Machine Learning: Optimization 1

Hao Tang

January 31, 2024
• For mean-squared error

\[ L = \sum_{i=1}^{n} (w^\top x_i - y_i)^2 = \|Xw - y\|_2^2, \]  

(1)

we know that

\[ w^* = (X^\top X)^{-1} X^\top y \]  

(2)

is the solution of \( \nabla_w L = 0 \).

• How do we know \( w^* \) is the optimal point?
• For log loss

$$L = \sum_{i=1}^{n} \log \left( 1 + \exp(-y_i w^T \phi(x_i)) \right)$$

we cannot even solve $\nabla_w L = 0$.

• How do we find the optimal solution?

• Could we find an approximate solution?
Convex optimization

\[ \frac{\partial L}{\partial w} = 0 \]

At the point \( w^* \), the gradient of the function \( L \) is zero, indicating a minimum or maximum point in the convex optimization problem.
Optimization

• Suppose $f : \mathbb{R}^d \to \mathbb{R}$.

• The goal is solve

\[
\min_x f(x).
\] (4)

• We want to find $x^*$ such that $f(x^*) = \min_x f(x)$.

• The point $x^*$ is called the optimal solution or the minimizer of $f$.

• There might not be a minimizer or there might have many, not just one. (In most case, we are content with finding one.)
Optimization

- Suppose $f : \mathbb{R}^d \to \mathbb{R}$.
- The goal is solve
  \[
  \min_x f(x).
  \tag{4}
  \]
- It means $\min_x f(x) \leq f(y)$ for any $y$. 

The point $x^*$ is called the optimal solution or the minimizer of $f$. There might not be a minimizer or there might have many, not just one. (In most cases, we are content with finding one.)
Optimization

• Suppose $f : \mathbb{R}^d \to \mathbb{R}$.

• The goal is solve

$$\min_x f(x).$$

(4)

• It means $\min_x f(x) \leq f(y)$ for any $y$.

• We want to find $x^*$ such that $f(x^*) = \min_x f(x)$.

• The point $x^*$ is called the **optimal solution** or the **minimizer** of $f$. 
Optimization

• Suppose $f : \mathbb{R}^d \to \mathbb{R}$.

• The goal is solve

$$\min_x f(x).$$  \hfill (4)

• It means $\min_x f(x) \leq f(y)$ for any $y$.

• We want to find $x^*$ such that $f(x^*) = \min_x f(x)$.

• The point $x^*$ is called the **optimal solution** or the **minimizer** of $f$.

• There might not be a minimizer or there might have many, not just one. (In most case, we are content with finding one.)
Convex functions

A function $f$ is **convex** if

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y),$$

(5)

for every $x, y$, and $0 \leq \alpha \leq 1$. 
\[ f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y) \]
\[ \alpha f(x) + (1 - \alpha) f(y) \leq f(\alpha x + (1 - \alpha) y) \]
\[ \alpha x + (1 - \alpha)y \leq f(\alpha x + (1 - \alpha)y) \]
\[ \alpha x + (1 - \alpha)y \]
\[ \alpha f(x) + (1 - \alpha) f(y) \leq f(\alpha x + (1 - \alpha)y) \]
Properties of convex functions

If $f$ is convex, then

$$f(x) \geq f(y) + \nabla f(y)^T (x - y), \quad (6)$$

for any $x$ and $y$. 
Properties of convex functions

If $f$ is convex, then

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y),$$

for any $x$ and $y$.

Proof:

$$f(\alpha x + (1 - \alpha) y) \leq \alpha f(x) + (1 - \alpha) f(y)$$

$$\alpha f(y) + f(y + \alpha (x - y)) - f(y) \leq \alpha f(x)$$

$$f(y) + \frac{f(y + \alpha (x - y)) - f(y)}{\alpha} \leq f(x)$$

$$f(y) + \nabla f(y)^\top (x - y) \leq f(x)$$
Supporting hyperplanes
Supporting hyperplanes
• Is the mean-squared error

\[ L = \| Xw - y \|_2^2 \]  

convex in \( w \)?

• The definition itself is not always easy to use for checking convexity.
A sufficient condition: Second derivative

• A matrix $H$ is positive semidefinite if $v^\top H v \geq 0$ for any $v$.

• If the Hessian of $f$ exists and is positive semidefinite everywhere, then $f$ is convex.
Convexity of squared distance

- The squared distance $\ell(s) = (s - s')^2$ is convex in $s$. 
Convexity of squared distance

- The squared distance $\ell(s) = (s - s')^2$ is convex in $s$.

$$\frac{\partial^2 \ell}{\partial s^2} = 2 \geq 0$$ (8)
Convexity of the $\ell_2$ norm

• Show that $f(x) = \|x\|_2^2$ is convex in $x$. 
Convexity of the $\ell_2$ norm

- Show that $f(x) = \|x\|_2^2$ is convex in $x$.

\[
\frac{\partial^2 \ell}{\partial x_i \partial x_j} = 0 \quad \frac{\partial^2 \ell}{\partial x_i^2} = 2 \tag{9}
\]
Affine transform preserves convexity

