Time: 30 minutes Max marks: 10

Instructions:
- Do not plagiarize. Do not assist your classmates in plagiarism.
- Show your full solution for the questions to get full credit.
- Attempt all questions that you can.
- In the unlikely case a question is not clear, discuss it with an invigilating TA. Please ensure that you clearly include any assumptions you make, even after clarification from the invigilator.

1. The sigmoid function is given as $\sigma(z) = \frac{1}{1+e^{-z}}$. A neural network as a binary classifier can be implemented in multiple ways. Assume that the backbone is connected to an output layer that comprises of (1) a single neuron with a sigmoid activation and (2) two neurons with a softmax activation. You may assume that the last layer of the backbone is 100-dimensional, i.e., has 100 neurons that are densely connected to the output layer in both the cases.
   (a) (1 point) Compute the number of *learnable parameters* for the two different cases of the output layers (1) single neuron with sigmoid activation and (2) two neuron with softmax activation.
   (b) (1 point) Differentiate the sigmoid function and then write the derivative ($\sigma'(z)$) in terms of the sigmoid function $\sigma(z)$ itself.
   (c) (½ point) Another popular activation function is the tanh() function $\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$. What is the range of the output of the tanh activation function?
   (d) (½ point) Write the tanh() function in terms of the sigmoid function $\sigma()$.
   (e) (1 point) Use the derivative ($\sigma'(z)$) from part (b) above to show that the derivative of the tanh() function is $\frac{d}{dz}(\tanh(z)) = 1 - \tanh^2(z)$

Total for Question 1: 4

Solution:

(a) (1 point) Given a 100-dimensional backbone, the number of *learnable parameters* will be equal to:
(1) 100 + 1 = 101 for 100 weight parameters and 1 bias parameter.
(2) 200 + 2 = 202 for 100 weight parameters and 1 bias parameter for each node in the output classification layer.
Note: If an additional 100 bias parameters are considered to be learnable for the backbone, the corresponding numbers will be 201 and 302 for (1) and (2) respectively. For full credit, the parameters should be duly specified.

(b) (1 point)

$$
\begin{aligned}
\sigma(z) &= \frac{1}{1+e^{-z}} \\
\sigma'(z) = \frac{\partial}{\partial z}\sigma(z) &= \frac{\partial}{\partial z}\left(\frac{1}{1+e^{-z}}\right) \\
&= \left(\frac{-1}{(1+e^{-z})^2}\right)\cdot\frac{\partial e^{-z}}{\partial z} \\
&= \left(\frac{-1}{(1+e^{-z})^2}\right)\cdot\frac{\partial}{\partial z}\left(1+e^{-z}\right) \\
&= \left(\frac{-1}{(1+e^{-z})^2}\right)\cdot\left(e^{-z}\cdot(-1)\right) \\
&= \left(\frac{-1}{1+e^{-z}}\right)\cdot\left(\frac{-e^{-z}}{1+e^{-z}}\right) \\
&= \left(\frac{1}{1+e^{-z}}\right)\cdot\left(\frac{e^{-z}}{1+e^{-z}}\right) \\
&= \left(\frac{1}{1+e^{-z}}\right)\cdot\left(\frac{1+e^{-z}-1}{1+e^{-z}}\right) \\
&= \left(\frac{1}{1+e^{-z}}\right)\cdot\left(1-\frac{1}{1+e^{-z}}\right) \\
&= \sigma(z)\cdot(1-\sigma(z))
\end{aligned}
$$

(c) (½ point) As $z \to \infty, \tanh(z) \to 1$, and as $z \to -\infty, \tanh(z) \to -1$, therefore the range of tanh() is $[-1, 1]$.

(d) (½ point)

$$
\begin{aligned}
\tanh(z) &= \frac{e^{z}-e^{-z}}{e^{z}+e^{-z}} \\
&= \frac{e^{z}(1-e^{-2z})}{e^{z}(1+e^{-2z})} \\
&= \frac{(1-1+1-e^{-2z})}{(1+e^{-2z})} \\
&= \frac{2-(1+e^{-2z})}{(1+e^{-2z})} \\
&= \frac{2}{1+e^{-2z}} - \frac{1+e^{-2z}}{1+e^{-2z}} \\
&= \frac{2}{1+e^{-2z}} - 1 \\
&= 2\sigma(2z) - 1
\end{aligned}
$$

(e) (1 point)

$$\begin{aligned}
\tanh(z) &= 2\sigma(2z) - 1 \\
\frac{\partial}{\partial z}(\tanh(z)) &= \frac{\partial}{\partial z}(2\sigma(2z) - 1) \\
&= 2\sigma'(2z) \cdot 2 \\
&= 4\sigma(2z) \cdot (1 - \sigma(2z)) \\
&= 4\sigma(2z) - 4(\sigma(2z))^2 + 1 - 1 \\
&= 1 - (-2 \cdot 2\sigma(2z) \cdot 1 + (2\sigma(2z))^2 + 1) \\
&= 1 - (2\sigma(2z) - 1)^2 \\
&= 1 - \tanh^2(z)
\end{aligned}$$

2. Assume that two points x_1 and x_2 are both independently drawn from a normal (Gaussian) distribution $\mathcal{N}(\mu, \sigma^2)$. Recall that the Gaussian distribution is given by $p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$.
   (a) (1 point) Show that if $\mu = 0$ and σ^2 is constant, the log joint probability of x_1 and x_2 is proportional to $x_1^2 + x_2^2 + \kappa$, where κ is a constant.
   (b) (1 point) Generalize the expression of the log joint probability for n independently drawn variables $(x_1, x_2, \ldots, x_n)$ from a zero-mean Gaussian distribution.

