ANSWERS VERSION 1

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.

   Correct answers: (b)

   Explanation: As stated in our validation lecture, when performing validation, we **select** a model or its (hyper)parameters based on maxmizing its performance on validation data. This biases the model to perform well on the validation data, as we selected it based on its performance.

   The other options, while possibly true, do not explain the optimistic nature of the J_{test}.

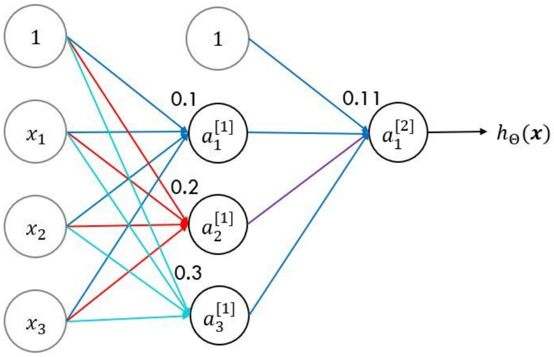
2. (MRQ; 3 marks) Which of the following statements are true regarding the ethics of data sharing?

   (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.

   Correct answers: (a), (b), (c), (d)

   Explanation: All of the options are important criteria to observe when considering data sharing. When collecting data from subjects, the subjects should be duly informed and give consent the data being explicitly collected for the purposes entailed. Data also needs proper document and have a clear provenance on its origin, (pre-)processing to aid downstream users towards its appropriate (and sometimes inappropriate usage) Before sharing the data, the originator should take responsibility for the dataset's quality and engineer collection filtering and/or quality checks that form an integral part of the provenance of the dataset. Both equitable and ethical access should be observed for in disseminating data so that models and results can be replicated in a responsible manner.

[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.

Correct answers: (c)

Explanation: This question first requires the calculation of the final output via forward propagation, so that a loss (if applicable) can be calculated. If there is a loss, then the weights should change to reflect weight updates to lessen the loss. This second phase is calculated through backpropagation.

In the forward propagation stage, the values are as follows:

$a_1^{[1]} = 1 \times -0.1 + 2 \times 0.3 + 3 \times 0.2 + 0.1 = ReLU(0.9) = 0.9,$
$a_2^{[1]} = 1 \times 0.3 + 2 \times -0.4 + 3 \times -0.2 + 0.2 = ReLU(-0.9) = 0,$
$a_3^{[1]} = 1 \times 0.2 + 2 \times -0.2 + 3 \times 0.2 + 0.3 = ReLU(0.7) = 0.7.$

We use these outputs to from $a^{[1]}$ to then calculate the results for $a^{[2]}$:

$a_1^{[2]} = ReLU(0.9 \times 0.6 + 0 \times -0.7 + 0.7 \times 0.5 + 0.11) = 1.0.$

So the predicted output is 1.0, but the correct value as stated in the problem is 0.8, so there is a loss. The loss is $(1.0 - 0.8)^2 = 0.2^2 = 0.04$. So, $J(\mathbf{a}^{[2]}, \mathbf{y}) = 0.04$.

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.

Correct answers: (c)

Explanation: We are given $\frac{\partial J(\mathbf{a}^{[2]},\mathbf{y})}{\partial a_1^{[2]}} = 0.5$, then asked to do backpropagation to obtain the partial gradient with respect to $\Theta_{1,1}^{[2]}$.

This is just taking the upstream partial of 0.5 and multiplying it by $a_1^{[1]}$'s contribution, which we calculated previously as 0.9. So, $0.9 \times 0.5 = 0.45$.

[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.

Correct answers: (a), (b), (d)

Explanation: We go over the true statements before detailing the false one.

RNNs (as taught in the simple, vanilla architecture in class) are used for classification tasks. While RNNs are used for modeling sequential input, due to their single state that must capture both current input and past context, RNNs are not effective for remembering long-term state information. Also, RNNs cannot be trained in parallel as they are recurrent; the current state depends on the previous state's calculation so training must proceed in a sequential (series) fashion.

RNNs most definitely can take sequences of differing lengths; this was the motivation behind the RNN architecture (sequential network), so the claim that it can take only a fixed-length input is incorrect.

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.

