MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY, WEST BENGAL

Paper Code : PCCAIML602 Deep Learning

UPID : 006917

Time Allotted : 3 Hours *Full Marks :70*

The Figures in the margin indicate full marks.
Candidate are required to give their answers in their own words as far as practicable

Group-A (Very Short Answer Type Question)

1. Answer *any ten* of the following : [1 x 10 = 10]

(I) With dropout technique what we can prevent?

(II) Which loss is used to train a MLP if the outputs are given as {1,2,3...} in this form?

(III) Which algorithm is used for likelihood computation in HMM?

(IV) Weight sharing is a procedure of reducing number of parameters. True or False?

(V) GPU stands for what?

(VI) You have to recognise digit 6. To handle overfit you use data augmentation. Which data augmentation you should not do?

(VII) Consider a trained logistic regression. It's weight vector is W and its test accuracy on a given dataset is A. Assuming no bias, dividing W by 2 what will be test accuracy?

(VIII) We cannot use mean squared error loss function in binary classification problem for ANN.True or False?

(IX) Viterbi Algorithm is used for decoding i.e. to find hidden sequence. True or False?

(X) Stride and number of filters are treated as hyperparameters True or False?

(XI) Is Big Data related to Deep Learning in symbolic way?

(XII) Mini batch gradient descent is advantageous comparing to full-batch gradient descent. True or false?

Group-B (Short Answer Type Question)

Answer *any three* of the following : [5 x 3 = 15]

2. What is Pooling? What is Max Pooling? What is Average pooling? Show two examples [5]

3. You have been asked to classify MNIST digits. Can you suggest a deep neural network architecture. You just need to specify the layers, and their sizes. The MNIST data set has the images having size say 28x28. [5]

4. What is tokenisation in text classification? What is the deep NN architecture for NLP? [5]

5. Say we have a neural network with 3 input neurons, 2 hidden layers each having 8 neurons, and 3 neurons at the output layer. Find the total number of biases. Find the total number of weights. Which loss and activation function for the output layer is best suited for the above-given network? [5]

6. Consider a MLP with 2 hidden layer. Only one output unit and 3 input unit. Show the backpropagation calculation for single step , i.e. from output 2nd hidden layer. [5]

Group-C (Long Answer Type Question)

Answer *any three* of the following : [15 x 3 = 45]

7. (a) Is there any difference between a single layer perceptron and a logistic regression.? If yes point the difference if no, justify [2]

(b) What is delta rule? Explain with expression. [2]

(c) If your logistic regression model sufferers from overfitting what will you do? [1]

(d) Say we have nonlinearly separable datapoints. We want to classify the with linear decision boundary. Is it possible? Justify your answer. [2]

(e) Suppose you have 5 logistic regression models. Can you make a strong classifier with these 5 models? If no say why. If yes justify your answer. [2]

(f) Why F1score is better than accuracy in learning problem? [2]

(g) We have a confusion matrix having all off diagonal elements are zero. What you can infer from this result for your designed classifier? [1]

(h) We have 10 set of non-linearly separable data with no labels are given. You want to classify this data set. What will be your approach? [2]

(i) How you can manage the vanishing gradient problem in deep network? Suggest one network which can help you to prevent the same.

8. (a) Why ReLU activation function leads to sparse activation maps? [2]

