Name: Ryan Toh
Student ID: 1005129

(Note: Question 1 is in the attached `.ipynb` file.)

Question 2: Neural Networks

2.1:

$$\hat{y} = \sigma(\mathbf{z}_2), \mathbf{z}_2 = \mathbf{w}_2\mathbf{a}_1 + b_2$$

Differentiating $\mathbf{z}_2$ with respect to $\mathbf{w}_2$:

$$\frac{\partial \mathbf{z}_2}{\partial \mathbf{w}_2} = \mathbf{a}_1 \tag{2.1.1}$$

Differentiating $\mathbf{z}_2$ with respect to b_2:

$$\frac{\partial \hat{y}}{\partial b_2} = 1 \tag{2.1.2}$$

$$\begin{aligned}\frac{\partial \hat{y}}{\partial \mathbf{w}_2} &= \frac{\partial \hat{y}}{\partial \mathbf{z}_2}\frac{\partial \mathbf{z}_2}{\partial \mathbf{w}_2} \\ &= \mathbf{a}_1 \frac{\partial \hat{y}}{\partial \mathbf{z}_2} \quad \text{using (2.1.1)}\end{aligned} \tag{2.1.3}$$

$$\begin{aligned}\frac{\partial \hat{y}}{\partial b_2} &= \frac{\partial \hat{y}}{\partial \mathbf{z}_2}\frac{\partial \mathbf{z}_2}{\partial b_2} \\ &= \frac{\partial \hat{y}}{\partial \mathbf{z}_2} \quad \text{using (2.1.2)}\end{aligned} \tag{2.1.4}$$

2.2:

Assume the activation function is the softplus function:

$$\begin{aligned}\hat{y} &= \sigma(\mathbf{z}_2) \\ &= \log(1 + e^{\mathbf{z}_2})\end{aligned}$$

Differentiating $\hat{y}$ with respect to $\mathbf{z}_2$, we have:

$$
\begin{aligned}
\frac{\partial \hat{y}}{\partial \mathbf{z}_2} &= \frac{\partial}{\partial \mathbf{z}_2}\Big(\log(1+e^{\mathbf{z}_2})\Big) \\
&= \frac{1}{1+e^{\mathbf{z}_2}}(e^{\mathbf{z}_2}) \\
&= \frac{e^{\mathbf{z}_2}}{1+e^{\mathbf{z}_2}} \\
&= \frac{1}{1+e^{-\mathbf{z}_2}}
\end{aligned} \tag{2.2.1}
$$

$$
\begin{aligned}
\frac{\partial \hat{y}}{\partial \mathbf{w}_2} &= \mathbf{a}_1 \frac{\partial \hat{y}}{\partial \mathbf{z}_2} \quad \text{using (2.1.3)} \\
&= \frac{\mathbf{a}_1}{1+e^{-\mathbf{z}_2}} \quad \text{using (2.2.1)}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \hat{y}}{\partial b_2} &= \frac{\partial \hat{y}}{\partial \mathbf{z}_2} \quad \text{using (2.1.4)} \\
&= \frac{1}{1+e^{-\mathbf{z}_2}} \quad \text{using (2.2.1)}
\end{aligned}
$$

2.3:

No, $\frac{\partial \hat{y}}{\partial \mathbf{x}}$ does not change.

$$
\begin{aligned}
\hat{y} = \sigma(\mathbf{z}_2), \mathbf{z}_2 &= \mathbf{w}_2\mathbf{a}_1 + b_2 \\
\mathbf{a}_1 = \sigma(\mathbf{z}_1), \mathbf{z}_1 &= \mathbf{w}_1\mathbf{x} + b_1
\end{aligned}
$$

$$
\frac{\partial \hat{y}}{\partial \mathbf{x}} = \frac{\partial \hat{y}}{\partial \mathbf{z}_2}\frac{\partial \mathbf{z}_2}{\partial \mathbf{a}_1}\frac{\partial \mathbf{a}_1}{\partial \mathbf{z}_1}\frac{\partial \mathbf{z}_1}{\partial \mathbf{x}} \tag{2.3.1}
$$