Total for Question 2: 2

Solution:

(a) (1 point) Assuming independently drawn samples, the log joint probability is given by

$$\begin{aligned}
\log(p(x_1, x_2)) &= \log(p(x_1) \cdot p(x_2)) \\
&= \log p(x_1) + \log p(x_2) \\
&= \log\left(e^{\frac{-1}{2}\left(\frac{x_1-0}{\sigma}\right)^2}\right) + \log\left(e^{\frac{-1}{2}\left(\frac{x_2-0}{\sigma}\right)^2}\right) + \kappa \text{ (constant terms)} \\
&= \frac{-1}{2}\left(\frac{x_1}{\sigma}\right)^2 + \frac{-1}{2}\left(\frac{x_2}{\sigma}\right)^2) + \kappa \text{ (constant terms)} \\
&= -\frac{1}{2\sigma^2}\left(x_1^2 + x_2^2\right) + \kappa \text{ (constant terms)}
\end{aligned}$$

(b) (1 point)

$$\begin{aligned}
\log(p(x_1, x_2, \ldots, x_n)) &= \log\left(\prod_{i=1}^{n} p(x_i)\right) \\
&= \sum_{i=1}^{n} \log p(x_i) \\
&= \sum_{i=1}^{n} \log\left(e^{\frac{-1}{2}\left(\frac{x_i - 0}{\sigma}\right)^2}\right) + \kappa' \text{ (constant terms)} \\
&= \sum_{i=1}^{n} \left(-\frac{1}{2}\left(\frac{x_i}{\sigma}\right)^2\right) + \kappa' \text{ (constant terms)} \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2 + \kappa' \text{ (constant terms)}
\end{aligned}$$

You might find it worth noting that the log-joint-probability (also known as the *log likelihood*) of the data samples drawn independently from a zero-mean, fixed variance, Gaussian distribution is proportional to the negative sum of squares of the samples (with some constant values added). From your ML study (either past or when you see it in future) of regression (or from estimation theory), you may recall that the maximum likelihood estimation requires you to minimize the sum of squared errors or *mean of squared errors* (MSE). The probabilistic interpretation of minimizing the MSE loss for regression tasks implicitly assumes that the error in your regressor's outputs is zero-mean, fixed variance, Gaussian!

3. Let $\mathbf{a}, \boldsymbol{b}, \mathbf{c}$ be column vectors with $\mathbf{a}, \mathbf{c} \in \mathbb{R}^n$ and $\boldsymbol{b} \in \mathbb{R}^m$.
   (a) (1 point) Let $\mathbf{B} = \boldsymbol{b}\mathbf{a}^\top$. What is the rank of the matrix $\mathbf{B}$? Can you give a numerical value that is more precise than $\min(m, n)$?
   (b) (1 point) Let $\mathbf{C} = \mathbf{c}\mathbf{a}^\top$. If $\mathbf{a}$ and $\mathbf{c}$ are orthogonal vectors, what is the trace of $\mathbf{C}$? {*Hint*: Recall that the trace is the sum of the diagonal elements of a square matrix.}

Total for Question 3: 2

Solution:

(a) (1 point) You may observe that each column of $\mathbf{B} = \boldsymbol{b}\mathbf{a}^\top$ is a scaled version of $\boldsymbol{b}$ and the scaling factors for each column are the elements of $\mathbf{a}$. Since each column is a multiple of another column, the rank of the matrix $\mathbf{B}$ is necessarily 1, unless $\boldsymbol{b} = 0$ or $\mathbf{a} = 0$, in which case the rank will be necessarily 0.

(b) (1 point) The trace of $\mathbf{C}$ is $\sum_{i=1}^{d} a_i c_i$, where d is the dimensionality of $\mathbf{a}$ and $\mathbf{c}$. We can see this is equal to $\mathbf{a}^\top\mathbf{c}$, which equals zero, since $\mathbf{a}$ and $\mathbf{c}$ are orthogonal. Therefore the trace of $\mathbf{C}$ will be 0.

4. A random variable X can take integer values between $[-2, 2]$ and each value is equally likely.
   (a) (1 point) Compute the Shannon's entropy for this random variable X.
   (b) (1 point) Let $Y = |X|$ be a derived random variable. What is the Shannon's entropy of Y?

Total for Question 4: 2

Solution:

(a) (1 point) Let X take values $x_i \in \{-2, -1, 0, 1, 2\}$. Since each value is equally likely, we have $P(X = x_i) = \frac{1}{5}, i = 1, 2, 3, 4, 5$. Therefore,

$$\begin{aligned} H(X) &= -\sum_{i=1}^{5} P(X = x_i) \log P(X = x_i) \qquad (1) \\ &= -\sum_{i=1}^{5} \frac{1}{5} \log \frac{1}{5} \\ &= -\frac{1}{5} \sum_{i=1}^{5} (-\log 5) \\ &= \log 5 \end{aligned}$$

(b) (1 point) $Y = |X|$ can take values $y_i \in \{0, 1, 2\}$ with probabilities $\{P(Y = 0) = \frac{1}{5}, P(Y = 1) = P(X = 1) + P(X = -1) = \frac{1}{5} + \frac{1}{5} = \frac{2}{5}, P(Y = 2) = P(X = 2) + P(X = -2) = \frac{1}{5} + \frac{1}{5} = \frac{2}{5}\}$. Therefore the entropy of Y is:

$$\begin{aligned} H(Y) &= -\left(\frac{1}{5} \log \frac{1}{5} + \frac{2}{5} \log \frac{2}{5} + \frac{2}{5} \log \frac{2}{5}\right) \\ &= -\left(\frac{1}{5} \log \frac{1}{5} + \frac{4}{5} \log \frac{2}{5}\right) \qquad (2) \end{aligned}$$