LING572 HW5: Gradient and Softmax

Due: 11pm on Feb 8, 2022

Author: DHWANI SERAI

A few notes about this assignment:

- The answers to the questions should be pretty short. I've left some space for you to fill in the answers. I've also made the LaTeX file available in case you want to add the answers to the latex file directly. In that case, you need to run pdf2latex, latexmk, or something like that to generate a pdf from the LaTeX file.
- If you prefer to write formulas on paper (instead of typing them with LaTeX or Word), it's ok. You just need to fill out the rest of the assignment, print out the file, insert formulas by hand, scan the paper, and then submit via Canvas.
- Since no programming is required, you only need to submit a single file. Please call it **readme.pdf**.
- The assignment has three parts:
  - Q1-Q2 are on the derivative of a univariate function (a function with a single variable), which should have been covered in a college-level calculus course (a prerequisite of LING572).
  - Q3 is on the partial derivates of a multivariate function. If you have not learned that topic before, you can look at the tutorials provided in Q3.
  - Q4 is on softmax, which has been covered in ling570. I include the url of a short tutorial on the function.
- There are tons of textbooks and online tutorials that cover those topics. If the links provided in Q3-Q4 do not work for you or you are still confused after going over them, feel free to read any calculus textbook or search the Internet for more info.

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

(a) 2 pts: What does f'(x) intend to measure?

$F'(x)$ measures the rate of change of $f(x)$ w.r.t. x.
It basically measures the slope of a function (on a graph) at any x

(b) 2 pts: Let $h(x) = f(g(x))$. What is $h'(x)$ in terms of f'(x) and g'(x)?

Using Chain Rule, $h'(x) = f'(g(x)) \cdot g'(x)$

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

Using Product Rule, $h(x) = f(x).g'(x) + f'(x).g(x)$

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

$f(x) = a^x$ — I

Taking logn on both sides

$\ln[f(x)] = \ln a^x$

$\therefore \ln[f(x)] = x \ln a$

$\rightarrow$ Differentiating both sides

$\frac{1}{f(x)} \cdot f'(x) = \ln a$

$\therefore f'(x) = f(x). \ln a = \boxed{a^x \ln a}$

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

$f(x) = x^{10} - 2x^8 + 4x^{-2} + 10$

$\therefore f'(x) = 10x^9 - 2\times 8\, x^7 + 4.(-2)x^{-3} + 0$ (using Power Rule)

$$\boxed{f'(x) = 10x^9 - 16x^7 - \frac{8}{x^3}}$$

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 $f'(x) = f(x)(1 - f(x))$.

$f(x) = (1+e^{-x})^{-1}$

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

$= \frac{-1}{(1+e^{-x})^2} . \left(0 + e^{-x} . \frac{d}{dx}(-x)\right)$

$\therefore f'(x) = \frac{-1 \times -1\, e^{-x}}{(1+e^{-x})^2}$

$= \frac{e^{-x}}{(1+e^{-x})^2}$ [LHS]

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

$= \frac{e^{-x}}{(1+e^{-x})^2}$ [RHS]

Hence, Proved.

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

$g(x) = (e^x - e^{-x})(e^x + e^{-x})^{-1}$

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

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

Proved

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

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

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

$\therefore$ LHS = RHS [Proved]

Q3 (45 points): Let f be a multi-variate function, and let x be one of the variables in f. Let us denote the partial derivative of f with respect to x by f'_x or $\frac{df}{dx}$ or $\frac{\partial f}{\partial x}$. Please answer the following questions:

(a) 15 free pts: Refresh your memory about gradient, partial derivative, chain rule. Here are some readings on this. Free free to skip them if you already know the content. On the other hand, if you need more info or cannot access the videos as youtube is blocked in your country, just search for "partial derivatives", "gradient", and "chain rule with partial derivatives". There should be tons of materials on those topics on the Internet.

- Khan Academy's page on partial derivatives:
  https://www.khanacademy.org/math/multivariable-calculus/multivariable-derivatives/partial-derivati
  a/introduction-to-partial-derivatives
- Khan Academy's page on the gradient:
  https://www.khanacademy.org/math/multivariable-calculus/multivariable-derivatives/partial-derivati
  a/the-gradient

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

The partial derivative f'_x is trying to measure the rate of change of function wrt x while keeping other variables constant.

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

The gradient of a function at point z is a vector of all the partial derivatives of all variables (n) in the n dinensional space. If z is $(x_1, x_2, x_3, \dots x_n)$ gradient$(\nabla f)_{(x_1, x_2, \dots x_n)} = \begin{bmatrix} f'_{x_1} \\ f'_{x_2} \\ f'_{x_n} \end{bmatrix}$

(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 $\frac{\partial z}{\partial t}$ using the chain rule of partial derivatives?

