C247 Midterm

ASHISH KUMAR SINGH
UID : 105479019

Answer :- 1

1a)

4 outer points will give wrong class

so error $= \frac{4}{14} = 0.2857$

Answer :- 2

2a) i)

$$L = -y \cdot s$$

$$\frac{\partial L}{\partial s} = -y$$

$$s = \frac{z^T \hat{z}}{\|z\|_2 \|\hat{z}\|_2}$$

Using derivative quotient rule,

$$\frac{\partial s}{\partial z} = \frac{\left(\|z\|_2 \|\hat{z}\|_2\right)\hat{z} - \left(z^T \hat{z}\right)\left(\frac{\|\hat{z}\|_2}{\|z\|_2}\right) z}{\|z\|_2^2 \ \|\hat{z}\|_2^2}$$

and $$\frac{\partial s}{\partial \hat{z}} = \frac{\left(\|z\|_2 \|\hat{z}\|_2\right) z - \left(z^T \hat{z}\right)\left(\frac{\|z\|_2}{\|\hat{z}\|_2}\right)\hat{z}}{\|z\|_2^2 \ \|\hat{z}\|_2^2}$$

Using chain rule,

$$\nabla_z L = \frac{\partial L}{\partial z} = \frac{\partial s}{\partial z} \cdot \frac{\partial L}{\partial s} = -\left(\frac{\partial s}{\partial z}\right) y$$

$$\nabla_{\hat{z}} L = \frac{\partial L}{\partial \hat{z}} = \frac{\partial s}{\partial \hat{z}} \cdot \frac{\partial L}{\partial s} = -\left(\frac{\partial s}{\partial \hat{z}}\right) y$$

2 a) ii)

Using chain rule for matrix derivatives,

$$\nabla_{w_2} L = \frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial z} h_1^T + \frac{\partial L}{\partial \hat{z}} \hat{h}_1^T$$

$$\nabla_{w_2} L = \delta_z h_1^T + \delta_{\hat{z}} \hat{h}_1^T$$

2a) iii)

$$\nabla_{h_1} L = \frac{\partial L}{\partial h_1} = w_2^T \frac{\partial L}{\partial z}$$

$$\delta_{h_1} = w_2^T \delta_z$$

$$\nabla_{\hat{h}_1} L = \frac{\partial L}{\partial \hat{h}_1} = w_2^T \frac{\partial L}{\partial \hat{z}}$$

$$\delta_{\hat{h}_1} = w_2^T \delta_{\hat{z}}$$

2a) iv)

$$\nabla_m L = \frac{\partial L}{\partial m} = \mathbb{I}(w_1 x \geq 0) \odot \delta_{h_1}$$

$$\nabla_n L = \frac{\partial L}{\partial n} = \mathbb{I}(w_1 \hat{x} \geq 0) \odot \delta_{\hat{h}_1}$$

where indicator func $\mathbb{I}(w_1 x \geq 0) = 1$ when $(w_1 x \geq 0)$ else 0

2 a) v)

$$\nabla_{W_1} L = \frac{\partial L}{\partial W_1} = \frac{\partial L}{\partial m} x^T + \frac{\partial L}{\partial n} \hat{x}^T$$

$$\nabla_{W_1} L = \delta_m x^T + \delta_n \hat{x}^T$$

---

2 b) i)

Given $y = +1$, $z^{(g)} = \hat{z}^{(g)}$

$$s = \frac{z^T \hat{z}}{\|z\|_2 \|\hat{z}\|_2}$$

Putting z

$$s = \frac{(z^{(g)})^T z^{(g)}}{\|z^{(g)}\|_2 \|z^{(g)}\|_2} = \frac{\|z^g\|_2^2}{\|z^g\|_2^2} = 1$$

$$L = -y \cdot s = -1$$

$$\boxed{L = -1}$$

---

2 b) ii)

When $z^{(g)}$ & $\hat{z}^{(g)}$ are orthogonal

$$(z^{(g)})^T \cdot \hat{z}^{(g)} = 0$$

$$s = 0$$

$$L = -y \cdot s = 0$$

$$\boxed{L = 0}$$

---

2 b) iii)

$$z^{(g)} = -\hat{z}^{(g)}$$

$$s = \frac{(z^{(g)})^T(-z^{(g)})}{\|z^{(g)}\|_2 \, \|-z^{(g)}\|_2}$$

$$s = -\frac{\|z^{(g)}\|_2^2}{\|z^{(g)}\|_2^2} = -1$$

$$L = -y \cdot s = 1$$

$$\boxed{L = 1}$$

2 c)

Yes,

Since $y = +1$, both signatures are genuine, so, when embedding are same loss should be minimum.

And when embedding are $z^{(g)} = -\hat{z}^{(g)}$ loss is maximum.

