Assignment 4, Written Part

Nicholas Renninger

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.

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

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

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

use chain rule:

$$\begin{aligned} \frac{dl_i}{dp_i} &= 2(y_i - p_i)\left(\frac{d}{dp_i}(y_i - p_i)\right) \\ &= 2(y_i - p_i)\,(-1) \\ &= \boxed{2(p_i - y_i)} \end{aligned}$$

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

use the chain rule:

$$\begin{aligned} \frac{dp_i}{da_i} &= -\frac{\frac{d}{da_i}(1 + e^{-a_i})}{(1 + e^{-a_i})^2} \\ &= -\frac{e^{-a_i}\frac{d}{da_i}(-a_i)}{(1 + e^{-a_i})^2} \\ &= \boxed{\frac{e^{-a_i}}{(1 + e^{-a_i})^2}} \end{aligned}$$

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

The inverse of the logistic function is the well known logit function: $\text{logit}(u) = \text{logistic}^{-1}(u) =$

$\log\left(\frac{p_i}{1-p_i}\right) = \log(p_i) - \log(1-p_i)$. Apply this to the result from above:

$$\begin{aligned}
\frac{dp_i}{da_i} &= \frac{e^{-a_i}}{(1+e^{-a_i})^2} \\
&= \frac{e^{-\log\left(\frac{p_i}{1-p_i}\right)}}{(1+e^{-\log\left(\frac{p_i}{1-p_i}\right)})^2} \\
&= \frac{e^{\log\left(\frac{1-p_i}{p_i}\right)}}{(1+e^{\log\left(\frac{1-p_i}{p_i}\right)})^2} \\
&= \frac{\left(\frac{1-p_i}{p_i}\right)}{(1+\left(\frac{1-p_i}{p_i}\right))^2} \\
&= \boxed{p_i(1-p_i)}
\end{aligned}$$

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

$$\frac{da_i}{dw} = \boxed{x_i}$$

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

Use the chain rule:

$$\begin{aligned}
\frac{dl_i}{dw} &= \frac{dl_i}{dp_i} \cdot \frac{dp_i}{da_i} \cdot \frac{da_i}{dw} \\
&= (2(p_i - y_i)) \cdot (p_i(1-p_i)) \cdot (x_i) \\
&= \boxed{2p_i x_i (p_i - y_i)(1-p_i)}
\end{aligned}$$

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.

$$\begin{aligned}
\lim_{p\to 0} \frac{dl_i}{dw} &= \lim_{p\to 0} (2p_i x_i (p_i - (1))(1-p_i)) \\
&= \lim_{p\to 0} (-x_i 2p_i (1-p_i)^2) \\
&= (-x_i 2(0)(1-(0))^2) \\
&= (0)
\end{aligned}$$

This is **good for learning**, as it means that when the classifier got the correct answer, there is no loss gradient, and thus the weights should not change.

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 $y_i a_i = 1$.]

Assume here that $y_i a_i > 1$:

$$\begin{aligned}\frac{dl_i}{da_i} &= \frac{d(\max(0, 1 - y_i a_i)) + w^2}{da_i} \\ &= \frac{d(0 + w^2}{da_i} \\ &= \boxed{0}\end{aligned}$$

Assume here that $y_i a_i < 1$:

$$\begin{aligned}\frac{dl_i}{da_i} &= \frac{d(\max(0, 1 - y_i a_i)) + w^2}{da_i} \\ &= \frac{d((1 - y_i a_i) + w^2}{da_i} \\ &= \boxed{-y_i}\end{aligned}$$

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

First we need $\dfrac{da_i}{dw} = x_i$. Then, using the chain rule:

$$\frac{dl_i}{dw} = \frac{dl_i}{da_i} \cdot \frac{da_i}{dw}$$

Assume here that $y_i a_i > 1$:

$$\begin{aligned}\frac{dl_i}{dw} &= \frac{dl_i}{da_i} \cdot \frac{da_i}{dw} \\ &= (0) \cdot (x_i) \\ &= \boxed{0}\end{aligned}$$

Assume here that $y_i a_i < 1$:

$$\begin{aligned}\frac{dl_i}{dw} &= \frac{dl_i}{da_i} \cdot \frac{da_i}{dw} \\ &= (-y_i) \cdot (x_i) \\ &= \boxed{-y_i x_i}\end{aligned}$$

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

The general rule for SGD is: $w_{new} = w_{old} - \eta \nabla l_i$. Thus, in this case as $\nabla l_i = \dfrac{dl_i}{dw}$.

