University of Glasgow

Saturday 14th August 2021
(24 hour open online assessment – Indicative duration 2 hours)

DEGREE OF MSc

Deep Learning for MSc (M)
COMPSCI 5103

(Answer ALL 3 questions)

This examination paper is worth a total of 60 marks.

1. Refer to the neural network shown in Figure 1 to answer the following questions. In this diagram, all input values $x_n^{[0]} = 1.0$ and all bias, b, and weight, w, parameters take on the value 0.1.

(a) Give plausible reasons why the network designer has chosen ReLU activation for some neurons in this network while using sigmoid for others. In particular, consider the case if this network was extended to be a much deeper network. [4]

> **Solution:** Sigmoid activation has been used in the output neuron. The most likely reason for this is that the network is doing a binary prediction and thus the sigmoid outputs a value in the range (0.0,1.0) which can be interpreted as the probability of the positive case [1 mark]. Together with binary cross-entropy loss function, this implements a principled maximum likelihood framework which provides suitable gradients for parameter optimization [1 mark]. ReLU activation has been used for internal neurons and this has the key advantage that it doesn't saturate like sigmoid or tanh activation for large and small input values which can cause parameter optimization to stagnate due to the generation of very small gradients [1 mark]. This is particularly significant if the network was extended to be deeper where multiplication of multiple gradients can result in the vanishing gradient problem [1 mark].

(b) This network contains 9 parameters (6 weights and 3 bias). How many passes of the network would be required to determine the gradients for all these parameters using forward autodiff and reverse autodiff. [2]

> **Solution:** The student should realize that forward autodiff and reverse autodiff are different procedures (and these terms do not refer to the forward pass and the reverse pass of the reverse autodiff procedure). The forward autodiff procedure would require 9 forward passes of the network [1 mark] and reverse autodiff would require 2 passes of the network [1 mark].

(c) Carry out a forward pass on this network to determine the output value $y_1^{[2]}$. Show intermediate values for hidden units. [2]

> **Solution:** $h_1^{[1]} = ReLU(0.1+0.1+0.1) = 0.3$, also $h_2^{[1]} = ReLU(0.1+0.1+0.1) = 0.3$ [1 mark]
> $y_1^{[2]} = \sigma(0.03+0.03+0.1) = \sigma(0.16) \approx 0.54$ [1 mark]

(d) (i) We define:

$$z_1^{[2]} = w_{1,1}^{[2]} h_1^{[1]} + w_{2,1}^{[2]} h_2^{[1]} + b_1^{[2]}$$

Do the first step of a backward pass to determine the numerical value of:

$$\frac{\partial y_1^{[2]}}{\partial h_1^{[1]}}$$

(Show your working, employing the $z_1^{[2]}$ term, and also state the final numerical value obtained.) [5]

Solution:
This type of calculation, employing the chain rule of calculus to do backpropagation, has been shown to students in lectures but this particular example is unseen to them.
Given $z_1^{[2]} = w_{1,1}^{[2]} h_1^{[1]} + w_{2,1}^{[2]} h_2^{[1]} + b_1^{[2]}$
Then we have: $y_1^{[2]} = \sigma(z_1^{[2]})$.
Then

$$\frac{\partial y_1^{[2]}}{\partial h_1^{[1]}} = \frac{\partial y_1^{[2]}}{\partial z_1^{[2]}} \frac{\partial z_1^{[2]}}{\partial h_1^{[1]}} = \sigma(z_1^{[2]})(1-\sigma(z_1^{[2]}))w_{1,1}^{[2]} = \sigma(0.16)(1-\sigma(0.16))(0.1) \approx 0.0248$$

[3 marks for working and 2 marks for final value.]

