Theory Report    Abhishek Sushil 2021441

Q1.

a) MSE → simply minimizes dist b/w real & predicted values
Based on squared $L_2$ norm of actual & predictⁿ
∴ No basis for false +ves or false -ve as it takes squared dist
Eg: Here labelling a sweet papaya is not sweet doesnt affect customer but the opposite does

b) $\text{BinCrossEntropy}(\underset{\text{real}}{y}, \underset{\text{pred}}{\hat{y}}) = -\left[ y \log_2 \hat{y} + (1-y)\log_2(1-\hat{y}) \right]$

real ← $y$, $\hat{y}$ → predicted probability

c) $\hat{y}$ for class 0 $= 0.9$ ∴ $\hat{y}$ for class 1 $= 0.1$

$$L(y,\hat{y}) = -\left[1 \log_2 0.1 + 0 \cdot \log_2 0.9\right]$$

$$= -\log_2 0.1 \approx 3.32$$

d) $S_1 \rightarrow L(1, 0.1) = 3.32$ as above

$S_2 \rightarrow L(0, 0.2) = -\left[0 \log 0.2 + 1 \log 0.8\right]$
$= 0.32$

$S_3 \rightarrow L(0, 0.7) = -\left[0 \log 0.7 + 1 \log 0.3\right]$
$= 1.73$

∴ Average $= \underline{1.79}$

e) $W = W - \alpha \cdot \frac{\partial L_{BCE}}{\partial W} + 2\lambda W$

More penalty as $2\lambda W$ term present $\therefore$ a decrease in weights for model A that uses $L_2$ regularizer.
Advantage for model A as $L_2$ prevents overfitting

f) KL Divergence
→ Used to see how far apart 2 probability distribution are based on how well one distribution is likely to generate samples from the other

$$D_{KL}(P(x)|Q(x)) = \sum_{x \in X} P(x) \left( \frac{\log(P(x))}{\cancel{\log}(Q(x))} \right)$$

Cross Entropy
→ Diff. b/w actual & predicted probability distribution

$$H(P(x), Q(x)) = -\sum_{x \in X} P(x) \cdot \log(P(x))$$

$$D_{KL} = \sum_{x \in X} P(x) \left( \frac{\log(P(x))}{\cancel{\log}(Q(x))} \right)$$

$$= \sum_{x \in X} P(x) \left[ \log(P(x)) - \log(Q(x)) \right]$$

$$= \sum_{x \in X} P(x) \log P(x) - \sum_{x \in X} P(x) \log Q(x)$$

$$D_{KL} = -H(P(x)) + H(P(x), Q(x))$$

$$\Rightarrow D_{KL}(P(x)||Q(x)) + H(P(x)) = H(P(x), Q(x))$$

Q2.

a) $\dim(W^{[2]}) = K \times Da$

$\dim(b^{[2]}) = K \times 1$

dim of hidden layer w/ 'm' samples = $Da \times m$

b) $$\hat{y}_K = \frac{e^{Z_K^{[2]}}}{\sum_{i=0}^{K} e^{Z_i^{[2]}}}$$

$$\therefore \frac{\partial \hat{y}_K}{\partial Z_K^{[2]}} = \frac{\partial \left[ \frac{e^{Z_K^{[2]}}}{\sum e^{Z_i^{[2]}}} \right]}{\partial Z_K^{[2]}}$$

$$= \frac{1}{\left(\sum_{i}^{K} e^{Z_i^{[2]}}\right)^2} \left[ \frac{\partial e^{Z_K^{[2]}}}{\partial Z_K^{[2]}} \cdot \sum_{i=0}^{K} e^{Z_i^{[2]}} - \frac{\partial \sum_{i=0}^{K} e^{Z_i^{[2]}}}{\partial Z_K^{[2]}} \cdot e^{Z_K^{[2]}} \right]$$ ... Division Rule

$$= \frac{1}{\left(\sum_{i=0}^{K} e^{Z_i^{[2]}}\right)^2} \left[ e^{Z_K^{[2]}} \cdot \sum_{i=0}^{K} e^{Z_i^{[2]}} - e^{Z_K^{[2]}} \cdot e^{Z_K^{[2]}} \right]$$

$$= \frac{e^{Z_K^{[2]}}}{\left(\sum_{i=0}^{K} e^{Z_i^{[2]}}\right)} - \left[ \frac{e^{Z_K^{[2]}}}{\sum_{i=0}^{K} e^{Z_i^{[2]}}} \right]^2$$

$$\underbrace{\frac{\partial \hat{y}_K}{\partial Z_K^{[2]}} = \hat{y}_K - (\hat{y}_K)^2 = \hat{y}_K (1 - \hat{y}_K)}$$

