CS2109s - Tutorial 8

Eric Han (TG12-TG15)

Apr 4, 2024

Annoucements

Important admin

- Tutorials left:
  - Tutorial 9: 11 Apr 2024
  - Tutorial 10: 18 Apr 2024
- Feel free to approach me to chat about research/module etc
  - Lunch / Coffee in school

Problem Set 5

- Q9 - Logistic regression using stochastic gradient descent
  - For iteration, update a randomly selected datapoint.
- Q11 - Stochastic gradient descent vs batch gradient descent
  - SGD faster BGD slower for every iteration
  - SGD *usually* reaches the same loss faster
  - SGD varies more due (better/worse) to the random updates (depending on point)
- Q16 - Linear SVM vs Gaussian Kernel SVM
  - Linear Separability / Model Complexity
  - Data Sparsity: More features » data points
  - Compute complexity is tricky - Time Taken for Linear is actually longer due to non-linear separability
  - Performance is also tricky - More data added, Gaussian performs worse (overfitting)

Student Feedback on Teaching (SFT)

NUS Student Feedback https://blue.nus.edu.sg/blue/, due **26 Apr**:

- Don't Mix module/grading/project feedback - **feedback only for teaching**.
- Feedback is confidential to university and anonymous to us.
- Feedback is optional but highly encouraged.
- Past student feedback improves teaching; see https://www.eric-han.com/teaching
  - ie. Telegram access, More interactivity.
- Your feedback is important to me, and will be used to improve my teaching.
  - Good > Positive feedback > Encouragement
    - Teaching Awards (nominate)
    - Steer my career path
  - Bad > Negative feedback (nicely pls) > Learning
    - Improvement
    - Better learning experience

Student Feedback on Teaching (SFT)

Your feedback is important to me; optional, but highly encouraged:

Figure 1: NUS Student Feedback on Teaching - https://blue.nus.edu.sg/blue/

Question 1

$$f^{[1]} = W^{[1]^T} X, \quad \hat{Y} = g^{[1]}(f^{[1]}), \quad \mathcal{E} = -\frac{1}{n}\sum_{i=0}^{n-1}\Big\{[Y_{0i} \cdot log(\hat{Y}_{0i})] + [(1-Y_{0i})log(1-\hat{Y}_{0i})]\Big\}$$

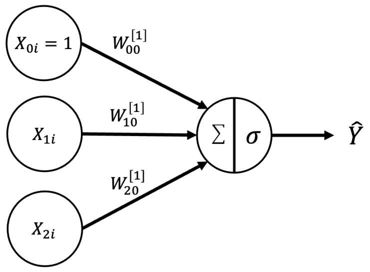

Figure 2: Simple Neural Network

Question 1a [G]

When $n = 1$:

i. $\frac{\partial \mathcal{E}}{\partial \hat{Y}} = \Big[-\frac{Y_{00}}{\hat{Y}_{00}} + \frac{1-Y_{00}}{1-\hat{Y}_{00}}\Big]$ (Given)

ii. $\frac{\partial \mathcal{E}}{\partial f^{[1]}} = \hat{Y} - Y$

iii. $\frac{\partial \mathcal{E}}{\partial W_{20}^{[1]}} = \left(\frac{\partial \mathcal{E}}{\partial f^{[1]}}\right)_{00} X_{20}$

Recap

- What is back propagation?
- How to perform forward propagation?
- How to perform back propagation?