Differentiate $\mathbf{z}_2$ with respect to $\mathbf{a}_1$:

$$
\frac{\partial \mathbf{z}_2}{\partial \mathbf{a}_1} = \mathbf{w}_2 \tag{2.3.2a}
$$

Differentiate $\mathbf{a}_1$ with respect to $\mathbf{z}_1$:

$$\begin{aligned}\mathbf{a}_1 &= \sigma(\mathbf{z}_1)\\ &= \log(1+e^{\mathbf{z}_1})\end{aligned}$$

$$\begin{aligned}\frac{\partial \mathbf{a}_1}{\partial \mathbf{z}_1} &= \frac{\partial}{\partial \mathbf{z}_1}\left(\log(1+e^{\mathbf{z}_1})\right)\\ &= \frac{1}{1+e^{-\mathbf{z}_1}} \quad \text{similar to (2.2.1)}\end{aligned} \tag{2.3.2b}$$

Differentiate $\mathbf{z}_1$ with respect to $\mathbf{x}$:

$$\frac{\partial \mathbf{z}_1}{\partial \mathbf{x}} = \mathbf{w}_1 \tag{2.3.2c}$$

Using $(2.2.1)$, $(2.3.2\text{a})$, $(2.3.2\text{b})$ and $(2.3.2\text{c})$ on $(2.3.1)$:

$$\begin{aligned}\frac{\partial \hat{y}}{\partial \mathbf{x}} &= \frac{\partial \hat{y}}{\partial \mathbf{z}_2}\frac{\partial \mathbf{z}_2}{\partial \mathbf{a}_1}\frac{\partial \mathbf{a}_1}{\partial \mathbf{z}_1}\frac{\partial \mathbf{z}_1}{\partial \mathbf{x}}\\ &= \left(\frac{1}{1+e^{-\mathbf{z}_2}}\right)(\mathbf{w}_2)\left(\frac{1}{1+e^{-\mathbf{z}_1}}\right)(\mathbf{w}_1)\\ &= \frac{\mathbf{w}_1\mathbf{w}_2}{(1+e^{-\mathbf{z}_2})(1+e^{-\mathbf{z}_1})}\end{aligned}$$

Since there is no b_2 value in this expression, a change in bias Δb_2 will not change the overall value.

Hence, $\frac{\partial \hat{y}}{\partial \mathbf{x}}$ does not change.

2.4:

Assume activation function is the logistic function:

$$\begin{aligned}\hat{y} &= \sigma(\mathbf{z}_2)\\ &= \frac{1}{1+e^{-\mathbf{z}_2}}\end{aligned} \tag{2.4.1}$$

$$\begin{aligned}\hat{y} &= \sigma(\mathbf{z}_2), \mathbf{z}_2 = \mathbf{w}_2\mathbf{a}_1 + b_2\\ \mathbf{a}_1 &= \sigma(\mathbf{z}_1), \mathbf{z}_1 = \mathbf{w}_1\mathbf{x} + b_1\end{aligned}$$

$$\frac{\partial \hat{y}}{\partial \mathbf{w}_1} = \frac{\partial \hat{y}}{\partial \mathbf{z}_2}\frac{\partial \mathbf{z}_2}{\partial \mathbf{a}_1}\frac{\partial \mathbf{a}_1}{\partial \mathbf{z}_1}\frac{\partial \mathbf{z}_1}{\partial \mathbf{w}_1} \tag{2.4.2}$$

