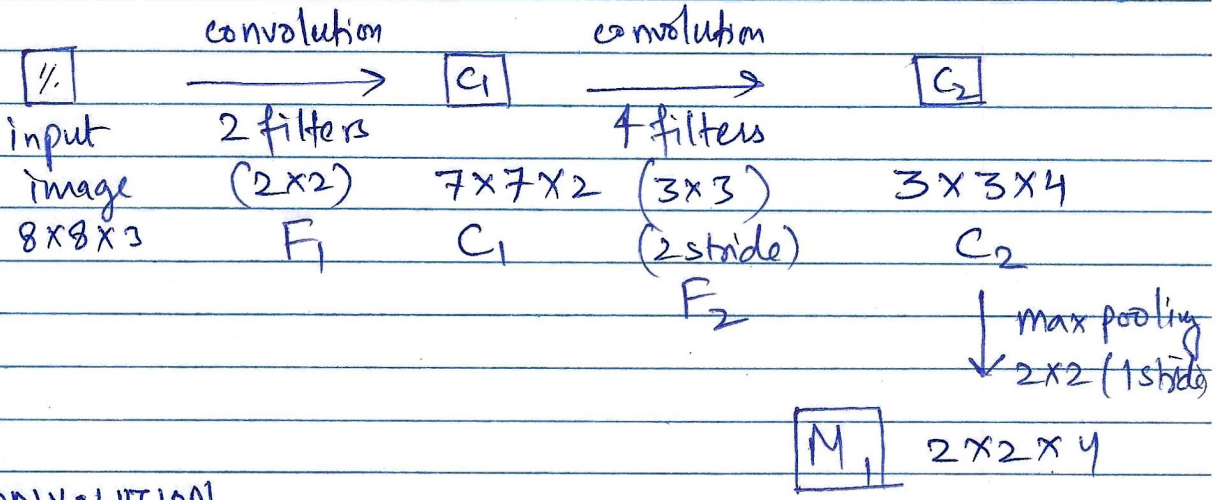
CSCI-567 HW02

BHASKAR GOYAL (USC ID: 6547367383)

PROBLEM - 1.1

1.1 The network is as follows.



CONVOLUTION

$C_1 \Rightarrow (8-2)+1 = 7$ and 2 filters.

So, C_1 has dimensions of $7 \times 7 \times 2$.

$C_2 \Rightarrow (7-4)/2 + 1 = 3$ and 4 filters.

So, C_2 has dimensions of $3 \times 3 \times 4$

MAX POOLING

$M_1 \Rightarrow (3-2)+1 = 2$ and 1 stride, so.

So, M_1 has dimensions of $2 \times 2 \times 4$.

TOTAL PARAMETERS

Parameters (without bais) $= F_1 \times 2 + F_2 \times 4$

$= 2 \times 2 \times 3 \times 2 + 3 \times 3 \times 2 \times 4$

$= 24 + 72$

$= 96$

Parameters (with bais) $= F_1 \times 2 + F_2 \times 4 +$ No. of filters.

$= (2 \times 2 \times 3 \times 2) + (3 \times 3 \times 2 \times 4) + (2+4)$

$= 96 + 6$

$= 102$

Hence,

total params (without bais) = 96

total params (with bais) = 102

(1.2) PROBLEM 1.2

Dimension after first convolution.

$C_1 \Rightarrow (8-2)+1 = 7$ and 2 filters.

So,

C_1 has dimensions of $7 \times 7 \times 2$.

Dimensions after second convolution.

$C_2 \Rightarrow$ ~~$(3-2)+1 = 2$~~

$(7-4)/2 + 1 = 3$ and 4 filtess

So,

C_2 has dimensions of $3 \times 3 \times 4$.

Dimensions after Max Pooling.

$M_1 \Rightarrow (3-2)+1 = 2$ and 1 stride,

So,

M_1 has dimensions of $2 \times 2 \times 4$.

∴ Final Dimension of Output is $2 \times 2 \times 4$.

PROBLEM 2.1

(2.1) GIVEN: $K(x_1, x_2) = (f(x_1) + f(x_2))^2$

To show that K is a valid kernel or not, let's consider 2 training points x_1 and x_2 only.

If K is valid kernel, then Kernel matrix K should be a positive semi definite.

