CS6910: Tutorial 2

VERSION I

1. Assume that we have five learnable parameters and we wish to use the Gradient Descent algorithm. Suppose that the gradient of the loss function w.r.t the parameters is: [0.51 0.01 -2.15 -0.26 0.10]. In order to minimise the loss, which of the following actions should we take (multiple actions may be valid)?

   (a) Decrease p_2 by 2.15
   (b) Decrease p_3 by 0.26
   (c) Increase p_0 by 0.51
   (d) Increase p_2 by 2.15

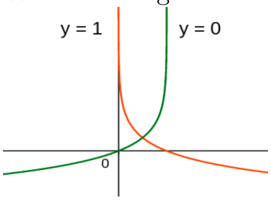
2. In the above question, according to the given gradient vector, changing which of the given five parameters would be the most effective in minimising the loss?

   (a) p_0
   (b) p_1
   (c) p_2
   (d) p_3

3. Taylor series gives a formula for approximating the value of the function $f(x)$ in a small neighborhood around it. Given that $3^3 = 27$, calculate the value of $(3.0001)^3$ using the first order and second order approximations given by Taylor series.

   (a) 27.0027, 27.00270009
   (b) 27.0018, 27.00180009
   (c) 27.0036, 27.00360009
   (d) 27.0045, 27.00450009

4. Which of the given loss functions corresponds to the below given plot?



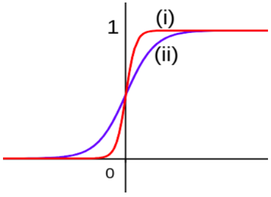
(a) $(y_i - \hat{y_i})^2$

(b) $-y_i * log(\hat{y_i}) - (1 - y_i) * log(1 - \hat{y_i})$

(c) $max(0, 1 - y_i * \hat{y_i})$

(d) None of the above

5. Choose the sigmoid functions which correspond to the graphs (i) and (ii), respectively shown in the figure below.



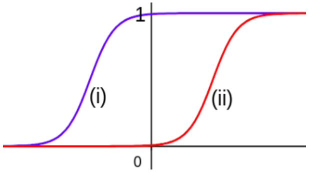
(a) $\frac{1}{1+e^{-(10x+5)}}, \frac{1}{1+e^{-(30x+5)}}$

(b) $\frac{1}{1+e^{-(30x)}}, \frac{1}{1+e^{-(10x)}}$

(c) $\frac{1}{1+e^{-(10x)}}, \frac{1}{1+e^{-(30x)}}$

(d) $\frac{1}{1+e^{-(30x+5)}}, \frac{1}{1+e^{-(10x+5)}}$

6. Choose the sigmoid functions which correspond to the graphs (i) and (ii), respectively shown in the figure below.



(a) $\frac{1}{1+e^{-(10x+5)}}, \frac{1}{1+e^{-(30x+5)}}$

(b) $\frac{1}{1+e^{-(10x+5)}}, \frac{1}{1+e^{-(10x-5)}}$

(c) $\frac{1}{1+e^{-(10x-5)}}, \frac{1}{1+e^{-(10x+5)}}$

(d) $\frac{1}{1+e^{-(30x+5)}}, \frac{1}{1+e^{-(10x+5)}}$

7. Which of the following activation functions is computationally the least expensive?

   (a) $sigmoid(x) = \frac{1}{1+e^{-(wx+b)}}$
   (b) $tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$
   (c) $relu(x) = max(0, x)$
   (d) All are equally expensive

8. Which of the following functions is even?

   (a) logistic sigmoid
   (b) hyperbolic tangent
   (c) linear
   (d) None of the above