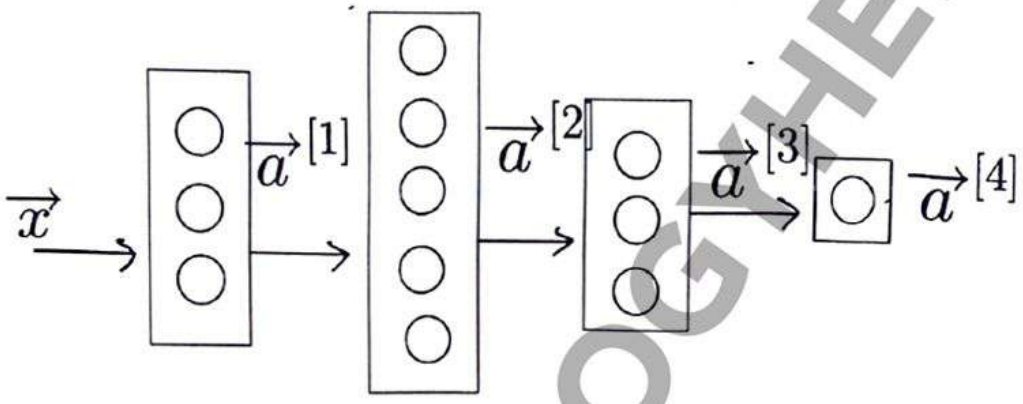
(b) We have a problem set with 4 class classification. Outputs are one hot encoded. Say for the 3rd class we have the softmax output =[0.3,0.02,0.6,0.08]. What will be the cross entropy loss? [3]

(c) A given cost function J(w) = 2w^2-4w+2. What is the weight update rule for gradient descent optimization at step t+1. Consider learning rate = 0.01 [3]

(d) What is the range of tanh activation function? [1]

(e) What is vanishing gradient? Which activation function(s) lead(s) to the same? Explain with example. When is gradient descent algorithm certain to find global minima? [6]

9. (a) Consider the above network. Answer the following
What is the size of W^[1]? [1]



(b) How a^[3] is computed? Show all components of a^[3] [2]

(c) How a^[2] is computed? Show all components of a^[2] [2]

(d) Why W^[1],W^[2],W^[3] are vectors or matrices? [2]

(e) Redraw the network with weights and links in each layer. [3]

(f) What is the size of W^[2]? [1]

(g) What is the size of W^[3]? [1]

(h) Redraw the network considering all activation functions are linear except the output. [3]

10. (a) Explain briefly what will happen if learning rate is equal to 0. [2]

(b) If g(x) represent sigmoid function find it's first derivative. Now show that at what point the gradient of g(x) is maximum. [5]

(c) If J = loss function, y = output, d = desired output, a = nonlinearity, f = network then how the derivative of the loss function with respect to the weights in a deep neural network is computed? [4]

(d) We are training a neural network using a normal gradient descent algorithm. We observe that the change in weights is small in successive iterations. What are the possible causes for the following phenomenon? Possibility 1: learning rate large, possibility 2 : learning rate small, possibility 3: weight change small, possibility 4 : weight change large. Justify your answer. [3]

(e) *Given z = Wx+b and a = g(z) , If z, a has dimension 3x1, and x has dimension nx1. What will be the dimension of b and W?* [1]

11. (a) In named identity recognition system we have to recognise the person's name from a given sentence. The input is given as " The best book of deep learning is written by Goodfellow , Aaron and Bengio" . What will be the corresponding output vector if you want to build a recurrent neural network? [1]

(b) We have a vocabulary of say 10000 words where ate position 2 the word is "Aaron", at position 1229 the word is "Bengio" and at position 2048 the word is "Goodfellow". We want to represent the input as one hot encoding. What will be the following vectors x<10>, x<11> and x<12>? [2]

(c) A standard network with multiple layers cannot handle this type of sequential problems. Why? [1]

(d) How recurrent neural network can overcome all the problems. Show a standard recurrent neural network architecture, where input length and output length are same. [2]

(e) How the above network in Q.(d) will be changed if the network is used for music generation where input is an integer and output is a sequence of data? [2]

(f) Show an architecture of recurrent neural network with weights for machine translation problem. [2]

(g) Show any RNN for same length input and output. Write the forward propagation expressions for any layer. [2]

(h) Draw a computational graph of RNN (input and output length same) showing the loss of individual layer and the backpropagation flow. [2]

(i) What is the advantage of Bidirectional RNN over standard RNN. Justify with an example. [1]

*** END OF PAPER ***