For orthogonal embeddings, ideally loss should be positive but it doesn't matter much as small perturbations will lead to non-orthogonal embeddings in next iterations

Answer :- 3

3 a)

B. & C.

~~Both tanh and sigmoid ...~~

Derivative of tanh & sigmoid is close to 0 for large feature values. which when multiplied in chain rule leads to vanishing gradients

3 b)

C. & D.

no more iteration,

have learnable params γ, β

non-linear unit stats

BN changes statistic to var & mean which can be seen as adding noise

BN is applicable at test time

3 c)

A, C, D

L1 leads to some 0 weights

add reg might increase. loss

3d)

A, B, D

We can augment imbalanced class more Multitask can be applicable even if we have small data for a task as we can augment.

3e)

(1) Loss
(2) Epochs / Number of Iterations
(3) Best validation accuracy / least val loss
(4) Early stopping iteration
(5) Validation loss
(6) Training Loss

Answer 1-4

4 a) i)

GD with momentum and Adam will get out of the trap because both of them have momentum component which will push θ towards right even when gradients are small in plateau region.

4 a) ii)

GD momentum > Adam > Adagrad

GD momentum uses varilla momomentum which will give greatest update to θ.

Adam uses exponentially weight momentum so magnitude will be less.

Adagrad will decrease magnitude in that direction as previous step were in same dire.

4b)

$$E[g_t] = \mu \quad , \quad t = 1, 2 \ldots\ldots$$

$$v_t = (1-\beta_1) \sum_{i=1}^{t} \beta_1^{t-i} g_i$$

$$E[v_t] = E\left[(1-\beta_1) \sum_{i=1}^{t} \beta_1^{t-i} g_i\right]$$

Using Linearity of expectations,

$$E[v_t] = (1-\beta_1) \sum_{i=1}^{t} \beta_1^{t-i} E[g_i]$$

$$E[v_t] = (1-\beta_1) \sum_{i=1}^{t} \beta_1^{t-i} \mu$$

$$E[v_t] = \mu(1-\beta_1)\left(1 + \beta_1 + \beta_1^2 \ldots + \beta_1^{t-1}\right)$$

Using GP sum formula

$$E[v_t] = \mu(1-\beta_1)\frac{(1-\beta_1^t)}{(1-\beta_1)}$$

$$E[v_t] = \mu(1-\beta_1^t)$$

$$\mu = \gamma_1 E[v_t] = \frac{E[v_t]}{(1-\beta_1^t)}$$

$$\boxed{\gamma_1 = \frac{1}{(1-\beta_1^t)}}$$

Similar calculation can be done for γ_2

$$\gamma_2 = \frac{1}{(1-\beta_2^t)}$$

Answer :- 5

$$\tilde{\theta}_{new} = \tilde{\theta}_{old} + v_{new} + \alpha(v_{new} - v_{old})$$
$$= \tilde{\theta}_{old} + v_{new}(1+\alpha) - \alpha v_{old}$$
$$= \tilde{\theta}_{old} + (1+\alpha)(\alpha v_{old} - \epsilon \nabla_\theta L(\tilde{\theta}_{old})) - \alpha v_{old}$$
~~$= \theta_{old} + (1+\alpha)$~~
$$\tilde{\theta}_{new} = \tilde{\theta}_{old} + \alpha v_{old} - \epsilon \nabla_\theta L(\tilde{\theta}_{old}) + \alpha^2 v_{old} - \epsilon\alpha \nabla_\theta L(\tilde{\theta}_{old}) - \alpha v_{old}$$

$$\tilde{\theta}_{new} = \tilde{\theta}_{old} + \alpha^2 v_{old} - (1+\alpha)\epsilon \nabla_\theta L(\tilde{\theta}_{old})$$

Substituting $\tilde{\theta}_{old} = \theta_{old} + \alpha v_{old}$

$$= \theta_{old} + \alpha v_{old} + \alpha^2 v_{old} - (1+\alpha)\epsilon \nabla_\theta L(\tilde{\theta}_{old})$$
$$= \theta_{old} + (1+\alpha)\left[\alpha v_{old} - \epsilon \nabla_\theta L(\theta_{old} + \alpha v_{old})\right]$$
$$\tilde{\theta}_{new} = \theta_{old} + (1+\alpha) v_{new}$$
$$\theta_{new} + \alpha v_{new} = \theta_{old} + (1+\alpha) v_{new}$$
$$\theta_{new} = \theta_{old} + v_{new}$$

which is same as vanilla nesterov momentum.

1 b) i)

False

for $k=1$ training error $= 0$

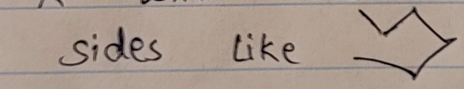
1 b) ii)

Method 2

because tuning params to test will not generalize for unseen data. So tuning hyper params should be done on validation data.

1 c)

B, D

Increasing k will have decision boundary with more sides like  , not smooth

1 d) i)

(3) overfit the data

1 d) ii)

A, D