9. Consider the ground truth and predicted values of a regression model below. What is the RMS error of the model?

Ground Truth	Predicted value
0.5	0.4
0.63	0.6
0.25	0.35
0.8	0.85
0.55	0.63

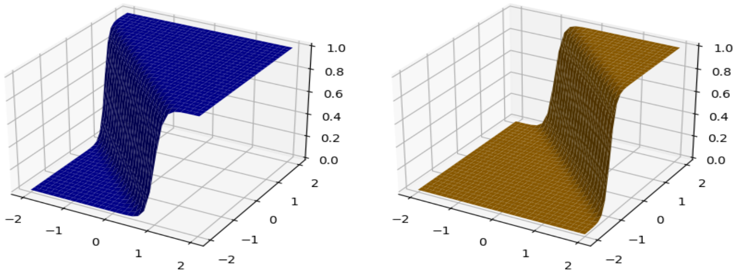
   (a) 0.0772
   (b) 0.0882
   (c) 0.0662
   (d) 0.0992

10. Consider the following class labels and predicted probability values of a classification model. Calculate the cross entropy error of the given model.

Class label	apple	banana	mango
apple	0.8	0.1	0.1
banana	0.15	0.75	0.1
apple	0.6	0.2	0.2
mango	0.2	0.1	0.7

   (a) 0.5585
   (b) 0.5235
   (c) 0.5765
   (d) 0.5985

11. Choose the sigmoid functions which correspond to the blue and orange graphs, respectively shown in the figure below:



(a) $\frac{1}{1+e^{-(10x+10y+10z+10)}}, \frac{1}{1+e^{-(10x+10y+10z-10)}}$
(b) $\frac{1}{1+e^{-(10x+10y-10)}}, \frac{1}{1+e^{-(10x+10y+10)}}$
(c) $\frac{1}{1+e^{-(10x+10y+10)}}, \frac{1}{1+e^{-(10x+10y-10)}}$
(d) $\frac{1}{1+e^{-(10x+10y+10z-10)}}, \frac{1}{1+e^{-(10x+10y+10z+10)}}$

12. State if the following statement is true or false: In batch gradient descent, the parameters get updated only after computing the gradient of the loss for all the training examples.

(a) False
(b) True

13. Gradient based optimization algorithms have the following template : (i) Initialize parameters randomly to θ_0 (ii) Iteratively update the parameters using a certain rule until convergence. Now consider the following optimization problem: $min(x^2 + y^2 + z^2 - 8)$. Suppose we start with the following initializations (x_0, y_0, z_0) = (-3, 0, -1). What will be the updated values of x,y,z after one iteration? Assume the learning rate to be 0.01.

(a) (-3.06, 0, -1.02)
(b) (-2.94, 0, -0.98)
(c) (-3.06, 0, -0.98)
(d) (-2.94, 0, -1.02)

14. Mark the statements which are True:

    (a) A network of perceptrons with a single hidden layer can be used to represent any continuous function precisely.
    (b) A network of sigmoid neurons with a single hidden layer can be used to represent any continuous function precisely.
    (c) A network of perceptrons with a single hidden layer can be used to represent any continuous function to any desired degree of precision.
    (d) A network of sigmoid neurons with a single hidden layer can be used to represent any continuous function to any desired degree of precision.

15. Suppose you are minimizing the following function $x^2 + y^2 + z^2$. Instead of using the standard gradient descent update suppose you use the following update rule: $x_{k+1} = x_k - [f^{''}(x_k)]^{-1} f'(x_k)$. Calculate the number of steps required to converge if your initial values for x, y, z are (i) (1,0,0) (ii) (1, 0, 1) (iii) (-3, 0, -1). Can you reason why you get this specific value for the number of steps?

    (a) 4
    (b) 2
    (c) 3
    (d) 1

    Reason: ________________

16. State if the following statement is true or false: In standard gradient descent algorithm, higher learning rate will always help in fast convergence of the algorithm.

    (a) True
    (b) False

17. What is the maximum value of the derivative of the function $\frac{1}{1+e^{-x}}$? At what value of the input would this maximum value be achieved?

