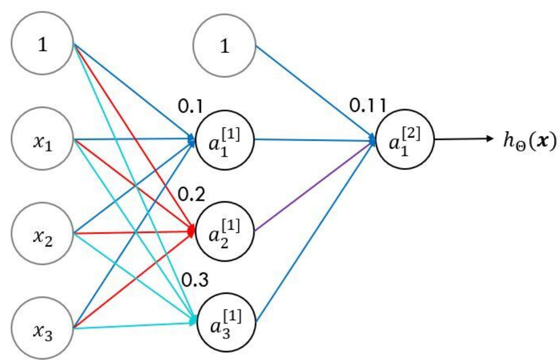
1. (MCQ; 2 marks) We stated that validation generally produces an optimistic estimate of J_{test}. Why?
   (a) Because possibly many parameters are tested in the validation process.
   (b) Because we choose the model based on their performance.
   (c) Because possibly many values of parameters are tested in the validation process.

2. (MRQ; 3 marks) Which of the following statements are true regarding the ethics of data sharing?
   (a) Ensuring the quality of collected data; and holding sources accountable for low-quality or actively misleading data.
   (b) Equitable and ethical access to data once it's collected.
   (c) Transparency around the collection of data and how this collected data will be used.
   (d) Clear provenance of data so that data scientists are always aware of where their datasets come from.

[Questions 3–4] Suppose we are using a neural network with an input vector of length 3, one hidden layer with three neurons and one output neuron. Additionally, the hidden neurons and the input include a bias. We use the ReLU function as the nonlinearity. The basic structure is shown below:



Suppose there is a data input $\mathbf{x} = (1, 2, 3)$ and the actual output label is $\mathbf{y} = (0.8)$. The bias weights are included in the figure, and the remaining weights for the network are:

$$\mathbf{\Theta}^{[1]} = \begin{bmatrix} -0.1 & 0.3 & 0.1 \\ 0.3 & -0.4 & -0.2 \\ 0.2 & -0.2 & 0.2 \end{bmatrix}, \mathbf{\Theta}^{[2]} = \begin{bmatrix} 0.6 & -0.7 & 0.5 \end{bmatrix},$$

3. (MCQ; 3 marks) Calculate the value of $J(\mathbf{a}^{[2]}, \mathbf{y})$ after forward propagation when using squared loss.
   (a) $(J(\mathbf{a}^{[2]}, \mathbf{y}) = 0.4$.
   (b) $(J(\mathbf{a}^{[2]}, \mathbf{y}) = 0.02$.
   (c) $(J(\mathbf{a}^{[2]}, \mathbf{y}) = 0.04$.
   (d) $(J(\mathbf{a}^{[2]}, \mathbf{y}) = 0.03$.
   (e) None of these are correct.

Continuing from the above, if we are given that $\frac{\partial J(\mathbf{a}^{[2]},\mathbf{y})}{\partial a_1^{[2]}} = 0.5$,

4. (MCQ; 3 marks) Calculate the gradient of $J(\mathbf{a}^{[2]}, \mathbf{y})$ with respect to $\Theta_{1,1}^{[2]}$.

   (a) $\frac{\partial J(\mathbf{a}^{[2]},\mathbf{y})}{\partial \Theta_{1,1}^{[2]}} = 0.55$.
   (b) $\frac{\partial J(\mathbf{a}^{[2]},\mathbf{y})}{\partial \Theta_{1,1}^{[2]}} = 0.30$.
   (c) $\frac{\partial J(\mathbf{a}^{[2]},\mathbf{y})}{\partial \Theta_{1,1}^{[2]}} = 0.45$.
   (d) $\frac{\partial J(\mathbf{a}^{[2]},\mathbf{y})}{\partial \Theta_{1,1}^{[2]}} = 0.40$.
   (e) None of these are correct.

[Questions 5–6] Consider the **vanilla** character level classification **Recurrent Neural Network** seen in our Deep Learning (W09) Colab notebook, which takes a name as input and outputs the score over each language.

5. (MRQ with 4 options; 4 marks) Which of the following statements are true?

   (a) It cannot regress continuous output.
   (b) While it can model short term history, it is ineffective at remembering long term states.
   (c) It can only take input of a fixed length.
   (d) It cannot be trained in parallel.

