Assignment 3: Duality and Backpropagation

AIGS/CSED515 Machine Learning
Instructor: Jungseul Ok
jungseul@postech.ac.kr

Due: 20:00pm Nov 5 , 2020

Remarks

- Group study and open discussion via LMS board are encouraged, however, assignment that your hand-in must be **of your own work**, and **hand-written** unless you're asked a coding task.
- Submit a scanned copy of your answer on LMS online in **a single PDF file**, to which you also print and add your code if apply, i.e., no zip file, just a single PDF containing everything.
- Delayed submission may get some penalty in score: 5% off for delay of $0 \sim 4$ hours; 20% off for delay of $4 \sim 24$ hours; and delay longer than 24 hours will not be accepted.

1. [8pt; Practice KKT] We learned how to make Lagrange dual problem from a constrained optimization problem, in which Karush-Kuhn-Tucker (KKT) conditions provide a set of necessary conditions on the optimization solution, c.f.,

https://www.cs.cmu.edu/ ggordon/10725-F12/slides/16-kkt.pdf
http://www.stat.cmu.edu/ ryantibs/convexopt-F16/scribes/kkt-scribed.pdf

Practice the use of KKT conditions (stationary, feasibility, and complementary slackness) and duality with the following optimization problems:

(a) [7pt] Write Lagrange function for the following minimization problem with dual variable μ. Write the stationary condition, and extract the property of (w_1, w_2) to guarantee the existence of μ. Combining the above observation and the feasibility condition, find the properties on (w_1, w_2). Check if $\mu > 0$ with (w_1, w_2) verifying all the above properties. State the complementary slackness condition, and specify optimal solution (w_1^*, w_2^*) from the KKT conditions.

$$\begin{aligned} \underset{w_1, w_2}{\text{minimize}} \quad & \sqrt{w_1^2 + w_2^2} \\ \text{subject to} \quad & w_1 + 2w_2 \geq 5 \end{aligned}$$

(b) [1pt] Find the optimal solution of the following optimization:

$$\begin{aligned} \underset{w_1, w_2}{\text{maximize}} \quad & \frac{1}{\sqrt{w_1^2 + w_2^2}} \\ \text{subject to} \quad & w_1 + 2w_2 \geq 5 \end{aligned}$$

2. [15pt; Backpropagation] We want to train a simple deep neural network $f_{\boldsymbol{w}}(x)$ with $\boldsymbol{w} = (w_1, w_2, w_3)^\top \in \mathbb{R}^3$ and $x \in \mathbb{R}$, defined as:

$$f_{\boldsymbol{w}}(x) := w_3 \sigma_2(w_2 \sigma_1(w_1 x))$$

where $\sigma_1(u) = \sigma_2(u) = \frac{1}{1+\exp(-u)}$, i.e., sigmoid activation. You may denote $x_1 := w_1 x$ and $x_2 := w_2 \sigma_1(x_1)$ for notational convenience.

(a) [1pt] Illustrate a directed acyclic graph corresponding to the computation of $f_{\boldsymbol{w}}(x)$.

(b) [2pt] Compute $\frac{\partial \sigma_1}{\partial u}$ and provide the answer in two different forms: (i) using only u and the exponential functions; and (ii) using only $\sigma_1(u)$.

(c) [2pt] Describe briefly what is meant by a *forward pass* and a *backward pass*?

(d) [2pt] Compute $\frac{\partial f_{\boldsymbol{w}}}{\partial w_3}$. Which result should we retain from the forward pass in order for efficiently computing this derivative?

(e) [3pt] Compute $\frac{\partial f_{\boldsymbol{w}}}{\partial w_2}$ using the second option in Problem 2b. Which results should we retain from the forward pass in order for efficiently computing this derivative?

(f) [5pt] Compute $\frac{\partial f_{\boldsymbol{w}}}{\partial w_1}$ using the second option in Problem 2b. Which results should we retain from the forward pass in order for efficiently computing this derivative? In what order should we compute the derivatives $\frac{\partial f_{\boldsymbol{w}}}{\partial w_1}$, $\frac{\partial f_{\boldsymbol{w}}}{\partial w_2}$ and $\frac{\partial f_{\boldsymbol{w}}}{\partial w_3}$ in order for maximizing computational efficiency? How is this order related to the forward pass?