Machine Learning Exercise Sheet 06

Optimization

Exercise sheets consist of two parts: homework and in-class exercises. You solve the homework exercises on your own or with your registered group and upload it to Moodle for a possible grade bonus. The in-class exercises will be solved and explained during the tutorial. You do not have to upload any solutions of the in-class exercises.

In-class Exercises

Problem 1: Prove or disprove whether the following functions $f : D \to \mathbb{R}$ are convex

a) $D = (1, \infty)$ and $f(x) = \log(x) - x^3$,

b) $D = \mathbb{R}^+$ and $f(x) = -\min(\log(3x+1), -x^4 - 3x^2 + 8x - 42)$,

c) $D = (-10, 10) \times (-10, 10)$ and $f(x, y) = y \cdot x^3 - y \cdot x^2 + y^2 + y + 4$.

a) The second derivative of f is $\frac{d^2 f(x)}{dx^2} = \frac{d}{dx}\left(\frac{1}{x} - 3x^2\right) = -\frac{1}{x^2} - 6x$, which is negative on the given set D and therefore f is not convex.

b) Transform min to max

$$-\min\{\log(3x+1), -x^4 - 3x^2 + 8x - 42\} = \max\{-\log(3x+1), x^4 + 3x^2 - 8x + 42\}\,.$$

$\max(g_1(x), g_2(x))$ is convex if both g_1 and g_2 are convex on $D = \mathbb{R}^+$ (see Exercise Sheet 6, Problem 1c). $g_1(x) = -\log(3x+1)$ is convex since the second derivative is positive on $\mathbb{R}^+$:

$$\frac{d^2}{dx^2}\left(-\log(3x+1)\right) = \frac{d}{dx}\left(-\frac{3}{3x+1}\right) = \frac{9}{(3x+1)^2} > 0$$

$g_2(x) = x^4 + 3x^2 - 8x + 42$ is also convex:

$$\frac{d^2}{dx^2}\left(x^4 + 3x^2 - 8x + 42\right) = \frac{d}{dx}\left(4x^3 + 6x - 8\right) = 12x^2 + 6 > 0$$

Thus f is convex.

c) For the function $f(x, y)$ to be convex (on D) it has to hold for all $x_1, x_2, y \in D$ and $\lambda \in (0, 1)$ that

$$\lambda f(x_1, y) + (1-\lambda) f(x_2, y) \geq f(\lambda x_1 + (1-\lambda)x_2, y)\,.$$

It does not hold in our case, consider $y = 1, x_1 = -4, x_2 = 0$ and $\lambda = 0.5$:

$$0.5 f(-4, 1) + 0.5 f(0, 1) = 0.5 \cdot (-74) + 0.5 \cdot 6 = -34$$
$$f\big(0.5 \cdot (-4) + 0.5 \cdot 0, 0.5 \cdot 1 + 0.5 \cdot 1\big)\big) = f(-2, 1) = -6 \;\; > -34$$

Thus $f(x, y)$ is not convex.

Upload a single PDF file with your homework solution to Moodle by 16.12.2020, 23:59 CET. We recommend to typeset your solution (using LaTeX or Word), but handwritten solutions are also accepted. If your handwritten solution is illegible, it won't be graded and you waive your right to dispute that.

Problem 2: Prove that the following function (the loss function of logistic regression) $f : \mathbb{R}^d \to \mathbb{R}$ is convex:

$$f(\boldsymbol{w}) = -\ln p\big(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{X}\big) = -\sum_{i=1}^{N} \big(y_i \ln \sigma(\boldsymbol{w}^T\boldsymbol{x}_i) + (1-y_i)\ln(1-\sigma(\boldsymbol{w}^T\boldsymbol{x}_i))\big) \ .$$

First, let's simplify the above expression. For this we will need the following two facts

$$\sigma(z) = \frac{1}{1+e^{-z}} = \frac{e^z}{1+e^z} \qquad \text{and} \qquad 1-\sigma(z) = \sigma(-z) = \frac{1}{1+e^z},$$

which implies that

$$\ln\sigma(z) = \ln\left(\frac{e^z}{1+e^z}\right) = z - \ln(1+e^z) \qquad \text{and} \qquad \ln(1-\sigma(z)) = -\ln(1+e^z).$$