The chain rule states that,

$$\frac{\partial z}{\partial t} = \frac{\partial f}{\partial x} \cdot \frac{dx}{dt} + \frac{\partial f}{\partial y} \cdot \frac{dy}{dt} = \frac{\partial f}{\partial x} \cdot g'(t) + \frac{\partial f}{\partial y} h'(t)$$

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

What is f'_x? What is f'_y? For the given function,

$$f'_x = 3x^2 + 3.2xy + 0 + 2 = 3x^2 + 6xy + 2$$

$$f'_y = 0 + 3x^2(1) + 3y^2 + 0 = 3x^2 + 3y^2$$

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

From the calculation above we know, $\nabla f(x,y) = \begin{bmatrix} 3x^2 + 6xy + 2 \\ 3x^2 + 3y^2 \end{bmatrix}$

$$\therefore \nabla f(1,2) = \begin{bmatrix} 3(1)^2 + 6(1)(2) + 2 \\ 3(1)^2 + 3(2)^2 \end{bmatrix} = \begin{bmatrix} 17 \\ 15 \end{bmatrix}$$

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

$$z = \sum_{i=1} w_i x_i = w_1 x_1 + w_2 x_2 + \dots + w_n x_n$$

$$\frac{\partial z}{\partial w_1} = x_1 \quad \frac{\partial z}{\partial w_2} = x_2 \dots \frac{\partial z}{\partial w_n} = x_n$$

Therefore, $\boxed{\frac{\partial z}{\partial w_i} = x_i}$

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

What is $\frac{\partial f}{\partial z}$? Here z is a single variable function so,

$\frac{\partial f}{\partial z} = \frac{df}{dz} = f'(z)$; From Q2 a) we know that for $f(z) = \frac{1}{1+e^{-z}}$ $f'(z) = f(z)(1-f(z$

So $\frac{\partial f}{\partial z} = f(z).(1 - f(z))$

What is $\frac{\partial f}{\partial w_i}$?

Using chain rule, $\frac{\partial f}{\partial w_i} = \frac{\partial f}{\partial z} \cdot \frac{\partial z}{\partial w_i} = f(z).(1-f(z)) \cdot x_i$ [From part (f) above]

$$\therefore \frac{\partial f}{\partial w_i} = f(z).(1-f(z)).x_i$$

Hint: Use chain rule and your answers should contain $f(z)$.

(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 $\frac{\partial E}{\partial w_i}$? Hint: the answer should contain $f(z)$.

$$\frac{\partial E}{\partial w_i} = \frac{\partial E}{\partial f(z)} \cdot \frac{\partial f(z)}{\partial z} \cdot \frac{\partial z}{\partial w_i} \quad \text{[using Chain Rule]}$$

$$= \frac{1}{2} \times 2(t - f(z)).(0 - f'(z)) \cdot f'(z) \cdot x_i$$

$$= [t - f(z)](-1)[f'(z)]^2 . x_i$$

$$= -[t - f(z)][f(z)(1 - f(z))]^2 \cdot x_i \quad \text{[From Q2 a)]}$$

Q4 (25 points): The softmax function: please read the short tutorial at https://deepai.org/machine-learning-glossary-and-terms/softmax-layer and answer the following questions:

(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?

Softmax function takes an input x and gives output y.

⇒ If x ~~is~~ represents a single variable then it can be any real number or else it can be a vector of real numbers.

⇒ y is either a real value between 0 and 1 (binary) or a vector (n classes) with a real value (0, 1)

(b) **5 pts:** In general which layer in neural network (NN) is the softmax function used and why?

The softmax function is generally used in the final layer of a neural network because it converts any real valued number to a real valued number between 0 and 1 [i.e. (0, 1)]. It helps in normalizing values to a probability distribution.

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

Sigmoid is basically a special type of Softmax function which is used for binary classification. In the formula for softmax if we set the input vector to be [x, 0] we get ~~s~~ sigmoid.

Softmax ~~Sigmoid~~ $(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$ [K classes] ; Sigmoid $((x, 0)) = \frac{e^x}{e^x + e^0} = \frac{1}{1 + e^{-x}}$

(d) **7 pts:** What is the relationship between the softmax function and the argmax function? In NN, when do you use softmax and when do you use argmax?

Softmax function gives a value between 0 and 1 [i.e. (0, 1)] for every input ~~va~~ in the vector, whereas argmax gives either 0 or 1 as the output for every input in the vector. Argmax only gives the value 1 for the input with max value and 0 for all others

~~(e) 6 pts: If a vector x is [1, 2, 3, -1, -4, 0], what is softmax(x)? What is argmax(x)~~?

In NN, softmax is generally used for training to get good optimization cost (differentiability) but it is switched to argmax for inference. During inference time it is easier to just look at a vector of 0's with only one 1 value to identify which class has the highest probability.

572. HW 5

Q4 part e) $x = [1, 2, 3, -1, -4, 0]$

To find: Softmax(x) , Argmax(x)

Solution: Argmax$(x) = [0, 0, 1, 0, 0, 0]$

So we know that the third class has the max value.

For Softmax(x),

Let the denominator (normalization factor) be d

So, $d = e^{1} + e^{2} + e^{3} + e^{-1} + e^{-4} + e^{0}$

$= \cancel{272}\ 31.579$

$\therefore$ Softmax$(x) = \left[\frac{e^{1}}{d}, \frac{e^{2}}{d}, \frac{e^{3}}{d}, \frac{e^{-1}}{d}, \frac{e^{-4}}{d}, \frac{e^{0}}{d}\right]$

0.012 0.0005

$= [0.086, 0.234, 0.636, \quad , \quad , 0.0316]$

Softmax$(x) = [0.086 \ \ 0.234 \ \ 0.636 \ \ 0.012 \ \ 0.0005 \ \ 0.0316]$