Indian Institute of Technology Jodhpur
Department of Computer Science and Engineering
Machine Learning - 2 (Code: CSL7050)

January 09, 2022

Instructions:
1. Read the questions carefully.
2. This is a closed-book exam, and plagiarism of any form will be severely penalized.
3. Write your answers on white paper and submit scanned copy (one single PDF file) in google-classroom. Submissions received in any other format or through any other medium will not be considered. Name the file as <YourRollNumber.pdf>
4. There will be a penalty for late submissions.
5. No queries will be answered during the exam. If you find anything unclear/incorrect in any question, make a reasonable assumption and proceed.

Time: 2 hours (including 10 minutes to upload) I-grade Exam Maximum Marks: 60

1. Suppose our prediction rule is given by $\hat{y} = \max(\sigma(wx + b), 0.2)$, where w is a learnable parameter, x is an input and b is a constant bias (all of them are scalars). Assuming that we have only one sample in our training set, calculate/derive the first-order derivative of the MSE loss w.r.t. 'w'. [6]

2. Compute the number of parameters in a convolution layer with an input volume of size $32 \times 32 \times 3$, with twenty 5×5 filters, stride $= 1$, and padding $= 2$. Include a bias parameter for the calculation. [4]

3. Write the equation of the ReLU activation function, calculate its derivative, and plot both of these. [5]

4. Early stopping is a popular (and perhaps the simplest) way to avoid overfitting, where the training procedure is stopped when there is an increase in the error on the validation set. Briefly explain why this method works. [5]

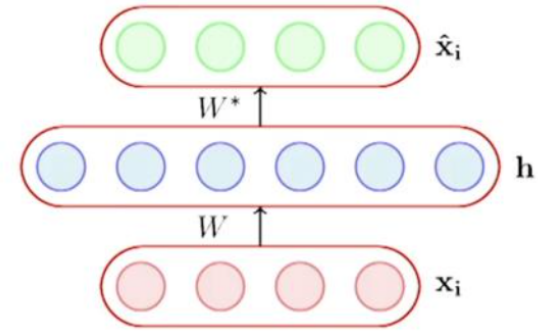
5. How can we use hierarchical clustering for network compression? Calculate the compression ratio achieved using this approach. [4]

6. What does proxy data refer to in context of neural architecture search? [2]

7. What is the difference between Continuous Bag of Words Model (CBOW) and Skip-gram models? Give a real world example where one would be advantageous over the other, justify your answer. [4]

8. Consider the autoencoder shown below where the bottleneck layer has higher dimension as compared to the input. Explain the problem associated with this architecture and suggest the solutions. [4]

Please go on to the next page...



9. Consider a deep neural network classifier developed for COVID vs. non-COVID chest x-ray classification. The network is trained on images of European population and performs well on the same. However, the network performs poorly on Indian population, what could be the possible reasons of the poor performance? Propose a solution to the problem. Are there any other ways of solving the problem, justify your choice? [6]

10. Consider a dataset $\{\mathbf{x}_i\}_{i=1}^n$ containing samples drawn from an unknown data distribution $p(\mathbf{x})$. Explain how we can use GAN framework for implicitly learning this distribution $p(\mathbf{x})$. [6]

11. Consider a dataset $\{\mathbf{x}_i\}_{i=1}^n$ containing samples drawn from an unknown data distribution $p(\mathbf{x})$, we want to learn a distribution $p_{\boldsymbol{\theta}}(\mathbf{x})$ that is as close as possible to the true distribution $p(\mathbf{x})$. We have to minimize the KL divergence between the distributions $p(\mathbf{x})$ and $p_{\boldsymbol{\theta}}(\mathbf{x})$ with respect to $\boldsymbol{\theta}$ to find the optimal values of the parameters $\boldsymbol{\theta}$. Show that minimizing the KL divergence is equivalent to maximize the log likelihood $\log p_{\boldsymbol{\theta}}(\mathbf{x})$ with the parameters $\boldsymbol{\theta}$. [6]

12. Given a set of points $\{\mathbf{v}_i\}_{i=1}^N$, we model $p_{\boldsymbol{\theta}}(\mathbf{v})$ that is as close as possible to true distribution $p_{\text{data}}(\mathbf{v})$. Here, all vectors $\mathbf{v}_i \in \{0,1\}^{m\times 1}$. To solve this generative modeling problem, we employed the restricted Boltzmann machine (RBM) framework. Here, we model the joint distribution of the visible variables $\mathbf{v}$ and hidden variables $\mathbf{h} \in \{0,1\}^{n\times 1}$ using the energy based model, i.e., $p(\mathbf{v},\mathbf{h}) = \frac{e^{-E(\mathbf{v},\mathbf{h})}}{Z}$. Here, the functions $E(\mathbf{v},h)$ denotes the energy function and defined as $E(\mathbf{v},\mathbf{h}) = -\mathbf{v}^\top\mathbf{W}\mathbf{h} - \mathbf{b}^\top\mathbf{v} - \mathbf{c}^\top\mathbf{h}$. Here, $\mathbf{W}$, $\mathbf{b}$, and $\mathbf{c}$ are the learnable parameters of the RBM. We maximize the log-likelihood function $\ell(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\log p_{\boldsymbol{\theta}}(\mathbf{v}_i)$ with respect to the parameters $\boldsymbol{\theta} = \{\mathbf{W},\mathbf{b},\mathbf{c}\}$. Show that [8]

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \mathbf{w}_{ij}} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{h}\sim p(\mathbf{h}|\mathbf{v}_i)}[v_i h_j] - \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{(\mathbf{v},\mathbf{h})\sim p(\mathbf{v},\mathbf{h})}[v_i h_j].$$

Now, assume that we already know the optimal parameters $\mathbf{b}$ and $\mathbf{c}$. Write the down the contrastive-divergence algorithm to find the optimal $\mathbf{W}$. Here, w_{ij} is the $(i,j)^{\text{th}}$ entry of the matrix $\mathbf{W}$.

End of exam