So,

$u^T K u \geq 0$, where $u = \begin{pmatrix} a \\ b \end{pmatrix}$

$$u^T A u = \begin{pmatrix} a & b \end{pmatrix} \begin{pmatrix} (f(x_1)+f(x_1))^2 & (f(x_1)+f(x_2))^2 \\ (f(x_2)+f(x_1))^2 & (f(x_2)+f(x_2))^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$

where $a, b \in \mathbb{R}$

$$= \begin{pmatrix} a & b \end{pmatrix} \begin{pmatrix} 4a(f(x_1))^2 + b(f(x_1)+f(x_2))^2 \\ a(f(x_2)+f(x_1))^2 + 4b(f(x_2))^2 \end{pmatrix}$$

$$= 4a^2(f(x_1))^2 + 2ab(f(x_1)+f(x_2))^2 + 4b^2(f(x_2))^2 \quad \text{---(1)}$$

The above should be ≥ 0 to be K to be a valid kernel, $\forall a, b \in \mathbb{R}$

But this is < 0 for the below values

$f(x) = x$

$x_1 = 2$

$x_2 = 3$

So, equation (1) becomes,

$= 16a^2 + 50ab + 36b^2$

This is < 0 for $a = 15$ and $b = -10$

$= 16(15)^2 + 50(15)(-10) + 36(-10)^2$

$= -300 < 0$

Hence, we have found values of some a, b which make $u^T A u < 0$,

Therefore, K is not a valid kernel as we have found a set of x that breaks the condition of positive semi-definiteness of the Gram matrix.

$\therefore K(x_1, x_2) = (f(x_1) + f(x_2))^2$

is NOT a valid kernel.

PROBLEM 2.2

(2.2) GIVEN: $k(x_1, x_2)$ is a valid kernel.

To Prove: $f(K(x_1, x_2)) = \sum_{i=0}^{p} c_i k^i(x_1, x_2)$; $c_i \geq 0$

Proof:

We know that $k(x_1, x_2)$ is a valid kernel,

so

$k(x_1, x_2) \cdot k(x_1, x_2)$ should also be a valid kernel.
$= k^2(x_1, x_2)$

(Because product of 2 kernel functions is also a kernel).

Similarity,

$k^3(x_1, x_2) = k(x_1, x_2) \cdot k^2(x_1, x_2)$ is also a valid kernel.

Also,

$k^p(x_1, x_2)$ should be a valid kernel, as it can be decomposed into a product of valid kernels.

Now, we also know that a linear combination of kernels is a valid kernel if coefficient are positive
$= \alpha k_1(\cdot,\cdot) + \beta k_2(\cdot,\cdot)$ if $\alpha, \beta > 0$.

$\therefore$ $c_0 k^0(x_1, x_2) + c_1 k^1(x_1, x_2) + \ldots + c_p k^p(x_1, x_2)$ is also a valid kernel as it is a linear combination of valid kernels with $c_0, c_1, c_2, \ldots c_p \geq 0$

Hence, $f(K(x_1, x_2)) = \sum_{i=0}^{p} c_i k^i(x_1, x_2)$; $c_i \geq 0$ is a valid kernel.

PROBLEM 3.1

(3.1) GIVEN: $\hat{p}(y=1|x) = \sigma(w^T x + b)$

$$\sigma = \frac{1}{1+e^{-z}}$$

$$L(w,b,x,y) = -y \log \hat{p}(y=1|x) - (1-y) \log [1 - \hat{p}(y=1|x)]$$

Update Rule of w;

$$\Rightarrow w = w - \eta \nabla_w L(w,b,x,y),$$

where

η is the learning rate

L is the loss function.

