DEEP LEARNING – WORKSHEET

5 Q1 to Q8 are MCQs with only one correct answer. Choose the correct option.

1. Which of the following are advantages of batch normalization?
   A) Reduces internal covariant shift.
   B) Regularizes the model and reduces the need for dropout, photometric distortions, local response normalization and other regularization techniques.
   C) allows use of saturating nonlinearities and higher learning rates.
   D) All of the above
   D) All of the above

2. Which of the following is not a problem with sigmoid activation function?
   A) Sigmoids do not saturate and hence have faster convergence
   B) Sigmoids have slow convergence.
   C) Sigmoids saturate and kill gradients.
   D) Sigmoids are not zero centered; gradient updates go too far in different directions, making optimization more difficult.
   C) Sigmoids saturate and kill gradients.

3. Which of the following is not an activation function?
   A) Swish B) Maxout C) SoftPlus D) None of the above
   B) Maxout

4. The tanh activation usually works better than sigmoid activation function for hidden units because the mean of its output is closer to zero, and so it centers the data better for the next layer. True/False?
   A) True B) False
   A) True

5. In which of the weights initialisation techniques, does the variance remains same with each passing layer?
   A) Bias initialisation B) Xavier Initialisation C) He Normal Initialisation D) None of these
   B) Xavier Initialisation

6. Which of the following is main weakness of AdaGrad?
   A) learning rate shrinks and becomes infinitesimally small
   B) learning rate doesn't shrink beyond a point
   C) change in learning rate is not adaptive
   D) AdaGrad adapts updates to each individual parameter
   A) learning rate shrinks and becomes infinitesimally small

7. In order to achieve right convergence faster, which of the following criteria is most suitable?
   A) momentum and learning rate both must be high
   B) momentum must be high and learning rate must be low
   C) momentum and learning rate both must be low
   D) momentum must be low and learning rate must be high
   D) momentum must be low and learning rate must be high

8. When is an error landscape is said to be poor(ill) conditioned?
   A) when it has many local minima
   B) when it has many local maxima
   C) when it has many saddle points and flat areas
   D) None of these
   A) when it has many local minima

Q9 and Q10 are MCQs with one or more correct answers. Choose all the correct options.

9. Which of the following Gradient Descent algorithms are adaptive?
   A) ADAM B) SGD C) NADAM D) RMS Prop.
   A) ADAM
   D) RMS Prop

10. When should an optimization function (gradient descent algorithm) stop training:
    A) when it reaches local minimum
    B) when it reaches saddle point
    C) when it reaches global minimum
    D) when it reaches a local minima which is similar to global minima (i.e. which has very less error distance with global minima)
    C) when it reaches global minimum
    D) when it reaches a local minima which is similar to global minima (i.e. which has very less error distance with global minima)

Q11 to Q15 are subjective answer type question. Answer them briefly.

11. What are convex, non-convex optimization?

    Convex optimization there can be only one optimal solution, which is globally optimal or we might prove that there is no feasible solution to the problem. Non-convex optimization may have multiple locally optimal points and it can take a lot of time to identify whether the problem has no solution or if the solution is global. Hence, the efficiency in time of the convex optimization problem is much better.

12. What do you mean by saddle point? Answer briefly.

    A point at which a function of two varibales has partial derivatives equal to zero but at which the function has neither a maximum nor a minimum value. When we optimize neural networks or any dimensional function. For most of the trajectory we optimize, the critical points(the points where the derivative is zero or close to zero) are saddle points. Saddle points, unlike local minima, are easily escapable.

13. What is the main difference between classical momentum and Nesterov momentum? Explain briefly.

    The distinction between momentum method and Nesterov momentum gradient updates was both methods are distinct ony when the learning rate is reasonably large. When the learning rate is relatively large, Nesterov momentum radient allows larger decay rate than momentum method, while preventing oscillations. The theorem also shows that both momentum method and Nesterov momentum gradient become equivalent when rate is small.

14. What is Pre initialisation of weights? Explain briefly.

15. What is internal covariance shift in Neural Networks?

    We define Internal Covariet shift as change in the distribution of network activations due to change in network parameters during training. In neural network, the output of the first layer feeds into the second layer, the output of the sencond layer feeds into third, and so on. When the parameters of a layer change, so does the distribution of inputs to subsequent layers. These shifts in input distributions can be problematic for neura; networks, especially deep neural networks that could have a large number of layers. Batch normalization is a method intened to imtigate internal covariate shift for neural networks.