(ii) Continue this backward pass to determine the numerical value of

$$\frac{\partial y_1^{[2]}}{\partial w_{1,1}^{[1]}}$$

showing how this is generated from the previous term you calculated:

$$\frac{\partial y_1^{[2]}}{\partial h_1^{[1]}}$$

(Again show your working, employing the intermediate value $z_1^{[1]}$, and state the final numerical value obtained.) [4]

Solution:
We have $z_1^{[1]} = w_{1,1}^{[1]} x_1^{[0]} + w_{2,1}^{[1]} x_2^{[0]} + b_1^{[1]}$
And: $h_1^{[1]} = ReLU(z_1^{[1]})$.
Then

$$\frac{\partial y_1^{[2]}}{\partial w_{1,1}^{[1]}} = \frac{\partial y_1^{[2]}}{\partial h_1^{[1]}} \frac{\partial h_1^{[1]}}{\partial z_1^{[1]}} \frac{\partial z_1^{[1]}}{\partial w_{1,1}^{[1]}} = \frac{\partial y_1^{[2]}}{\partial h_1^{[1]}} (1.0)(x_1^{[0]}) \approx (0.0248)(1.0)(1.0) \approx 0.0248$$

Essentially it has the same gradient as the previous calculation since input to ReLU unit is positive and x input is 1.0.
[2 marks for working and 2 marks for final value.]

(e) Assume these are the weights and biases set at initialization of a gradient descent optimization. What key problem exists with these initial settings in this case? How could it be resolved? [3]

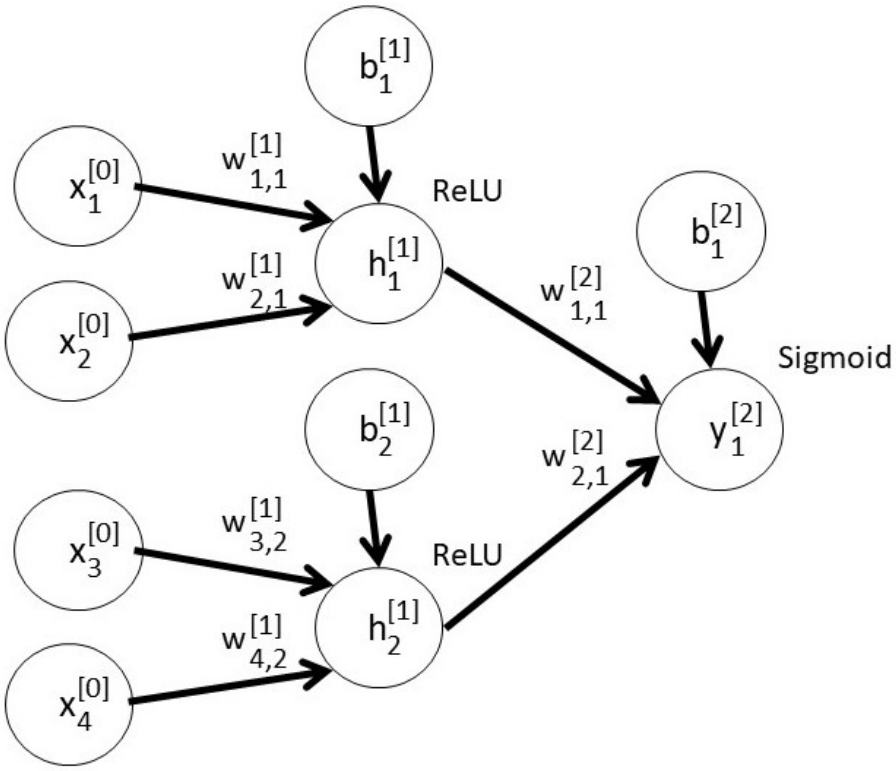

Figure 1: Example feedforward neural network with parameters.

Solution: The key problem is that the weights are perfectly symmetric (all 0.1) and so is the network - so it is unlikely the optimiser would break this symmetry and it may sit at an unstable maximum [2 marks]. This can be very simply resolved by assigning small random weights [1 mark]. The students may give more detail about weight intialization such as the Glorot & Bengio method given in lectures (essentially reproducing bookwork) - but the key idea is recognizing the problem of symmetry and also how it is typically resolved with random initialization.