Answer ii

Since $n = 1$, $\frac{\partial \mathcal{E}}{\partial f^{[1]}} = \frac{\partial \mathcal{E}}{\partial f_{00}^{[1]}} = \frac{\partial \mathcal{E}}{\partial \hat{Y}_{00}} \frac{\partial \hat{Y}_{00}}{\partial f_{00}^{[1]}}$ (chain rule)

Since $\hat{Y}_{00} = \sigma(f_{00}^{[1]}) \implies \frac{\partial \hat{Y}_{00}}{\partial f_{00}^{[1]}} = \sigma(f_{00}^{[1]})\Big(1 - \sigma(f_{00}^{[1]})\Big) = \hat{Y}_{00}(1 - \hat{Y}_{00})$

From (i), $\frac{\partial \mathcal{E}}{\partial \hat{Y}_{00}} = \Big[-\frac{Y_{00}}{\hat{Y}_{00}} + \frac{1-Y_{00}}{1-\hat{Y}_{00}} \Big]$

$$\begin{aligned}
\frac{\partial \mathcal{E}}{\partial f^{[1]}} &= \Big[\frac{\partial \mathcal{E}}{\partial \hat{Y}_{00}} \frac{\partial \hat{Y}_{00}}{\partial f^{[1]}_{00}} \Big] \\
&= \Big[-\frac{Y_{00}}{\hat{Y}_{00}} + \frac{1-Y_{00}}{1-\hat{Y}_{00}} \Big] \Big[\hat{Y}_{00}(1-\hat{Y}_{00}) \Big] \\
&= \Big[-Y_{00}(1-\hat{Y}_{00}) + (1-Y_{00})\hat{Y}_{00} \Big] \\
&= \Big[\hat{Y}_{00} - Y_{00} \Big] \\
&= \hat{Y} - Y.
\end{aligned}$$

Answer iii

Since $n = 1$, $\frac{\partial \mathcal{E}}{\partial W^{[1]}_{20}} = \frac{\partial \mathcal{E}}{\partial f^{[1]}_{00}} \frac{\partial f^{[1]}_{00}}{\partial W^{[1]}_{20}}$ (chain rule)

$f^{[1]}_{00} = W^{[1]^T} X = \sum_{i=0}^{2} (W^{[1]^T})_{0i} X_{i0} = \sum_{i=0}^{2} W^{[1]}_{i0} X_{i0} \implies \frac{\partial f^{[1]}_{00}}{\partial W^{[1]}_{20}} = X_{20}$

$$\begin{aligned}
\frac{\partial \mathcal{E}}{\partial W^{[1]}_{20}} &= \frac{\partial \mathcal{E}}{\partial f^{[1]}_{00}} \frac{\partial f^{[1]}_{00}}{\partial W^{[1]}_{20}} \\
&= \frac{\partial \mathcal{E}}{\partial f^{[1]}_{00}} X_{20} \\
&= \left(\frac{\partial \mathcal{E}}{\partial f^{[1]}} \right)_{00} X_{20}
\end{aligned}$$

Note: $\frac{\partial \mathcal{E}}{\partial \hat{Y}}$, and $\frac{\partial \mathcal{E}}{\partial f^{[1]}}$ are matrices since $\mathcal{E}$ is a scalar, but $\hat{Y}$ and $f^{[1]}$ are matrices. However, $\frac{\partial \mathcal{E}}{\partial W^{[1]}_{20}}$ is a scalar since $W^{[1]}_{20}$ is a scalar.

Question 1b-e [G]

b. Derive an expression for $\frac{\partial \mathcal{E}}{\partial W^{[1]}}$, how does back propagation work?
c. Let us consider a general case where $n \in \mathbb{N}$, find $\frac{\partial \mathcal{E}}{\partial f^{[1]}}$.
d. Why do the hyper-parameters α and β? How to set their values?

$$\mathcal{E} = -\frac{1}{n} \sum_{i=0}^{n-1} \Big\{ \alpha [Y_{0i} \cdot log(\hat{Y}_{0i})] + \beta [(1 - Y_{0i}) \cdot log(1 - \hat{Y}_{0i})] \Big\}$$