From $2.4.1$, we differentiate $\hat{y}$ with respect to $\mathbf{z}_2$:

$$\begin{aligned}
\frac{\partial \hat{y}}{\partial \mathbf{z}_2} &= \frac{\partial}{\partial \mathbf{z}_2}\big((1+e^{-\mathbf{z}_2})^{-1}\big) \\
&= -(1+e^{-\mathbf{z}_2})^{-2}e^{-\mathbf{z}_2}(-1) \\
&= \frac{e^{-\mathbf{z}_2}}{(1+e^{-\mathbf{z}_2})^2} \\
&= \frac{1}{1+e^{-\mathbf{z}_2}} \cdot \frac{e^{-\mathbf{z}_2}}{1+e^{-\mathbf{z}_2}} \\
&= \frac{1}{1+e^{-\mathbf{z}_2}} \cdot \Big(1 - \frac{1}{1+e^{-\mathbf{z}_2}}\Big) \\
&= \hat{y}(1-\hat{y}) \qquad (2.4.3a)
\end{aligned}$$

Differentiate $\mathbf{z}_2$ with respect to $\mathbf{a}_1$:

$$\frac{\partial \mathbf{z}_2}{\partial \mathbf{a}_1} = \mathbf{w}_2 \qquad (2.4.3b)$$

Differentiate $\mathbf{a}_1$ with respect to $\mathbf{z}_1$:

$$\begin{aligned}
\frac{\partial \mathbf{a}_1}{\partial \mathbf{z}_1} &= \sigma(\mathbf{z}_1) \\
&= \mathbf{a}_1(1-\mathbf{a}_1) \quad \text{similar to (2.4.3a)} \qquad (2.4.3c)
\end{aligned}$$

Differentiate $\mathbf{z}_1$ with respect to $\mathbf{w}_1$:

$$\frac{\partial \mathbf{z}_1}{\partial \mathbf{w}_1} = \mathbf{x} \qquad (2.4.3d)$$

Using $(2.4.3a)$, $(2.4.3b)$, $(2.4.3c)$ and $(2.4.3d)$ on $(2.4.2)$:

$$\begin{aligned}
\frac{\partial \hat{y}}{\partial \mathbf{w}_1} &= \frac{\partial \hat{y}}{\partial \mathbf{z}_2}\frac{\partial \mathbf{z}_2}{\partial \mathbf{a}_1}\frac{\partial \mathbf{a}_1}{\partial \mathbf{z}_1}\frac{\partial \mathbf{z}_1}{\partial \mathbf{w}_1} \\
&= \big(\hat{y}(1-\hat{y})\big)(\mathbf{w}_2)\big(\mathbf{a}_1(1-\mathbf{a}_1)\big)(\mathbf{x}) \\
&= \mathbf{a}_1\mathbf{w}_2\mathbf{x}\hat{y}(1-\mathbf{a}_1)(1-\hat{y})
\end{aligned}$$

Question 3:

Let feature vector $x \in X$ be such that:

- x_0 represents *'Refund'*, i.e. $x_0 \in \{\text{"Yes"}, \text{"No"}\}$
- x_1 represents *'Martial Status'*, i.e. $x_1 \in \{\text{"Single"}, \text{"Married"}, \text{"Divorced"}\}$
- x_2 represents *'Taxable Income'*.

We want to predict label Y, which represents whether or not an individual would evade taxes.

Given a new individual $x^{(n+1)}$, we want to classify them with the *maximum a posteriori estimate* decision rule, by finding the class label $\hat{Y}$ that maximises the *posterior probability* $p(Y \mid X)$.

In other words, we want to find $\hat{Y}$ so:

$$\hat{Y} = \underset{Y}{\operatorname{argmax}}\ p(Y \mid X)$$

With Bayes' Theorem, we have:

$$\begin{aligned} p(Y \mid X) &= \frac{p(X \mid Y)\, p(Y)}{p(X)} \\ &\propto p(X \mid Y)\, p(Y) \end{aligned}$$

- We can remove $p(X)$ as it is constant given the input, and we only care about the proportionality since we are trying to find $\hat{Y}$.
- With the *naïve assumption*, we assume that features X are *conditionally independent given* class label Y, hence:

$$p(X \mid Y) = \prod_{j=1}^{n} p(X_j \mid Y)$$

Hence, we have:

$$\begin{aligned}\hat{Y} &= \underset{Y}{\text{argmax}}\ p(Y \mid X) \\ &= \underset{Y}{\text{argmax}}\ p(Y) \prod_{j=1}^{n} p(X_j \mid Y)\end{aligned}$$