2. You are developing a deep neural network to detect a rare cancer which can be cured if detected early. The input to the method is microscope images of biopsy tissue. Any positive cases detected would undergo further manual checking to confirm them. You have 10000 microscope images of normal tissue (negative samples) and 100 images of the cancer (positive samples). You randomly sample 2000 normal images and 20 cancer images from these data to produce a test set and you intend to employ 4-fold cross validation using the remaining data to optimize hyperparameters using random search.

(a) You are considering three different measures of success: i) accuracy, ii) precision & recall and iii) F1 score. Explain whether each measure is appropriate given this scenario and justify your answers. [5]

> **Solution:** i) Accuracy is not appropriate since the data is not balanced and you could get high accuracy by just predicting all samples negative (not a useful predictor). [1 mark]
>
> ii) Precision/recall are the most appropriate measures since they successfully deal with unbalanced data. In particular, this scenario would want very high recall (detecting all cancer cases) whereas precision does not need to be particularly high since detected cases go to further manual checking to confirm them. [2 marks]
>
> iii) The F1 score uses precision and recall but this single score cannot be used in this scenario since a high F1 score may indicate high recall and low precision (which is desirable in this case), but it could equally indicate high precision and low recall (which is not desirable). [2 marks]

(b) Explain why you cannot use one of the measures of success given in part (a) as the loss function for the network even though this is what you want to optimize. What loss function should you use in this scenario and what are you hoping by optimising it? [3]

> **Solution:**
>
> The key issue is that these measures are not differentiable (or even continuous) so gradients cannot be determined to drive gradient descent optimization [1 mark]. Loss functions need to be differentiable (although some loss function may have isolated points that are not differentiable such as the ridge loss function). This is clearly a binary classification problem and binary cross-entropy should be used [1 mark]. The hope is that the loss function is a suitable proxy (or surrogate) for the actual score measure and by optimizing the loss function then the score measure will also be optimized (although it need not be) [1 mark].

(c) (i) You adjust some common hyperparameters during hyperparameter optimization. One of them generates the loss curves given in Figure 2 as it is adjusted. What is the hyperparameter you are most likely adjusting here and explain whether it is taking a large/medium/small value for the given loss curves A), B) and C). Which of these curves shows the best hyperparameter value and explain why this particular value

results in the best learning. [3]

> **Solution:** This hyperparameter is likely to be the learning rate [1 mark]. The numerical instability/explosion in A) indicates a high learning rate, B) has an appropriate medium learning rate and C) has a small learning rate [1 mark]. B) is again the most appropriate value since it is small enough to be numerically stable (unlike A) but it is sufficiently large to make progress during optimization (unlike C) [1 mark].

(ii) What hyperparameter are you most likely adjusting in Figure 3 and explain whether it is taking a large/medium/small value in the cases of loss curves A), B) and C). Which of these curves shows the best hyperparameter value and explain why this particular value results in the best learning. [3]

> **Solution:** This hyperparameter is likely to be the mini-batch size [1 mark]. The higher noise in A) indicates a small mini-batch size, with B) having a medium mini-batch size and the smooth C) having a large mini-batch size [1 mark]. B) reaches the lowest loss since the gradient from the mini-batch is sufficiently accurate while the introduced randomness in the gradients help avoid local minima so it finds a lower minima loss value [1 mark].

(d) How should you generate your final model from cross-validation and determine its overall final score? Be explicit about the number and size of training and validation sets used, how the best hyperparameter values would be determined, and how the final model would be generated and scored using the data. [6]

> **Solution:**
> The remaining data (after the test set has been extracted) consists of 80,000 negative samples and 80 positive samples. 4-fold cross validation would split this dataset into 4 parts each consisting of 20,020 samples (with appropriate stratification to ensure each has equal numbers of negative and positive samples) [1 mark]. It says random hyperparameter search is used so for each run, random hypervalues within certain ranges will be chosen for all the hyperparameters for the run [1 mark], and then a full cross validation run would be carried out (i.e. the model would be trained 4 times and tested each time using the chosen measure) with the overall mean success measure value across the four models used to determine the best hyperparameter values [1 mark]. A final model would then be generated using *all the training data* (i.e. no validation data) using the best hyperparameter values determined previously [2 marks] and the unused test set would be used to determine its overall score [1 mark].