Answer 1b

From (a), $\frac{\partial \mathcal{E}}{\partial W^{[1]}_{i0}} = \left(\frac{\partial \mathcal{E}}{\partial f^{[1]}} \right)_{00} X_{i0}$

$$\begin{aligned}
\frac{\partial \mathcal{E}}{\partial W^{[1]}} &= \left[\frac{\partial \mathcal{E}}{\partial W^{[1]}_{00}}, \frac{\partial \mathcal{E}}{\partial W^{[1]}_{10}}, \frac{\partial \mathcal{E}}{\partial W^{[1]}_{20}} \right]^T = \left(\frac{\partial \mathcal{E}}{\partial f^{[1]}} \right)_{00} [X_{00}, X_{10}, X_{20}]^T = \left(\frac{\partial \mathcal{E}}{\partial f^{[1]}} \right)_{00} X \\
&= \Big(\hat{Y} - Y \Big) X = \Big(g^{[1]}(f^{[1]}) - Y \Big) X = \Big(g^{[1]}(W^{[1]^T} X) - Y \Big) X
\end{aligned}$$

Intuition behind back propagation $W^{[1]} = W^{[1]} - \alpha \frac{\partial \mathcal{E}}{\partial W^{[1]}}$:

- Change in first layer weighted sum $f^{[1]}$
- Change in predicted value $\hat{Y}$

- Change of loss $\mathcal{E}$
- Decrease the loss by changing the weights

Answer 1c

From (a), $\frac{\partial \mathcal{E}}{\partial f_{0i}^{[1]}} = \left[\frac{\partial \mathcal{E}}{\partial \hat{Y}_{0i}} \frac{\partial \hat{Y}_{0i}}{\partial f_{0i}^{[1]}}\right]$

$$\frac{\partial \mathcal{E}}{\partial \hat{Y}} = \left[\frac{\partial \mathcal{E}}{\partial \hat{Y}_{00}}, \frac{\partial \mathcal{E}}{\partial \hat{Y}_{01}}, \cdots, \frac{\partial \mathcal{E}}{\partial \hat{Y}_{0n}}\right] = \left[\cdots, \frac{1}{n}\Big(-\frac{Y_{0i}}{\hat{Y}_{0i}} + \frac{1-Y_{0i}}{1-\hat{Y}_{0i}}\Big), \cdots\right]$$

$$\frac{\partial \hat{Y}_{0i}}{\partial f_{0i}^{[1]}} = \sigma(f_{0i}^{[1]})\Big(1-\sigma(f_{0i}^{[1]})\Big) = \hat{Y}_{0i}(1-\hat{Y}_{0i})$$

$$\begin{aligned} \frac{\partial \mathcal{E}}{\partial f^{[1]}} &= \left[\frac{\partial \mathcal{E}}{\partial f_{00}^{[1]}}, \frac{\partial \mathcal{E}}{\partial f_{01}^{[1]}}, \cdots, \frac{\partial \mathcal{E}}{\partial f_{0n}^{[1]}}\right] \\ &= \frac{1}{n}\Big[(\hat{Y}_{00}-Y_{00}), (\hat{Y}_{01}-Y_{01}), \dots, (\hat{Y}_{0n}-Y_{0n})\Big] \\ &= \frac{1}{n}(\hat{Y}-Y) \end{aligned}$$

Answer 1d

Weighted Error:

$$\mathcal{E} = -\frac{1}{n}\sum_{i=0}^{n-1}\Big\{\alpha[Y_{0i} \cdot log(\hat{Y}_{0i})] + \beta[(1-Y_{0i}) \cdot log(1-\hat{Y}_{0i})]\Big\}$$

Apply a weight to how much each class contributes to the loss function:

- Error due to Cultiva A ($p_A = 100/1100$): $Y_{0i} \cdot log(\hat{Y}_{0i})$
- Error due to Cultiva B ($p_B = 1000/1100$): $(1-Y_{0i}) \cdot log(1-\hat{Y}_{0i})$

Since we have unbalanced dataset, we can weight using the ratio $\frac{\alpha}{\beta} = \frac{1/100}{1/1000}$:

- $\alpha = 1/100$
- $\beta = 1/1000$

We punish the model more heavily if it misclassifies A, so the model won't be biased towards predicting all samples as B.