Since x_2 is assumed to follow a class-conditional normal distribution, the probability density function is likely close to zero. Hence, we need to use the *log-sum-exp trick*.

$$\begin{aligned}\hat{Y} &= \underset{Y}{\text{argmax}}\ p(Y \mid X) \\ &= \underset{Y}{\text{argmax}}\ \log p(Y \mid X) \quad \text{(proportionality)} \\ &= \underset{Y}{\text{argmax}}\ \log p(Y) + \sum_{j=1}^{n} \log p(X_j \mid Y)\end{aligned}$$

The *prior* $p(Y)$ can be found:

$$\begin{aligned}p(Y = \text{"Yes"}) &= \frac{\sum_{i=1}^{m} \mathbb{1}\{y^{(i)} == \text{"Yes"}\}}{m} \\ &= \frac{3}{10} \\ p(Y = \text{"No"}) &= \frac{\sum_{i=1}^{m} \mathbb{1}\{y^{(i)} == \text{"No"}\}}{m} \\ &= \frac{7}{10}\end{aligned}$$

For $j = 0$ (i.e. '*Refund*'), we have the following...

Frequency Table (Refund):

	$Y = \text{"Yes"}$	$Y = \text{"No"}$
$x_0 = \text{"Yes"}$	0	3
$x_0 = \text{"No"}$	3	4
Total	3	7

Likelihood Table (Refund): $p(X_0 \mid Y)$

	$Y =$ "Yes"	$Y =$ "No"
$x_0 =$ "Yes"	0/3	3/7
$x_0 =$ "No"	3/3	4/7
Total	3/3	7/7

Since the naïve prediction requires that each conditional probability is zero, we will need to perform *Laplace smoothing* on the likelihood table for when $Y =$ "Yes". We set $\alpha = 1$ in this case.

	$Y =$ "Yes"	$Y =$ "No"
$x_0 =$ "Yes"	1/5	3/7
$x_0 =$ "No"	4/5	4/7
Total	5/5	7/7

For $j = 1$ (i.e. '*Marital Status*'), we have...

Frequency Table (Marital Status):

	$Y =$ "Yes"	$Y =$ "No"
$x_1 =$ "Single"	2	2
$x_1 =$ "Married"	0	4
$x_1 =$ "Divorced"	1	1
Total	3	7

Likelihood Table (Marital Status): $p(X_1 \mid Y)$

Again, we perform Laplace smoothing for when $Y =$ "Yes".

	$Y =$ "Yes"	$Y =$ "No"
$x_1 =$ "Single"	3/6	2/7

	$Y = \text{"Yes"}$	$Y = \text{"No"}$
$x_1 = \text{"Married"}$	1/6	4/7
$x_1 = \text{"Divorced"}$	2/6	1/7
Total	6/6	7/7

Finally, for $j = 2$ (i.e. "*Taxable Income*"), we assume:

$$p(X_2 \mid Y = c) \sim \mathcal{N}(\mu_c, \sigma_c^2)$$

- where $c \in \{\text{"Yes"}, \text{"No"}\}$

$$p(X = x_2 \mid Y = c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp(-\frac{(x_2 - \mu_c)^2}{2\pi\sigma_c^2})$$

- where μ_c is the sample mean, $\mu_c = \frac{1}{n_c} \sum_{i=1}^{n_c} x_i$
- where σ_c^2 is the sample variance, $\sigma_c^2 = \frac{1}{n_c - 1} \sum_{i=1}^{n_c} (x_i - \mu_c)^2$

We split the data based on the value of $Y = c$:

$Y = \text{"Yes"}$	$Y = \text{"No"}$
95K	125K
85K	100K
90K	70K
	120K
	60K
	220K
	75K

Based on the data, we calculate the sample mean and variance.