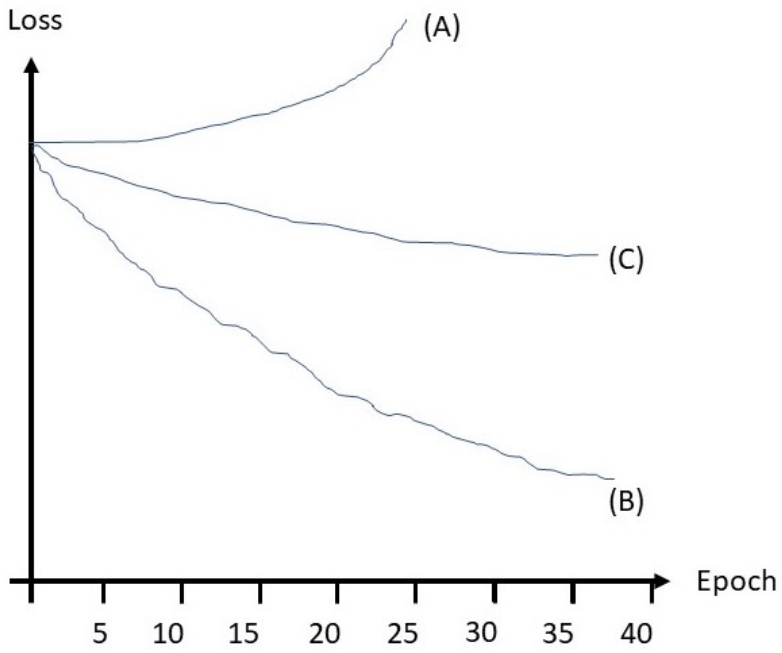

Figure 2: Example loss curves when changing a particular hyperparameter.

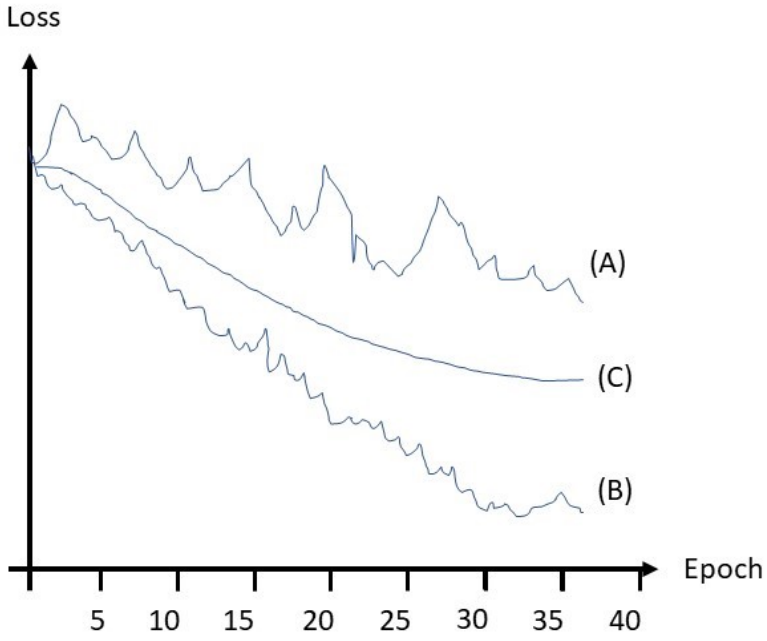

Figure 3: Example loss curves when changing a different hyperparameter.

3. A hospital has a large collections of images which could form a training set for a particular diagnostic classification task. However, they have only been able to afford to label 10% of the datapoints. The diagnostic system is intended to analyse thousands of people every day and speed up diagnosis by making decisions on cases it is certain about, and referring cases it is uncertain about to a human specialist. The baseline rate for a positive diagnosis which requires medical intervention is 1% of tests.