Plugging this into the definition of the loss function we obtain

$$\begin{aligned}
f(\boldsymbol{w}) &= -\sum_{i=1}^{N} \big(y_i \ln \sigma(\boldsymbol{w}^T\boldsymbol{x}_i) + (1-y_i)\ln(1-\sigma(\boldsymbol{w}^T\boldsymbol{x}_i))\big) \\
&= -\sum_{i=1}^{N} \Big(y_i \left(\boldsymbol{w}^T\boldsymbol{x}_i - \ln(1+e^{\boldsymbol{w}^T\boldsymbol{x}_i})\right) - (1-y_i)\ln(1+e^{\boldsymbol{w}^T\boldsymbol{x}_i}))\Big) \\
&= \sum_{i=1}^{N} \left(-y_i(\boldsymbol{w}^T\boldsymbol{x}_i) + \ln(1+e^{\boldsymbol{w}^T\boldsymbol{x}_i})\right)
\end{aligned}$$

We know that $\boldsymbol{w}^T\boldsymbol{x}_i$ is a convex (and concave) function of $\boldsymbol{w}$. Therefore, the first term $-y_i(\boldsymbol{w}^T\boldsymbol{x}_i)$ is also convex.

Now, if we show that $\ln(1+e^z)$ is a <u>nondecreasing</u> and <u>convex</u> function of z on $\mathbb{R}$, we will be able to use the convexity preserving operations to prove that $f(\boldsymbol{w})$ is convex.

The first derivative of $\ln(1+e^z)$ is

$$\frac{d}{dz}\ln(1+e^z) = \frac{e^z}{1+e^z} = \sigma(z),$$

which is positive for all $z \in \mathbb{R}$, which means that $\ln(1+e^z)$ is an nondecreasing function.

The second derivative is

$$\frac{d^2}{dz^2}\ln(1+e^z) = \frac{d}{dz}\sigma(z) = \sigma(z)\sigma(-z),$$

which is also positive for all $z \in \mathbb{R}$, which means that $\ln(1+e^z)$ is a convex function.

Using the following two facts

1. Sum of convex functions is convex
2. Composition of a convex function with a convex nondecreasing function is convex

we can verify that $f(\boldsymbol{w})$ is indeed convex in $\boldsymbol{w}$ on $\mathbb{R}^d$.

Problem 3: Prove that for differentiable convex functions each local minimum is a global minimum. More specifically, given a differentiable convex function $f : \mathbb{R}^d \to \mathbb{R}$, prove that

a) if $\boldsymbol{x}^*$ is a local minimum, then $\boldsymbol{\nabla} f(\boldsymbol{x}^*) = \mathbf{0}$.

b) if $\boldsymbol{\nabla} f(\boldsymbol{x}^*) = \mathbf{0}$, then $\boldsymbol{x}^*$ is a global minimum.

We will show that if the gradient at some point $\boldsymbol{x}^*$ is not equal to zero, then this point cannot be a local optimum — we could simply follow the direction of the negative gradient and end up in a point with a lower value of the function.

More formally, suppose $\nabla f(\boldsymbol{x}^*) \neq \mathbf{0}$ for some $\boldsymbol{x}^*$. Then by Taylor's theorem for a sufficiently small $\varepsilon > 0$ we get

$$\begin{aligned} f(\boldsymbol{x}^* - \varepsilon \boldsymbol{\nabla} f(\boldsymbol{x}^*)) &= f(\boldsymbol{x}^*) - (\varepsilon \boldsymbol{\nabla} f(\boldsymbol{x}^*))^T \boldsymbol{\nabla} f(\boldsymbol{x}^*) + O(\varepsilon^2 \|\boldsymbol{\nabla}^2 f(\boldsymbol{x}^*)\|_2^2) \\ &= f(\boldsymbol{x}^*) - \varepsilon ||\boldsymbol{\nabla} f(\boldsymbol{x}^*)||_2^2 + O(\varepsilon^2 \|\boldsymbol{\nabla} f(\boldsymbol{x}^*)\|_2^2) \\ &< f(\boldsymbol{x}^*) \end{aligned}$$

Which means that $\boldsymbol{x}^*$ is not a local optimum. Therefore, the gradient must be equal to zero for any local optimum $\boldsymbol{x}^*$.

We will prove (b) using the first-order criterion for convexity:

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + (\boldsymbol{y} - \boldsymbol{x})^T \boldsymbol{\nabla} f(\boldsymbol{x}).$$

If we plug in $\boldsymbol{x}^*$ and use the fact that $\boldsymbol{\nabla} f(\boldsymbol{x}^*) = 0$ we get: $f(\boldsymbol{y}) \geq f(\boldsymbol{x}^*)$ for all $\boldsymbol{y}$, meaning $\boldsymbol{x}^*$ is a global minimum.