Let's find the value of $\nabla_w L(w,b,x,y)$

$$\Rightarrow \nabla_w L(w,b,x,y) = \frac{\partial L}{\partial \hat{p}} \cdot \frac{\partial \hat{p}}{\partial z} \cdot \frac{\partial z}{\partial w}$$

$$= \left[\frac{-y}{\hat{p}(y=1|x)} + \frac{(1-y)}{(1-\hat{p}(y=1|x))} \right] \cdot \frac{e^{-z}}{(1+e^{-z})^2} \cdot \frac{\partial}{\partial w}(w^T x + b)$$

where, $z = w^T x + b$

$$= \left[\frac{(-y \times (1+e^{-z})) + (1-y)}{1 - \frac{1}{1+e^{-z}}} \right] \cdot \frac{e^{-z}}{(1+e^{-z})^2} \cdot x$$

$$= \left[-y(1+e^{-z}) + \frac{(1-y)(1+e^{-z})}{e^{-z}} \right] \cdot \frac{e^{-z}}{(1+e^{-z})^2} \cdot x$$

$$= \frac{(1+e^{-z})}{e^{-z}}\left[-ye^{-z} + (1-y)\right] \cdot \frac{e^{-z}}{(1+e^{-z})^2} \cdot x$$

$$= \frac{1 - y(1+e^{-z})}{(1+e^{-z})} \cdot x$$

$$= \left[\frac{1}{(1+e^{-z})} - y\right] \cdot x \qquad \text{, where } z = w^T x + b$$

$$\nabla_w L = \left[\frac{1}{1+e^{-(w^T x+b)}} - y\right] \cdot x$$

$$= \left[\hat{p}(y=1|x) - (y)\right] \cdot x$$

Using the value of $\nabla_w L$ in update rule we get.

$$w = w - \eta \nabla_w L$$

$$w = w - \eta\left(\hat{p}(y=1|x) - y\right) \cdot x$$

∴, the update rule of w is

$$w = w - \eta\left[\hat{p}(y=1|x) - y\right] x$$

PROBLEM 3.2

(3.2) The update rule of w is,

$$w = w - \eta \left[\hat{p}(y=1|x) - y \right] x$$

Where,
$(\hat{p}(y=1|x) - y)$ denotes the error.

When all examples are normalized, ie. $\forall x_i : \|x_i\| = 1$, then SGD makes a large update to the weights when the error is large.

∴, SGD makes large update to the weights when the error is large, which makes w move towards the direction of the optimal value.

PROBLEM - 4

4.1 By applying chain rule,

$$\frac{\partial L}{\partial v_{12}} = \frac{\partial L}{\partial \hat{y}_1} \cdot \frac{\partial \hat{y}_1}{\partial o_1} \cdot \frac{\partial o_1}{\partial v_{12}} + \frac{\partial L}{\partial \hat{y}_2} \cdot \frac{\partial \hat{y}_2}{\partial o_1} \cdot \frac{\partial o_1}{\partial v_{12}}$$

4.2

$$\frac{\partial L}{\partial \hat{y}_i} = \frac{\partial}{\partial \hat{y}_i} \left(-\sum_{i=1}^{2} y_i \log \hat{y}_i \right)$$

$$= -y_i \frac{\partial}{\partial \hat{y}_i} \log \hat{y}_i$$

$$\boxed{\frac{\partial L}{\partial \hat{y}_i} = \frac{-y_i}{\hat{y}_i}}$$

Hence, $\frac{\partial L}{\partial \hat{y}_1} = \frac{-y_1}{\hat{y}_1}$ and $\frac{\partial L}{\partial \hat{y}_2} = \frac{-y_2}{\hat{y}_2}$

4.3

$$\frac{\partial L}{\partial \hat{y}_2} = \frac{-y_2}{\hat{y}_2}$$

Now, $y_2 = 0$ and $y_1 = 1$

So,

$$\frac{\partial L}{\partial \hat{y}_2} = \frac{-y_2}{\hat{y}_2} = 0$$

$$\frac{\partial L}{\partial \hat{y}_1} = \frac{-y_1}{\hat{y}_1} = \frac{-1}{\hat{y}_1}$$

Simplification of part one, using the previous results.

