DEPARTMENT OF COMPUTER SCIENCE
BITS Pilani, K.K. Birla Goa Campus

Neural Networks – BITS F312

Comprehensive Exam, **Date: 06 December 2018**, **Weight: 40%**, **Closed Book**

The answers to subparts should not interleave between two different questions.

1. (10 points) The following questions expects precise/to-the-point answers.

    1. You have designed a RBF network. Suppose that you change your RBF network so that you keep only the top-k activations in the hidden layer, and set the remaining activations to 0. List out point-wise why such an approach might provide improved classification accuracy with limited data.
    2. Consider a 1-dimensional time-series with values 2, 1, 3, 4, 7. Perform a convolution with a 1-dimensional filter 1, 0, 1 and no padding.
    3. Show that the derivative of the sigmoid activation function is at most 0.25, irrespective of the value of its argument. At what value of its argument does the sigmoid activation function take on its maximum value?
    4. Mention methods for overcoming the problem of (a) vanishing gradients, (b) exploding gradient. (If you dont know the exact terminology, that is fine; just write how would you overcome?)
    5. Propose an approach for using RBMs for outlier detection.

    (5×2)

2. (12 points) Answer the following questions:

    (a) Consider a social network with a large volume of messages sent between sender-receiver pairs. We are interested only in the messages containing an identifying keyword, referred to as a *hashtag*. Create a real-time model using an RNN, which has the capability to recommend hashtags of interest to each user, together with potential followers of that user who might be interested in receiving messages related to that hashtag. Structural description with diagrams of your RNN model(s) is sufficient.

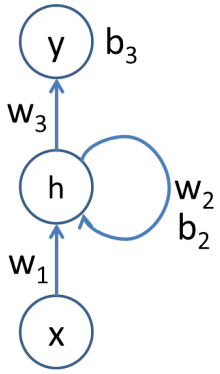
    (b) Consider a problem in which each sentence is treated as a training (or test) instance for classification purpose. This is a typical description of a sentiment analysis system. Assume that there are two different classes of sentences: *Positive*, *Negative*. Create RNN model for this problem: (a) draw the network/block structure, (b) explain the inputs and outputs, (c) describe the activation functions, (d) describe the loss function, (e) derive the weight update equations.

    (c) Recall that an RNN takes in an input vector $\mathbf{x}(t)$ and a state vector $\mathbf{h}(t-1)$ and returns a new state vector $\mathbf{h}(t)$ and an output vector $\mathbf{y}(t)$:

    $$h(t) = f(w_1 x(t) + w_2 h(t-1) + b_2)$$

$$y(t) = g(w_3 h(t) + b_3)$$

Where f and g are activations functions. Refer the following figure.



Given $(x \geq 0)?f(x) = 1 : f(x) = 0$; $g(x) = x$, and $h(0) = 0$; find the values or conditions on weights so that RNN initially outputs 0 and as soon as it receives input 1, it starts producing 1 as output. For example, if input sequence is 00100101, then output sequence would be 00111111.

(d) Answer the following: (i) If the training data set is re-scaled by a particular factor, do the learned weights of either batch normalization or layer normalization change? (ii) What would be your answer if only a small subset of points in the training data set are re-scaled?

(4+4+2+2)

3. (10 points) Answer the following questions:

   (a) Propose a method to extend Radial-basis Function (RBF) networks to unsupervised learning with autoencoders: describe (a) the neural structure diagram, (b) a suitable loss function, (c) derive the weight update equations using suitable optimisation algorithm.

   (b) Discuss how you can modify the RBF autoencoder in the question above to perform semi-supervised classification, when you have a lot of unlabeled data, and a limited amount of labeled data. You may choose to write a formal procedure.

   (5+5)

4. (8 points) You have an energy model of neural networks for unsupervised learning called a Restricted Boltzmann Machine (RBM).

   (a) Assuming Bernoulli features for inputs and Bernoulli code state, derive an expression for the RBM distribution over the space of inputs, denoted as $P(\mathbf{x})$.