Question 2 [G]

When $n = 1$, compute $\frac{\partial \mathcal{E}}{\partial W_{11}^{[1]}}$, where $f^{[1]} = W^{[1]^T}X$, $a^{[1]} = g^{[1]}(f^{[1]})$, $f^{[2]} = W^{[2]^T}a^{[1]}$, $\hat{Y} = g^{[2]}(f^{[2]})$, $g^{[1]}(s) = ReLU(s)$, $g^{[2]}(s) = \sigma(s) = \frac{1}{1+e^{-s}}$, $W^{[1]} \in \mathbb{R}^{3\times 2}$, $W^{[2]} \in \mathbb{R}^{2\times 1}$.

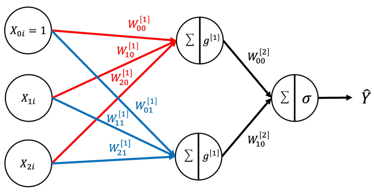

Figure 3: Complex NN

[@] Is ReLU continous/discontinuous, not/differentiable; Can we use discontinuous activation functions?

Answer

Intuition: Plot the forward path and take the derivative.

$$W_{11}^{[1]} \xrightarrow{f^{[1]}=W^{[1]^T}X} f_{10}^{[1]} \xrightarrow{a^{[1]}=g^{[1]}(f^{[1]})} a_{10}^{[1]} \xrightarrow{f^{[2]}=W^{[2]^T}a^{[1]}} f_{00}^{[2]} \xrightarrow{\hat{Y}=g^{[2]}(f^{[2]})} \hat{Y}_{00} \rightarrow \mathcal{E}$$

$$f^{[1]} = \begin{bmatrix} W_{00}^{[1]} & W_{01}^{[1]} \\ W_{10}^{[1]} & W_{11}^{[1]} \\ W_{20}^{[1]} & W_{21}^{[1]} \end{bmatrix}^T \begin{bmatrix} X_{00} \\ X_{10} \\ X_{20} \end{bmatrix} = \begin{bmatrix} \sum_i W_{i0}^{[1]} X_{i0} \\ \sum_i W_{i1}^{[1]} X_{i0} \end{bmatrix}$$

$$f^{[2]} = \begin{bmatrix} W_{00}^{[2]} \\ W_{10}^{[2]} \end{bmatrix}^T \begin{bmatrix} a_{00}^{[1]} \\ a_{10}^{[1]} \end{bmatrix} = \begin{bmatrix} \sum_i W_{i0}^{[2]} a_{i0}^{[1]} \end{bmatrix}$$

Expand using chain rule: $\frac{\partial \mathcal{E}}{\partial W_{11}^{[1]}} = \frac{\partial \mathcal{E}}{\partial \hat{Y}_{00}} \frac{\partial \hat{Y}_{00}}{\partial f_{00}^{[2]}} \frac{\partial f_{00}^{[2]}}{\partial a_{10}^{[1]}} \frac{\partial a_{10}^{[1]}}{\partial f_{10}^{[1]}} \frac{\partial f_{10}^{[1]}}{\partial W_{11}^{[1]}}$

Find each of the terms in $\frac{\partial \mathcal{E}}{\partial W_{11}^{[1]}} = \frac{\partial \mathcal{E}}{\partial \hat{Y}_{00}} \frac{\partial \hat{Y}_{00}}{\partial f_{00}^{[2]}} \frac{\partial f_{00}^{[2]}}{\partial a_{10}^{[1]}} \frac{\partial a_{10}^{[1]}}{\partial f_{10}^{[1]}} \frac{\partial f_{10}^{[1]}}{\partial W_{11}^{[1]}}$:

- $\frac{\partial \mathcal{E}}{\partial \hat{Y}_{00}} = -\frac{\alpha Y_{00}}{\hat{Y}_{00}} + \frac{\beta(1-Y_{00})}{1-\hat{Y}_{00}}$
- $\frac{\partial \hat{Y}_{00}}{\partial f_{00}^{[2]}} = \sigma(f_{00}^{[2]})\Big(1-\sigma(f_{00}^{[2]})\Big)$
- $\frac{\partial f_{00}^{[2]}}{\partial a_{10}^{[1]}} = W_{10}^{[2]}$
- $\frac{\partial a_{10}^{[1]}}{\partial f_{10}^{[1]}} = \begin{cases} 0, \text{if } f_{10}^{[1]} \le 0 \\ 1, \text{otherwise} \end{cases} = \mathbb{1}_{f_{10}^{[1]}>0}$
- $\frac{\partial f_{10}^{[1]}}{\partial W_{11}^{[1]}} = X_{10}$

where $\mathbb{1}_{f_{10}^{[1]}>0}$ is an indicator function. Therefore,

$$\frac{\partial \mathcal{E}}{\partial W_{11}^{[1]}} = \Big[-\frac{\alpha Y_{00}}{\hat{Y}_{00}} + \frac{\beta(1-Y_{00})}{1-\hat{Y}_{00}}\Big]\sigma(f_{00}^{[2]})\Big(1-\sigma(f_{00}^{[2]})\Big)W_{10}^{[2]}\mathbb{1}_{f_{10}^{[1]}>0}X_{10}$$

Question 3 [G]

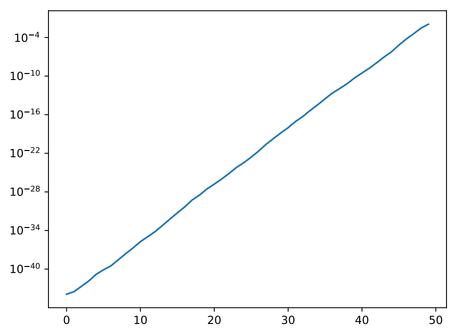

Figure 4: Layers / Max Abs Gradient, using sigmoid

a. Gradient magnitudes of the first few layers are extremely small, what's the problem?
b. Based on what we have learnt thus far, how can we mitigate this problem?
   - [@] Other sophisticated ways to resolve the issue, and why does it work?

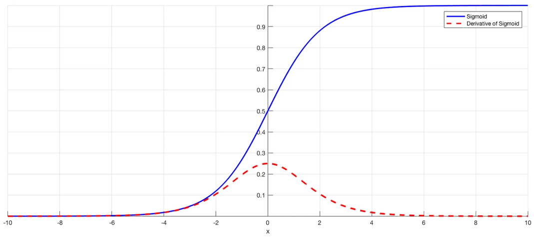

Figure 5: Sigmoid Function

Answer 3a

a. The eariler the weight, more terms needed to compute its update.
   - we need to take product of many, many derivatives
   - derivatives of sigmoid is in $(0, 1/4]$
   - ending up with a really small number
   - causing convergence to be slow.

Answer 3b

Use ReLU.

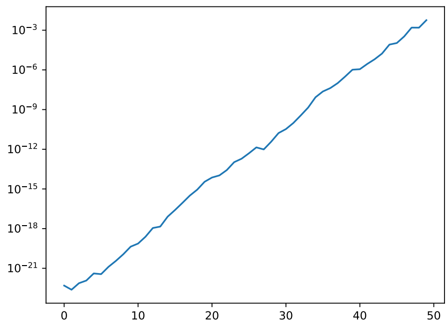

Figure 6: Layers / Max Abs Gradient, using ReLU

Bonus Qn

Investigate for exploding gradient as per question 3, use the code given for tutorial as a starting point.

Tasks

1. Implement a neural network that exhibits exploding gradient.
2. Plot the magnitudes for all layers like done in Question 3.
3. Analyse ways to mitigate the issue.

Buddy Attendance Taking

1. [@] and Bonus declaration is to be done here; You should show bonus to Eric.
2. Attempted tutorial should come with proof (sketches, workings etc…)
3. Guest students must inform Eric and also register the attendance.

Figure 7: Buddy Attendance: https://forms.gle/jsGfFyfo9PTgWxib6