- If $f$ is convex, then $g(x) = f(Ax + b)$ is also convex.
**Affine transform preserves convexity**

- If $f$ is convex, then $g(x) = f(Ax + b)$ is also convex.

\[
g(\alpha x + (1 - \alpha)y) = f(\alpha(Ax + b) + (1 - \alpha)(Ay + b)) \\
\leq \alpha f(Ax + b) + (1 - \alpha)f(Ay + b) = \alpha g(x) + (1 - \alpha)g(y)
\]
Nonnegative weighted sum of convex functions

• If $f_1, \ldots, f_k$ are convex, then $f = \beta_1 f_1 + \cdots + \beta_k f_k$ is also convex when $\beta_1, \ldots, \beta_k \geq 0$
Nonnegative weighted sum of convex functions

• If \( f_1, \ldots, f_k \) are convex, then \( f = \beta_1 f_1 + \cdots + \beta_k f_k \) is also convex when \( \beta_1, \ldots, \beta_k \geq 0 \)

\[
f(\alpha x + (1 - \alpha)y) = \beta_1 f_1(\alpha x + (1 - \alpha)y) + \cdots + \beta_k f_k(\alpha x + (1 - \alpha)y) \leq \beta_1 \alpha f_1(x) + \beta_1(1 - \alpha)f(y) + \cdots + \beta_k \alpha f_k(x) + \beta_k(1 - \alpha)f_k(y) = \alpha(\beta_1 f_1(x) + \cdots + \beta_k f_k(x)) + (1 - \alpha)(\beta_1 f_1(y) + \cdots + \beta_k f_k(y)) = \alpha f(x) + (1 - \alpha)f(y)
\]
Convexity of MSE

• The mean-squared error is

\[ L = \sum_{i=1}^{n} (w^\top x_i - y_i)^2 = \|Xw - y\|_2^2. \]  

(16)

• We know that the squared distance is convex.

• Use the affine transform and nonnegative weighted sum to obtain the mean-squared error.
Optimality condition

If $f$ is convex and

$$\nabla f(x^*) = 0 \quad (17)$$

at $x^*$, then $x^*$ is the minimizer of $f$. 

Optimality condition

If $f$ is convex and

$$\nabla f(x^*) = 0$$

at $x^*$, then $x^*$ is the minimizer of $f$.

Proof: Suppose $\nabla f(x^*) = 0$. For any $x$,

$$f(x) \geq f(x^*) + (x - x^*)^\top \nabla f(x^*) = f(x^*).$$
Optimal solution of MSE

• The mean-squared error is

$$L = \sum_{i=1}^{n} (w^T \phi(x_i) - y_i)^2 = \|Xw - y\|_2^2. \quad (19)$$

• The solution to $\nabla_w L = 0$ is $w^* = (X^T X)^{-1} X^T y$.

• Because $L$ is convex in $w$, $w^*$ is a minimizer of $L$. 
Convexity of log loss

• The log loss in the binary case is

\[ L = \sum_{i=1}^{N} \log \left( 1 + \exp(-y_i w^\top x_i) \right). \]  

(20)

• We just need to show \( \ell(s) = \log(1 + \exp(-s)) \) is convex in \( s \).

• Use affine transform and nonnegative weighted sum to obtain the log loss.
\[
\frac{\partial \ell}{\partial s} = \frac{-\exp(-s)}{1 + \exp(-s)} = \frac{1}{1 + \exp(-s)} - 1
\]  

(21)

\[
\frac{\partial^2 \ell}{\partial s^2} = \frac{1}{1 + \exp(-s)} \frac{\exp(-s)}{1 + \exp(-s)} = \frac{1}{1 + \exp(-s)} \left(1 - \frac{1}{1 + \exp(-s)}\right) \geq 0
\]  

(22)
A function $f$ is \textbf{strictly convex} if

$$f(\alpha x + (1 - \alpha)y) < \alpha f(x) + (1 - \alpha)f(y),$$

for every $x \neq y$, and $0 \leq \alpha \leq 1$.  

(23)
Properties of strictly convex functions

• If $f$ is strictly convex, then

$$f(x) > f(y) + \nabla f(y)^T (x - y),$$  \hspace{1cm} (24)

for any $x \neq y$. 

Properties of strictly convex functions

• If $f$ is strictly convex, then

$$f(x) > f(y) + \nabla f(y)^T(x - y),$$

for any $x \neq y$.

• A matrix $H$ is positive definite if $\nu^T H \nu > 0$ for any $\nu \neq 0$.

• If the Hessian of $f$ is positive definite, then $f$ is strictly convex.
A strictly convex function $f$ has a unique minimizer.
A strictly convex function $f$ has a unique minimizer.

Proof: Suppose $x^*$ is a minimizer of $f$, i.e., $\nabla f(x^*) = 0$. Since $f$ is strictly convex,

$$f(x) > f(y) + \nabla f(y)^T (x - y)$$ (25)

for any $x \neq y$. In particular, if we let $y = x^*$