LING572 Hw7 Solution

Q1 (12 points): Let $f'(x)$ denote the derivative of a function $f(x)$ w.r.t. the variable x.

(a) 2 pts: What does f'(x) intend to measure?
$\Rightarrow$ it measures the rate of change in f(x) w.r.t. the rate of change in x.

(b) 2 pts: Let $h(x) = f(g(x))$. What is $h'(x)$?
$\Rightarrow h'(x) = f'(g(x))g'(x)$

(c) 2 pts: Let $h(x) = f(x)g(x)$. What is $h'(x)$?
$\Rightarrow h'(x) = f'(x)g(x) + f(x)g'(x)$

(d) 3 pts: Let $f(x) = a^x$, where $a > 0$. What is $f'(x)$?
$\Rightarrow a^x ln(a)$

(e) 3 pts: Let $f(x) = x^{10} - 2x^8 + \frac{4}{x^2} + 10$. What is $f'(x)$?
$\Rightarrow 10x^9 - 16x^7 - \frac{8}{x^3}$

Q2 (18 points): The logistic function is $f(x) = \frac{1}{1+e^{-x}}$. The tanh function is $g(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.

(a) 6 pts: Prove that $f'(x) = f(x)(1 - f(x))$.
$\Rightarrow f'(x) = \frac{1}{(1+e^{-x})^2} * e^{-x} = \frac{e^{-x}}{(1+e^{-x})^2}$

$f(x)(1 - f(x)) = \frac{1}{1+e^{-x}} * \frac{1+e^{-x}-1}{1+e^{-x}} = \frac{e^{-x}}{(1+e^{-x})^2}$

(b) 6 pts: Prove that $g'(x) = 1 - g^2(x)$.

$\Rightarrow g'(x) = \frac{(e^x - e^{-x})'}{e^x + e^{-x}}$ - $\frac{e^x - e^{-x}}{(e^x + e^{-x})^2}(e^x + e^{-x})'$
$= \frac{e^x + e^{-x}}{e^x + e^{-x}} - \frac{(e^x - e^{-x})^2}{(e^x + e^{-x})^2}$
$= 1 - g^2(x)$

(c) 6 pts: Prove that $g(x) = 2f(2x) - 1$

$(e^x - e^{-x})(1 + e^{-2x}) = (e^x + e^{-x})(1 - e^{-2x}) = e^x - e^{-3x}$
$\Rightarrow \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{1 - e^{-2x}}{1 + e^{-2x}}$

$2f(2x)-1=\frac{2-1-e^{-2x}}{1+e^{-2x}}=\frac{1-e^{-2x}}{1+e^{-2x}}=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}=g(x).$

Thus, $g(x)=2f(2x)-1$.

Q3 (45 points): Let us denote the partial derivative of a multi-variate function f w.r.t. one of its variables, x, by f'_x or $\frac{df}{dx}$ or $\frac{\partial f}{\partial x}$.

(a) 15 free pts: refresh your memory about gradient, chain rule, etc.

(b) 3 pts: What is f'_x trying to measure?

$\Rightarrow$ the rate of change in f(x) when x changes while other variables remain constant.

(c) 3 pts: How do you calculate the gradient of f at a point z?

$\Rightarrow$ Suppose the input of f is a vector $x=(x_1,x_2,...,x_m)$. To find the gradient of f at a point $z=(z_1,z_2,...,z_m)$, just substitute x_i with z_i for every i in the vector $(\frac{df}{dx_1}, \ldots\ \frac{df}{dx_m})$.

(d) 5 pts: Suppose that $x=g(t)$ and $y=h(t)$ are differentiable functions of t and $z=f(x,y)$ is a differentiable function of x and y. How do you calculate $\frac{dz}{dt}$ using the chain rule of partial derivatives?

$\Rightarrow \frac{dz}{dt}=\frac{dz}{dx}\frac{dx}{dt}+\frac{dz}{dy}\frac{dy}{dt}$

(e) 6 pts: Let $f(x,y)=x^3+3x^2y+y^3+2x$.

What is f'_x? What is f'_y?

$\Rightarrow f'_x=3x^2+6xy+2$
$f'_y=3x^2+3y^2$

What is the gradient of $f(x,y)$ at point (1, 2)?

$\Rightarrow$ Just plug in (1, 2) into (f'_x,f'_y), which is $(3x^2+6xy+2, 3x^2+3y^2)$, so the answer is (17, 15).

(f) 3 pts: Let $z=\sum_{i=1}^{n} w_i x_i$. What is $\frac{dz}{dw_i}$?

$\Rightarrow \frac{dz}{dw_i}=x_i$

(g) 5 pts: Let $f(z) = \frac{1}{1+e^{-z}}$ and $z = \sum_{i=1}^{n} w_i x_i$.

What is $\frac{df}{dz}$?

$\Rightarrow \frac{df}{dz} = f(z)(1 - f(z))$

What is $\frac{df}{dw_i}$?

$\Rightarrow \frac{df}{dw_i} = f(z)(1 - f(z))x_i$

(h) 5 pts: Let $E(z) = \frac{1}{2}(t - f(z))^2$, $f(z) = \frac{1}{1+e^{-z}}$ and $z = \sum_{i=1}^{n} w_i x_i$. What is $\frac{dE}{dw_i}$?

$\Rightarrow \frac{dE}{dw_i} = -(t - f(z))f(z)(1 - f(z))x_i$

Q4 (25 points): The softmax function:

(a) 2 pts: The softmax function is a function that takes the input x and produces the output y. What is the type of x? What is the type of y?

⇒ Both x and y are vectors, and they have the same number of dimensions.

(b) 5 pts: In general where in NN is the softmax function used and why?

⇒ The softmax function is often used in the output layer of an NN, in order to turn a vector of real numbers into a probability distribution.

(c) 5 pts: What is the relationship between the softmax function and the sigmoid function?

⇒ The sigmoid function takes a scalar as input and produces a scalar value as output, where softmax takes a vector as input and produces a vector as output. The former can be seen as a special case of the latter for a classifier with only two classes.

(d) 7 pts: What is the relationship between the softmax function and the argmax function? When do you use softmax? When do you use argmax?

⇒ Both softmax and argmax take a vector as input and produce a vector as output. In softmax, all the elements in the output vector are in [0, 1], and add up to one. In argmax, all the elements in the output vector are 0 except one element is 1.

Softmax is differentiable and thus is used in training. That layer can be switched to argmax in inference if we want the system to output a single predicted value rather than a probability.

(e) 6 pts: If a vector x is [1, 2, 3, -1, -4, 0], what is the value of softmax(x)?

You can use the following python code to calcalute softmax:

```
np.exp(x) / np.sum(np.exp(x), axis=0)
```

softmax(x) $= [8.607859e{-}02, 2.339858e{-}01, 6.360395e{-}01, 1.164947e{-}02, 5.799929e{-}04, 3.166654e{-}02]$

argmax(x) = [0, 0, 1, 0, 0, 0].