Scanned with CamScanner

c) $i \neq k$

$$\hat{y}_k = \frac{e^{z_k^{[2]}}}{\sum_{i=0}^{k} e^{z_i^{[2]}}}$$

$$\frac{\partial \hat{y}_k}{\partial z_i^{[2]}} = \frac{\partial \left[\frac{e^{z_k^{[2]}}}{\sum_{i}^{k} e^{z_i^{[2]}}}\right]}{\partial z_i^{[2]}} \quad \rightarrow \quad \frac{\partial e^{z_k}}{\partial z_i} = 0$$

$$= \frac{1}{\left(\sum_{i=0}^{k} e^{z_i^{[2]}}\right)^2} \left[ 0 - e^{z_k^{[2]}} \cdot e^{z_i^{[2]}} \right]$$

$$= \frac{-e^{z_k^{[2]}}}{\sum_{i=0}^{k} e^{z_i^{[2]}}} \cdot \frac{e^{z_i^{[2]}}}{\sum_{i=0}^{k} e^{z_i^{[2]}}}$$

$$\frac{\partial \hat{y}_k}{\partial z_i^{[2]}} = -\hat{y}_k \cdot \hat{y}_i$$

d) i) $i = k$

$$\frac{\partial L}{\partial z_k^{[2]}} = \frac{\partial L}{\partial \hat{y}_k} \cdot \frac{\partial \hat{y}_k}{\partial z_k^{[2]}} \quad \text{chain rule}$$

$$= \frac{\partial \left[-\sum_{i=0}^{k} y_i \log \hat{y}_i\right]}{\partial \hat{y}_k} \cdot \left(\hat{y}_k (1-\hat{y}_k)\right) \quad \text{from b}$$

$$= \frac{y_k \log \hat{y}_k}{\partial \hat{y}_k} \left(\hat{y}_k (1-\hat{y}_k)\right)$$

$$= \frac{y_k}{y_k} (\hat{y}_k)(1-\hat{y}_k)$$

$$\frac{\partial L}{\partial z_k^{[2]}} = \underbrace{y_k}_{1} (1-\hat{y}_k) = (1-\hat{y}_k)$$

Scanned with CamScanner

ii) B $i \neq k$

$$\frac{\partial L}{\partial z_i^{[2]}} = \frac{\partial L}{\partial \hat{y}_k} \cdot \frac{\partial \hat{y}_k}{\partial z_i^{[2]}}$$

$$= \frac{\partial \left[ -\sum_{i=0}^{k} y_i \log \hat{y}_i \right]}{\partial B \hat{y}_k} \cdot \left( -\hat{y}_i \ \hat{y}_k \right) \text{ --- from (c)}$$

$$= \frac{y_k}{\hat{y}_k} \cdot -1 \times \hat{y}_i \cdot \hat{y}_k$$

$$= -1 \cdot y_k \cdot \hat{y}_i$$

$$\frac{\partial L}{\partial z_i^{[2]}} = -\hat{y}_i$$

e) Numerical Instability can be encountered when dealing w/ very large or very small values. These can when taken in exponents very large values lead to numerical overflow & very small nos. to underflow

Assume the final layer has a vector w/ values in a similar range we can normalise the vector before soft-maxing process

$[a, b, c] \rightarrow$ last layer $\quad \bar{x} = \text{mean} = \frac{a+b+c}{3}$

$\sigma = \text{std dev}(a, b, c)$

$\left[ \frac{a-\bar{x}}{\sigma}, \frac{b-\bar{x}}{\sigma}, \frac{c-\bar{x}}{\sigma} \right] \rightarrow$ modified last lay

Now apply soft max on the modified layer