    (a) 0.25, 0.5
    (b) 0.5, 0
    (c) 0.25, 0
    (d) 0.5, 0.25

18. What is the maximum value of the derivative of the function $\frac{e^x-e^{-x}}{e^x+e^{-x}}$? At what value of the input would this maximum value be achieved?

    (a) 1, e
    (b) 0.5, 0
    (c) 1, 0
    (d) 0.5, e

19. What is the relationship between logistic function ($f(x)$) and hyperbolic tangent function ($g(x)$)?

    (a) $f(x) = 0.5 * g(x)$
    (b) $g(x) = 2 * f(x) - 1$
    (c) $f(x) = 2 * g(2x) - 1$
    (d) $g(x) = 2 * f(2x) - 1$

20. A perceptron function is _______, whereas a logistic sigmoid function is _______.

    (a) continuous, differentiable
    (b) not continuous, not differentiable
    (c) not continuous, differentiable
    (d) differentiable, not differentiable

21. Consider the following points of the form (x_1, x_2). Suppose the points (-1, -1) and (1,1) belong to the positive class and the points (1,-1) and (-1,1) belong to the negative class. Notice that the points are not linearly separable, can you suggest a simple non-linear transformation, $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ so that the points become linearly separable?

    multiply

22. Which of the following sigmoid functions is closest to the perceptron step function?

    (a) $\frac{1}{1+e^{-10x}}$
    (b) $\frac{1}{1+e^{-10x+5}}$
    (c) $\frac{1}{1+e^{-100x}}$
    (d) $\frac{1}{1+e^{-100x+5}}$

23. Write the pseudo code for the perceptron learning algorithm.

24. What are the derivatives of $f(x) = \frac{1}{1+e^{-x}}$ and $g(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$?

   (a) $f(x)(1 - f(x))$, $1 - (g(x))^2$
   (b) $1 - (f(x))^2$, $g(x)(1 - g(x))$
   (c) $f(x)(1 + f(x))$, $1 - (g(x))^2$
   (d) $f(x)(1 - f(x))$, $1 + (g(x))^2$

25. Prove that for a quadratic loss function of a single parameter, the optimal learning rate is the reciprocal of the second order derivative of the loss function.

26. Consider $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$. Compute the gradient of $x^T Ax$ w.r.t A.

   (a) A^{-1}
   (b) $-A^{-1}xx^TA^{-1}$
   (c) $A^{-1}xx^T$
   (d) None of the above

27. Consider the task of assigning one of the following 3 labels to an image : apple, banana, mango. Further, consider we are given n images for training such that each image belongs to one of these three categories. We have a model which assigns a probability distribution to each training example: $\mathbf{q} = [q_{apple}, q_{banana}, q_{mango}]$. We train the model by minimizing the cross entropy between the true distribution and the predicted distribution for each training example. Show that this is the same as maximizing the log likelihood of the training data.

28. Which of the following loss functions is differentiable?

   (a) $\sum_{i=1}^{N} |y_i - \hat{y}_i|$
   (b) $\sum_{i=1}^{N} max(0, 1 - y_i * \hat{y}_i)$
   (c) $\sum_{i=1}^{N} max(-y_i * \hat{y}_i, 0)$
   (d) None of the above

29. Consider the following points (x,y): (1, 1), (-0.5, 2.8), (0.2, -1), (-2, 0), (-3, -1). Which of the following model is a better fit for the data? Use squared error loss to compare the two models.
Model 1: $y = -3x - 3$
Model 2: $y = -2x - 2$

(a) Model 1 is better.
(b) Model 2 is better.
(c) The two models are equivalent.

30. Consider the input $x \in \mathbb{R}^n$, output $y \in \mathbb{R}$ and learnable matrices $W_1, W_2 \in \mathbb{R}^{n \times n}$. You approximate the relationship between y and x using one of the models shown below. Which of these is not a linear model?

(a) $W_1 x$
(b) $x^T W_1 x$
(c) $W_1 W_2 x$
(d) $x^T W_1 x + W_2 x$

END