3).