$$\frac{\partial L}{\partial v_{12}} = \frac{\partial L}{\partial \hat{y}_1} \cdot \frac{\partial \hat{y}_1}{\partial o_1} \cdot \frac{\partial o_1}{\partial v_{12}} + \frac{\partial L}{\partial \hat{y}_2} \cdot \frac{\partial \hat{y}_2}{\partial o_1} \cdot \frac{\partial o_1}{\partial v_{12}}$$

$$= -\frac{y_1}{\hat{y}_1} \cdot \frac{\partial \hat{y}_1}{\partial o_1} \cdot \frac{\partial o_1}{\partial v_{12}} + \frac{-y_2}{\hat{y}_2} \cdot \frac{\partial \hat{y}_2}{\partial o_1} \cdot \frac{\partial o_1}{\partial v_{12}}$$

$$= -\frac{1}{\hat{y}_1} \cdot \frac{\partial \hat{y}_1}{\partial o_1} \cdot \frac{\partial o_1}{\partial v_{12}} + 0$$

$$\boxed{\frac{\partial L}{\partial v_{12}} = -\frac{1}{\hat{y}_1} \cdot \frac{\partial \hat{y}_1}{\partial o_1} \cdot \frac{\partial o_1}{\partial v_{12}}} \quad \text{(Simplification)}$$

(4.4)

$$\frac{\partial \hat{y}_1}{\partial o_1} = \frac{\partial}{\partial o_1}\left(\frac{e^{o_1}}{e^{o_1} + e^{o_2} + e^{o_3}}\right)$$

Applying product rule, we get

$$= \frac{e^{o_1}}{e^{o_1} + e^{o_2} + e^{o_3}} + \frac{-e^{o_1} \cdot e^{o_1}}{(e^{o_1} + e^{o_2} + e^{o_3})^2}$$

$$= \left(\frac{e^{o_1}}{e^{o_1} + e^{o_2} + e^{o_3}}\right)\left[1 - \frac{e^{o_1}}{e^{o_1} + e^{o_2} + e^{o_3}}\right]$$

$$\frac{\partial \hat{y}_1}{\partial o_1} = \hat{y}_1(1 - \hat{y}_1)$$

$$\frac{\partial o_1}{\partial v_{12}} = \frac{\partial}{\partial v_{12}} \left(v_{11} h_1 + v_{12} h_2 + v_{13} h_3 \right)$$

$$= \frac{\partial}{\partial v_{12}} \left(v_{12} h_2 + 0 + 0 \right)$$

$$\frac{\partial o_1}{\partial v_{12}} = h_2$$

So,

$$\boxed{\frac{\partial \hat{y}_1}{\partial o_1} = \hat{y}_1 (1 - \hat{y}_1)}$$

$$\boxed{\frac{\partial o_1}{\partial v_{12}} = h_2}$$

4.5

$$\frac{\partial L}{\partial v_{12}} = \frac{\partial L}{\partial \hat{y}_1} \cdot \frac{\partial \hat{y}_1}{\partial o_1} \cdot \frac{\partial o_1}{\partial v_{12}} + \frac{\partial L}{\partial \hat{y}_2} \cdot \frac{\partial \hat{y}_2}{\partial o_1} \cdot \frac{\partial o_1}{\partial v_{12}}$$

$$= \frac{-y_1}{\hat{y}_1} \cdot \hat{y}_1 (1 - \hat{y}_1) \cdot h_2 + \frac{-y_2}{\hat{y}_2} \cdot \frac{\partial \hat{y}_2}{\partial o_1} \cdot \frac{\partial o_1}{\partial v_{12}}$$

$$= -y_1 (1 - \hat{y}_1) \cdot h_2 + 0$$

$$\boxed{\frac{\partial L}{\partial v_{12}} = (\hat{y}_1 - 1) h_2}$$

SGD rule to update v_{12} is,

$$v_{12} = v_{12} - \eta \frac{\partial L}{\partial v_{12}}$$

$$\boxed{v_{12} = v_{12} - \eta (\hat{y}_1 - 1) h_2}$$

Hence, this is how SGD update rule does to v_{12} w.r.t input h_2 as it is linearly dependent and also depends on error rate $(\hat{y}_1 - 1)$.