(a) Describe a way that the hospital could use the unlabelled data to improve the classification performance. [3]

Solution: Unsupervised pretraining: They could set up an autoencoding task for the images where they generate an image $\hat{x}_i$ when fed an image x_i [1]. The bottleneck layer z of the network can be used as the input to a classification task [1], and this would allow the early feature detecting layers to benefit from many more examples [1]. This is an obvious example. If students come up with plausible alternatives, credit should be given.

(b) The manual labelling of the image classes in the training set is subject to uncertainty. How could this be taken into account in the training process, and what advantages would it bring? [4]

Solution: Example of label smoothing: we explicitly model the noise in the labels, incorporating the assumption of an error rate e into the cost function analytically, rather than sampling, via label smoothing [1]. Label smoothing regularises a softmax-based model by using targets of $e/(k-1)$ and $(1-e)$ rather than 0 and 1.[1] An advantage is that a softmax function can never predict 0 or 1 – the weights just keep getting larger as it asymptotically approaches the limit [1]. This can be prevented by using regularisation like l_2, but label smoothing has the advantage of preventing hard probabilities without discouraging correct classification [1].

(c) Give three examples of visualisation methods that could help a doctor interpret the decision making process of a deep network on a specific image. Explain the limitations of this approach to explaining classification behaviour. [5]

Solution: Saliency maps look at the derivative of the output class wrt to the input features. They can show where the network is sensitive, but are noisy [1]. Deconvolution with the guided back-propagation (derivative but filtering structures which do not contribute positively) can improve results [1]. Class activation heat maps (CAMs) visualise class-discriminative regions of the image [1]. These methods typically show the regions of the image the classification is sensitive to, which allows detection of obvious bias in a classifier looking at the wrong features, but does not make it obvious how exactly the features in that region are used [2]. Students may mention neural attention mechanisms which learn to provide weightings to information that is relevant

for a decision, and that would be fine as well.

(d) A university is considering using a deep machine learning system to rank applicants for admission based on a set of features which describe the candidate. As a training set they plan to use historical performance data from students accepted in the past, for whom they have the same feature data. It will give students instant feedback about their probability of being accepted, and the success prediction will be shared with staff to help them allocate support for students. Once the system is running, accepted students and their performance will be automatically added to the training data for future classifications.

What are the ethical and practical risks associated with the use of machine learning in such an admission system? [8]

Solution: There are multiple aspects that students could comment on. Arguments made should cover elements following 4 main topics: Bias in Training data, Impact on specific Groups, Feedback issues, Explainability. (i.e. it is not enough to list 8 points in one topic to get 8 marks, they should cover relevant elements from each topic). A first class answer (6-8 marks) should cover the full range of topics in coherent and well-argued manner. Some examples of specific arguments are given below, but this list is not exclusive - other appropriate arguments are possible. A pass mark (4 marks) might just have one topic from each area, with a lacklustre or unbalanced presentation of the argument.

Here are some examples. Training data bias: historical examples of the relevant outcomes will almost always reflect historical prejudices against certain social groups, prevailing cultural stereotypes, and existing demographic inequalities [1], so the use of the system may sustain or exacerbate these inequalities [1]. Already underrepresented groups will have less data about them, so the system will be less reliable with their predictions [1]. Potential for gaming the system: instant feedback means people or schools can potentially learn to adapt their submissions until successful [1], similarly some of the features may depend on other algorithms which change independently over time (e.g. school grading systems getting easier, or changing in a given year because of a special incident such as the covid pandemic) [1]. Feedback: If there is a bias in the network, adding accepted students to the training set has the potential to increase that bias over time [1]. Self-fulfilling predictions: Fact that staff and students themselves are given prediction of success means that this may bias their expectations in a way that is detrimental for performance [1]. Explainability: An ethical decision-making process might require the ability to explain a prediction or decision, which might not be feasible with complex DL black-box models [1].