Correct answers: (a)

Explanation: Using truncated BTT, rather than do the full backpropagation till the beginning of each sequence we stop after a set number of backpropagation steps or till we have fully backpropagated the loss through an input instance. The latter happens when the sequence is longer than the threshold set of the BTT truncation.

The rest of the statements are false. The weights at the beginning of sufficiently short sequences may be trained. Training may or may not converge faster, depending on the weight updates per epoch. Incorrect outputs in the beginning of the sequence still count the same as later outputs, when they are not trimmed by the truncation.

[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 $n + 1$.

   Correct answers: (a)

   Explanation: True because the attributes are categorical and can each be split only once.

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

   Correct answers: (b)

   Explanation: False (because the attribute may become relevant further down the tree when the instances are restricted to some value of another attribute; e.g. XOR).

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

   Correct answers: (a)

   Explanation: True because that attribute will have perfect information gain. If an attribute has perfect information gain it must split the records into pure buckets which can be split no more.

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.

    Correct answers: (a), (c), (d)

    Explanation: PCA is unsupervised and selects the components with the maximal variance so that the components are ordered to maximize capturing as much information of the dataset (to minimize reconstruction loss). As raw features may be unnormalised, it is important to first normalise the features to unit values so that the variance between dimensions is comparable. We can employ PCA to decompose our dataset and retain only the first principal components to do feature selection and visualize in a lower dimensional setting.

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.

    Correct answers: (a), (b), (c)

    Explanation: Feature selection can change the inputs used for the clustering and hence influences the clustering output. The distance metric also can change how the distance between points are measured, so it also influences the output. The overall clustering algorithm used also can affect the clustering output. However, labels do not concern clustering, as clustering is an unsupervised learning task, and hence their quality is irrelevant.

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.

Explanation: After the first calculation for the centroids' location, we determine $C1$'s centroid as $(4,4)$ and $C2$'s centroid as $(2,2)$. $C3$'s centroid is at $(6,6)$. We then recalculate membership of points to centroids, and we see that the membership of points originally in $C1$ and $C2$ shift such that $C1$ only consist of a single point $(4,4)$, and the remainder of the original points are assigned to $C2$'s centroid at $(2,2)$. $C3$'s centroid does not shift. After this first iteration, the clustering is stable and terminates.

The final answer is:

- $\mathrm{C1} : \{(4,4)\} \rightarrow (4,4)$.
- $\mathrm{C2} : \{(0,4), (4,0), (2,2)\} \rightarrow (2,2)$.
- $\mathrm{C3} : \{(6,6), (10,10)\} \rightarrow (8,8)$.

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.

13. (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.

Explanation: For each cell in Y we calculate each individual filter response's centered over the appropriate input cell and sum together the two responses.

A value in the Y is equal to the sum of all products of all values in the each filter with the corresponding X when the center of the filter is at the required index. Therefore:

For $Y_{1,1}$ = (2*1 + 2*1 + 0*-1 + -1*-1 + 1*1 + -1*-1 + -1*-1 + -2*1 + 1*1) + (2*-1 + -2*1 + 0*1 + -1*-1 + 1*-1 + -2*-1 + -1*1 + 0*-1 + 1*1) = 7 + (-2) = 5

For $Y_{2,2}$ = (1*1 + -1*1 + 0*-1 + -2*-1 + 1*1 + 0*-1 + 0*-1 + 0*1 + 0*1) + (1*-1 + -2*1 + 0*1 + 0*-1 + 1*-1 + 2*-1 + 0*1 + 2*-1 + 1*1) = 3 + (-7) = -4

For $Y_{3,3}$ = (1*1 + 0*1 + 0*-1 + 0*-1 + 0*1 + 0*-1 + 1*-1 + 2*1 + 0*1) + (1*-1 + 2*1 + 2*1 + 2*-1 + 1*-1 + 0*-1 + 1*1 + -2*-1 + 0*1) = 2 + 3 = 5

14. (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.

Explanation: There are many valid responses to this question, two sample answers are given. Prof. Lee Wee Sun's interests are in reinforcement learning and game playing; Prof. Terence Sim's interest are in biometrics and bioauthetication. Terence introduced his research on continuous verification using ensemble techniques.

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