6. (MCQ; 3 marks) If we decide to use truncated backpropagation in the above RNN, rather than the full backpropagation through time (BTT),

   (a) The training time per epoch will be more uniform, regardless of input length.
   (b) Training will necessarily converge faster.
   (c) The weights at the beginning of the sequence will never be trained.
   (d) Incorrect outputs at the beginning of the sequence will count for less in the loss.

[Questions 7–9] (MCQ; 2 marks each) Mark (a) for true and (b) for false for each of the following statements on **Decision Trees**.

Let's examine decision tree learning as taught in lecture, with categorical inputs X and output Y. Here we assume we do not employ pruning.

7. The depth of the tree cannot exceed $n + 1$.

8. If $IG(Y|X_i) = 0$, then X_i will not be used in the decision tree.

9. Suppose one of the attributes has a unique value in each instance. Then the decision tree must have depth 0 or 1.

10. (MRQ with 4 options; 3 marks) Which of the following statements are true regarding Principal Component Analysis (PCA)?
    (a) We can visualize our high-dimensional data by using PCA to project them to a low-dimensional space.
    (b) In PCA, we should select the principal components with minimum variance.
    (c) Before using PCA, we should perform feature normalization.
    (d) In PCA, we should select the principal components with maximum variance.

11. (MRQ with 4 options; 3 marks) Which of the following factors can impact the accuracy of clustering?
    (a) Algorithm, e.g., use K-Means or hierarchical clustering.
    (b) Distance metric, such as Euclidean distance and Manhattan distance.
    (c) Feature selection.
    (d) The quality of labels.

Assume the following dataset of data points is given: $(0,4)$, $(2,2)$, $(4,0)$, $(4,4)$, $(6,6)$ and $(10,10)$. K-Means is run with $k = 3$ to cluster the dataset. Moreover, Euclidean distance is used as the distance function to compute distances between centroids and points in the dataswet. The centroid for a set of n data points $((x_i, y_i), i = 1, 2, \ldots, n)$ can be calculated as $(\frac{\sum_1^n x_i}{n}, \frac{\sum_1^n y_i}{n})$.
At some iteration, C1, C2 and C3 of K-Means are as follows:

- $C1 : \{(2,2),(4,4)\}$
- $C2 : \{(0,4),(4,0)\}$
- $C3 : \{(6,6),(10,10)\}$

12. (Calculation Response; 5 marks) If we run K-Means to completion, what are the a) resulting clusters and b) cluster centroids? Show your work.

13. (Code Response; 7 marks) Write pseudocode or Python code for the bootstrapping method which takes two parameters: a dataset D (a matrix of size $m \times n$) and sample size s (an non-negative integer), and returns a dataset D_tilde (of dimension $s \times n$).

Let us define a **Convolutional Neural Network** with a filter $F1$ and $F2$ for 2 channels as specified below:

$$F1 = \begin{bmatrix} 1 & 1 & -1 \\ -1 & 1 & -1 \\ -1 & 1 & 1 \end{bmatrix}, F2 = \begin{bmatrix} -1 & 1 & 1 \\ -1 & -1 & -1 \\ 1 & -1 & 1 \end{bmatrix}$$

Let us now give a 2-channel input defined by the 2 matrices, respectively:

$$X1 = \begin{bmatrix} 2 & 2 & 0 & 1 & 1 \\ -1 & 1 & -1 & 0 & 2 \\ -1 & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 2 & 2 & 1 & 2 & 0 \end{bmatrix}, X2 = \begin{bmatrix} 2 & -2 & 0 & -1 & 1 \\ -1 & 1 & -2 & 0 & 2 \\ -1 & 0 & 1 & 2 & 2 \\ 0 & 0 & 2 & 1 & 0 \\ 1 & 2 & 1 & -2 & 0 \end{bmatrix}$$

The stride is 1 and there is no padding and let Y be the output matrix. $Y_{i,j}$ refers to the element in the ith row and jth column where we start indexing from 1.

14. (Calculation Response; 10 marks) Calculate the 3 missing values in the output matrix: $Y_{1,1}$, $Y_{2,2}$ and $Y_{3,3}$. Show your work.

15. (Text Response; 3 marks) Name one of the guest stars featured in the lectures and briefly describe their research interests that they mention in their outro video.

This marks the end of this part of the exam.
These is no additional material beyond this point.