Assignment 4, Written Part

Please turn in the answers to this written part of assignment 4 by either

- Typesetting your answers inline with LaTeX (.tex file provided).
- Write out your answers with tablet / stylus, and submit the annotated pdf.
- Print out the assignment, write answers by hand, and scan / photograph your work. The image must be clearly legible, and all pages must be combined into one file.
- For the written response questions, clearly justify all conclusions to receive full credit. A correct answer with no supporting work will receive no credit.

1. Consider a linear model for classification in which we use a logistic activation, but instead of cross-entropy loss, we use squared error loss. Assume a 1-dimensional input x, a single weight w and an outcome $y_i \in \{0, 1\}$. We will ignore the intercept term.

$$\begin{aligned} a_i &= w x_i \\ p_i &= \text{logistic}\,(a_i) \\ l_i &= (y_i - p_i)^2 \end{aligned}$$

Recall that logistic $(u) = \frac{1}{1+e^{-u}}$. Calculate the following:

a. $\dfrac{dl_i}{dp_i}$

b. $\dfrac{dp_i}{da_i}$, as a function of a_i

c. $\dfrac{dp_i}{da_i}$, rewritten as a function of p_i only

d. $\dfrac{da_i}{dw}$

e. $\dfrac{dl_i}{dw}$

f. Assume that $y_i = 1$. What is $\lim_{p \to 0} \frac{dl_i}{dw}$? Is this good or bad for learning? Explain why.

2. Consider a linear model for classification based on the hinge loss, with a penalty for weight magnitude. This is the basic support vector machine (don't worry if you haven't studied it). Unlike question 1, we will now assume that $y_i \in \{-1, 1\}$. Again, assume a single input variable x_i, and ignore the intercept term.

$$a_i = wx_i$$
$$l_i = \max(0, 1 - y_i a_i) + w^2$$

Calculate the following:

a. $\dfrac{\partial l_i}{\partial a_i}$ [Note: This technically should be a subgradient. Only worry about the two cases of $y_i a_i < 1$ and $y_i a_i > 1$. Don't worry about the non-differentiable point where $y_i a_i = 1$.]

b. $\dfrac{dl_i}{dw}$ [Again, there are two cases.]

c. Assume that $y_i = 1$. What is update rule for w for stochastic gradient descent?

d. Contrast this rule with the update rule for the perceptron.

3.

$$y = x + \frac{1}{wx+b}$$

Draw the computation graph for calculating y from x, w and b, Fill in the blanks for the reverse mode AD table at $x = 0.3, w = 0.5, b = 0.1$

Part 1 - Computation Graph

Forward Primal Trace

v_{-2}	$= x$	$= 0.3$
v_{-1}	$= w$	$= 0.5$
v_0	$= b$	$= 0.1$
v_1	$= v_{-2}v_{-1}$	$= 0.15$
v_2	$= v_1 + v_0$	$= 0.25$
v_3	$= \frac{1}{v_2}$	$= 4$
v_4	$= v_{-2} + v_3$	$= 4.3$
y	$= v_4$	$= 4.3$

Part 2 - Reverse Adjoint Trace

$\bar{v}_{-2}$	$=$	$=$
$\bar{v}_{-1}$	$=$	$=$
$\bar{v}_0$	$=$	$=$
$\bar{v}_1$	$=$	$=$
$\bar{v}_2$	$=$	$=$
$\bar{v}_3$	$=$	$=$
$\bar{v}_4$	$=$	$= 1$