Birla Institute of Technology & Science - Pilani, Hyderabad Campus
Second Semester 2020-21
BITS F441: Selected Topics from Computer Science (Deep Learning)
Test 3

Type: Open (Online) **Time: 30 mins** **Max Marks: 24** **Date: 21.11.2020**
All parts of the same question should be answered together.

BITS Id No:

Name:

Code of Honor: I hereby agree to the fact that I will not give or receive any aid during the examination. This includes, but is not limited to, viewing the answering of others, sharing answers with others, and making unauthorized use of internet while taking the exam. If I violate any of these in any manner, I am ready to receive the consequences of punishment awarded for using unfair means in examination.

1. What is the fundamental difference between the AdaGrad and RMSProp algorithm and how do you reason out that this fundamental difference to make RMSProp to perform better in certain scenarios? [3 Marks]

   Sol: The typical gradient accumulation in AdaGrad is changed to exponentially weighted moving averages to discard history from the extreme past so that it can converge rapidly (typically in the case of non-convex function).

2. What are the arguments that you put forth to push the algorithms like AdaGrad / RMSProp /ADAM on top of classical stochastic gradient / SGD with Momentum / SGD with Nesterov momentum. [4 Marks]

   Sol: The cost function is highly sensitive to some directions in parameter space and insensitive to others. If it is believed that the directions of sensitivity are axis aligned, it makes sense to use separate learning rate to each parameter and automatically adapt these learning rates throughout the course of learning. In all cases where the assumption of directions of sensitivity are axis aligned it is worth using the algorithms like AdaGrad / RMSProp /ADAM.

3. Discuss the significance of surrogate loss function. Discuss a scenario where surrogate function is to be used. [3 Marks]

   Sol: If the true loss function is a 0-1 loss function which is typically intractable then a surrogate loss function (that is close to the initial loss function) is proposed and the corresponding optimization problem is solved.

4. As discussed in the class, explain the difficulties in minimizing the following error function:

   $$J^*(\theta) = E_{(x,y)\sim p_{\text{data}}} \left(L(f(x;\theta), y) \right)$$

   where L is the per-example loss function, f(x; Θ) is the predicted output when the input is x and p_{data} is data-generating distribution.

   How is this problem being addressed in practice to provide a reasonable solution? [4 Marks]

Sol: The P_{data} is the data distribution which is not known. Hence the training data is taken as empirical distribution and mean loss of training examples is taken as the objective function to be minimized and the optimization problem is solved.

5. Discuss the scenarios in which an ensembler of k regression models will give optimal results.
   Note: You may assume reasonable assumptions to arrive at the solution of the problem.
   [4 Marks]

   Sol: Suppose each constituent models in the emsembler makes error €$_i$ on each example, with the errors drawn from zero-mean multivariate normal distribution with E[€^2_i] = v and covariance E[€$_i$, €$_l$] = c. If c is 0 then the ensemble will have optimal performance.

6. What is 'Representational Sparsity'? Explain how the same is being achieved by solving an optimization problem. You need to write the optimization problems with all necessary details.
   [3 Marks]

   Sol: Explain the loss function with terms in detail.
   $\tilde{J}(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}) = J(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}) + \alpha\Omega(\boldsymbol{h})$

7. What do you mean by i.i.d. examples whenever it is used in deep learning? Without this assumption, what are the issues that you foresee in formulating and solving the optimization problem for the learning problem? [3 Marks]
   Sol: If examples are drawn independently from an identical distribution then we say that examples are i.i.d. examples. Most of the times the error function is negative likelihood function i.e. - log (p(y1, y2, . . , yN/x1, x2, . . , xN) and this cannot be written as –log(p(y1/x1) p(y2/x2) . . p(yN/xN)) if examples are not i.i.d. examples and hence the loss function cannot be – $SUM_{1\leq n\leq N}$(log(yi/xi)).