Thus given $y_i = 1$:

$$\begin{aligned}\frac{dl_i}{dw} &= \begin{cases} 0, & \text{for } wx_i > 1 \\ -(1)x_i, & \text{for } wx_i < 1 \end{cases} \\ &= \begin{cases} 0, & \text{for } wx_i > 1 \\ -x_i, & \text{for } wx_i < 1 \end{cases}\end{aligned}$$

Therefore, we can re-write the SGD update rule:

$$w_{new} = \begin{cases} 0, & \text{for } w_{old} x_i > 1 \\ w_{old} + \eta x_i, & \text{for } w_{old} x_i < 1 \end{cases}$$

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

The perceptron update rule generally looks like the following: On a mistake, update as follows:

- Mistake on predicted label being true $\implies y_i = 1$: $w_{new} \leftarrow w_{old} + \eta x_i$
- Mistake on predicted label being false $\implies y_i = -1$: $w_{new} \leftarrow w_{old} - \eta x_i$

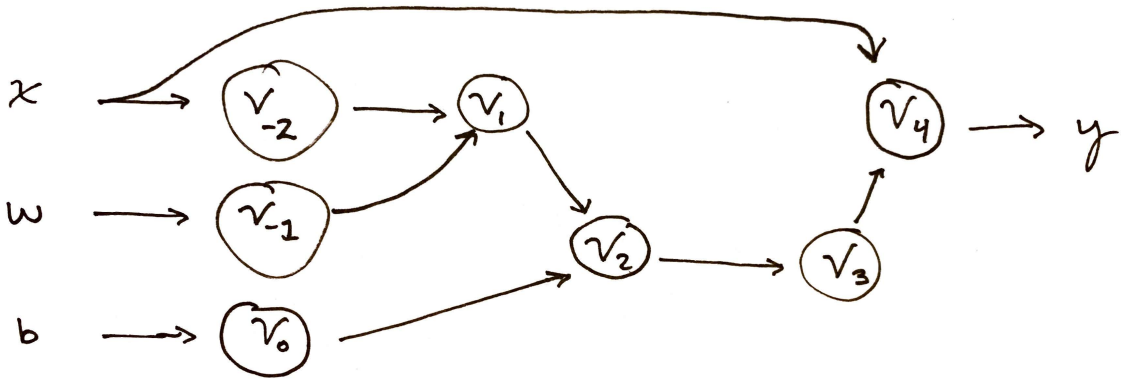
For the SVM, we basically computed the loss given that the predicted label was true, and we recovered the exact perceptron update rule! So under this scenario, the two are equivalent.

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

$\bar{v_{-2}}$	$= \bar{v_4}\ \frac{\partial v_4}{\partial v_{-2}} + \bar{v_1}\ \frac{\partial v_1}{\partial v_{-2}} =$ (1) (1) + (-4) (v_{-1}) = 1 + (-4) (0.5)	= -1
$\bar{v_{-1}}$	$= \bar{v_1}\ \frac{\partial v_1}{\partial v_{-1}} =$ (-4) (v_{-2}) = (-4) (0.3)	= -1.2
$\bar{v_0}$	$= \bar{v_2}\ \frac{\partial v_2}{\partial v_0} =$ (-4) (1)	= -4
$\bar{v_1}$	$= \bar{v_2}\ \frac{\partial v_2}{\partial v_1} =$ (-4) (1)	= -4
$\bar{v_2}$	$= \bar{v_3}\ \frac{\partial v_3}{\partial v_2} =$ (1) $(-1/(v_2))$ = -1 / 0.25	= -4
$\bar{v_3}$	$= \bar{v_4}\ \frac{\partial v_4}{\partial v_3} =$ (1) (1)	= 1
$\bar{v_4}$	=	= 1