4.6 Applying chain rule to get $\frac{\partial L}{\partial v_{23}}$.

$$\boxed{\frac{\partial L}{\partial v_{23}} = \frac{\partial L}{\partial \hat{y}_1} \cdot \frac{\partial \hat{y}_1}{\partial o_2} \cdot \frac{\partial o_2}{\partial v_{23}} + \frac{\partial L}{\partial \hat{y}_2} \cdot \frac{\partial \hat{y}_2}{\partial o_2} \cdot \frac{\partial o_2}{\partial v_{23}}}$$

4.7 We know that from part two.

$$\frac{\partial L}{\partial \hat{y}_i} = -\frac{y_i}{\hat{y}_i}$$

So,

$$\boxed{\frac{\partial L}{\partial v_{23}} = \frac{-y_1}{\hat{y}_1} \cdot \frac{\partial \hat{y}_1}{\partial o_2} \cdot \frac{\partial o_2}{\partial v_{23}} + \frac{-y_2}{\hat{y}_2} \cdot \frac{\partial \hat{y}_2}{\partial o_2} \cdot \frac{\partial o_2}{\partial v_{23}}}$$

4.8

$$\frac{\partial \hat{y}_1}{\partial o_2} = \frac{\partial}{\partial o_2}\left(\text{softmax}(o_1)\right)$$

$$= \frac{\partial}{\partial o_2}\left(\frac{e^{o_1}}{e^{o_1}+e^{o_2}+e^{o_3}}\right)$$

$$= \frac{-e^{o_1}e^{o_2}}{(e^{o_1}+e^{o_2}+e^{o_3})^2}$$

$$= -\hat{y}_1\hat{y}_2$$

$$\boxed{\frac{\partial \hat{y}_1}{\partial o_2} = -\hat{y}_1\hat{y}_2}$$

$$\frac{\partial o_2}{\partial v_{23}} = \frac{\partial}{\partial v_{23}}\left(v_{21}h_1 + v_{22}h_2 + v_{23}h_3\right)$$

$$= \frac{\partial}{\partial v_{23}}\left(v_{23}h_3\right) + 0 + 0$$

$$= \frac{\partial}{\partial v_{23}}\left(v_{23}h_3\right)$$

$$= h_3$$

$$\boxed{\frac{\partial o_2}{\partial v_{23}} = h_3}$$

(4.9) $$\frac{\partial L}{\partial r_{23}} = \frac{\partial L}{\partial \hat{y}_1} \cdot \frac{\partial \hat{y}_1}{\partial o_2} \cdot \frac{\partial o_2}{\partial r_{23}} + \frac{\partial L}{\partial \hat{y}_2} \cdot \frac{\partial \hat{y}_2}{\partial o_2} \cdot \frac{\partial o_2}{\partial r_{23}}$$

Calculating

$$\frac{\partial \hat{y}_1}{\partial o_2} = \frac{\partial}{\partial o_2}\left(\frac{e^{o_1}}{e^{o_1}+e^{o_2}+e^{o_3}}\right)$$

$$= \frac{-e^{o_1} e^{o_2}}{(e^{o_1}+e^{o_2}+e^{o_3})^2}$$

$$= -\hat{y}_1 \hat{y}_2$$

Hence,

$$\frac{\partial L}{\partial r_{23}} = \frac{-y_1}{\hat{y}_1}(-\hat{y}_1 \hat{y}_2)(h_3) + \frac{(-y_2)}{\hat{y}_2}(\hat{y}_2)(1-\hat{y}_2)h_3$$

Putting values $y_1 = 1$ and $y_2 = 0$

$$= +\hat{y}_2 h_3 + 0$$

$$\boxed{\frac{\partial L}{\partial r_{23}} = \hat{y}_2 h_3}$$

4.10

SGD update rule for v_{23},

$$v_{23} = v_{23} - \eta \frac{\partial L}{\partial v_{23}}$$

$$\boxed{v_{23} = v_{23} - \eta\, \hat{y}_2 h_3}$$

And we also know from part 4.5, that SGD rule for v_{12} is..

$$\boxed{v_{12} = v_{12} - \eta(\hat{y}_1 - 1) h_2}$$

We can clearly see that v_{23} is dependent on h_3 and $\hat{y}_2$, while v_{12} is dependent on h_2 and $\hat{y}_1$.

∴ , v_{23} is linear dependent on h_3 and $\hat{y}_2$ while v_{12} is linear dependent on h_2 and $\hat{y}_1$.

So, v_{23} depends on error rate $(\hat{y}_2 - 0)$ where $y_2 = 0$

and

v_{12} depends on error rate $(\hat{y}_1 - 1)$ where $y_1 = 1$ and $y_2 = 0$.

∴ v_{12} depends on the error in predicting y_1,
v_{23} depends on the error in predicting y_2