$$
\begin{aligned}
\mu_{c=\text{"Yes"}} &= \frac{1}{n_{c=\text{"Yes"}}} \sum_{i=1}^{n_{c=\text{"Yes"}}} x_i \\
&= \frac{1}{3}(95+85+90)(1000) \\
&= 90\text{K}
\end{aligned}
$$

$$
\begin{aligned}
\mu_{c=\text{"No"}} &= \frac{1}{n_{c=\text{"No"}}} \sum_{i=1}^{n_{c=\text{"No"}}} x_i \\
&= \frac{1}{7}(125+100+\ldots+220+75)(1000) \\
&= 110\text{K}
\end{aligned}
$$

$$
\begin{aligned}
\sigma^2_{c=\text{"Yes"}} &= \frac{1}{n_{c=\text{"Yes"}}-1} \sum_{i=1}^{n_{c=\text{"Yes"}}} (x_i - \mu_{c=\text{"Yes"}})^2 \\
&= \frac{1}{2} \sum_{i=1}^{n_{c=\text{"Yes"}}} (x_i - 90\text{K})^2 \\
&= 25\text{K}
\end{aligned}
$$

$$
\begin{aligned}
\sigma^2_{c=\text{"No"}} &= \frac{1}{n_{c=\text{"No"}}-1} \sum_{i=1}^{n_{c=\text{"No"}}} (x_i - \mu_{c=\text{"No"}})^2 \\
&= \frac{1}{6} \sum_{i=1}^{n_{c=\text{"No"}}} (x_i - 110\text{K})^2 \\
&= 2975\text{K}
\end{aligned}
$$

Hence, we can now predict if an individual is likely to evade taxes based on:

$$
\hat{Y} = \underset{Y}{\text{argmax}}\ \log p(Y) + \sum_{j=1}^{3} \log p(X_j \mid Y)
$$

Given an individual where:

- $x_0 = \text{"Yes"}$
- $x_1 = \text{"Married"}$
- $x_2 = 79\text{K}$

We can find:

$$
\begin{aligned}
\sum_{j=1}^{3} \log p(X_j \mid Y = \text{"Yes"}) &= \log(p(X_0 = \text{"Yes"} \mid Y = \text{"Yes"})) + \\
&\quad \log(p(X_1 = \text{"Married"} \mid Y = \text{"Yes"})) + \\
&\quad \log(p(X_2 = 79\text{K} \mid Y = \text{"Yes"})) \\
&\approx -1.6094 - 1.7918 - 3.2987 \\
&= -6.6999
\end{aligned}
$$

$$
\begin{aligned}
\log p(Y = \text{"Yes"} \mid X) &= \log(p(Y = \text{"Yes"})) + \sum_{j=1}^{3} \log p(X_j \mid Y = \text{"Yes"}) \\
&\approx -1.2040 - 6.6999 \\
&\approx -7.90
\end{aligned}
$$

$$
\begin{aligned}
\sum_{j=1}^{3} \log p(X_j \mid Y = \text{"No"}) &= \log(p(X_0 = \text{"Yes"} \mid Y = \text{"No"})) + \\
&\quad \log(p(X_1 = \text{"Married"} \mid Y = \text{"No"})) + \\
&\quad \log(p(X_2 = 79\text{K} \mid Y = \text{"No"})) \\
&\approx -0.8473 - 0.5596 - 4.9693 \\
&= -6.3762
\end{aligned}
$$

$$
\begin{aligned}
\log p(Y = \text{"No"} \mid X) &= \log(p(Y = \text{"No"})) + \sum_{j=1}^{3} \log p(X_j \mid Y = \text{"No"}) \\
&\approx -0.3567 - 6.3762 \\
&\approx -6.73
\end{aligned}
$$

Hence, we can find $\hat{Y}$:

$$
\begin{aligned}
\hat{Y} &= \underset{Y}{\operatorname{argmax}} \ \log p(Y) + \sum_{j=1}^{3} \log p(X_j \mid Y) \\
&= \text{"No"}
\end{aligned}
$$

since $\log p(Y = \text{"No"} \mid X) > \log p(Y = \text{"Yes"} \mid X)$.