   (b) Describe mathematically how can an RBM be interpreted as a standard feed-forward network with one layer of nonlinear processing neurons.

   (4+4)

ANSWER

1. (points) Answers:

   (a) This approach is a form of regularization of the hidden layer, and it will remove the effect of low activations that are noisy anyway. Therefore, the accuracy of the approach might improve when there is limited data.

   (b) Output length= 5 -3 +1 = 3. The sequence is 5, 5, 10.

   (c) The derivative is $f(x)(1-f(x))$. $f(x)$ is between $[0, 1]$. By differentiating, it is easy to show that this function takes on its maximum value at $f(x) = 0.5$ at $x = 0$.

   (d) (a) Vanishing gradient: Using a cell-state (like LSTM cells). (b) Exploding gradient: Gradient Clipping

   (e) It is easy to show that RBMs can be used for dimensionality reduction. Points with large reconstruction error are outliers.

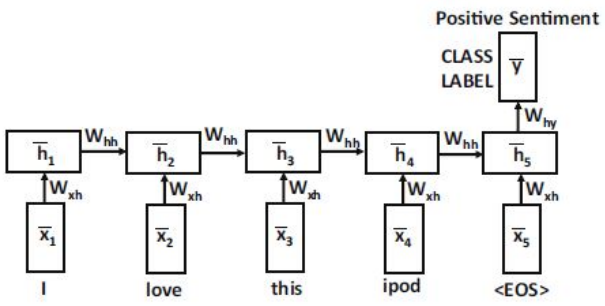
2. (points) Answers:

   (a) I would train two RNNs separately. One architecture has input as the sender/receiver identifier, and the output as the hashtag-identifier and the receiver/sender identifier at the other end. Another uses the sender/receiver and hashtag as input and outputs the receiver/sender. Therefore, for each message we obtain two training points, as we treat them in a symmetric way (asymmetric models are also possible). An RNN is trained incrementally as sender-receiver pairs and hashtags are received.

   The input of the first RNN uses the sender/receiver as input. The output of the first RNN uses two softmax units, one for the hashtag and the other for the receiver/sender identifier.

   The input of the second RNN uses the sender/receiver and hashtag as input. The output of the second RNN outputs a receiver/sender identifier. At any given point, we can use both the models in "test mode" where a potential sender is input to the first RNN in order to obtain receiver identifier and hashtag recommendations. The potential sender and the recommended hastag is then input to the second RNN in order to obtain possible receivers for that hashtag.

   (b) I will use a sentence-level RNN that takes inputs as words. We have a maximum K length sentence. Here is the structural diagram of the RNN:



   Inputs are words. Output is a class. Activation functions are tanh and *sigmoid* in hidden layer and output layer respectively. Backpropagation equations can be referred from the class notes.

(c) $b_3 = 0, w_3 + b_3 = 1 (i.e.\ w_3 = 1)$, $b2 < 0$, $w_1 + w_2 + b_2 \geq 0$, $w_1 + b_2 \geq 0$, $w_2 + b_2 \geq 0$.

(d) (i) Data set rescaling does not affect either normalization method. (ii) However, if only a subset of the data is scaled it will affect batch normalization but not layer normalization. This is because the normalization factors across a batch will change.

3. (points) Answers:

   (a) In this case, we create an output layer with as many units as the number of RBF units. The target values in the output layer are set to the same values as the RBF units. We also have a trained hidden layer between the unsupervised RBF hidden layer and the output layer, which has fewer units than the layers on either side of it. The weights on either side of this layer are tied, as in a conventional autoencoder. These weights are trained with backpropagation (refer class notes).

   (b) The RBF autoencoder from the previous question is first trained with unlabeled data. Then, its decoder is removed and the innermost hidden layer (with reduced representation) is capped with a classification layer (with sigmoid or softmax activation, depending on class label). The weights of this added layer are trained with the labeled data. Furthermore, the RBF autoencoder weights can also be re-tuned. The resulting approach provides a semi-supervised model for classification.

4. (points) Refer class notes for the answers.