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

January 08, 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) E-grade Exam Maximum Marks: 60

1. Briefly discuss four benefits of RNN/LSTM. [2]
2. 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]
3. (a) List two factors that consume most of the training time while learning a deep neural network using the stochastic gradient algorithm. [6]
   (b) How can we address each of these two (mention one way for each factor respectively)?
4. Which of the following is/are true in the context of "regularization"? Justify your answer for each. [6]
   1. It helps in expressing preference to a simpler model.
   2. It provides different perspectives on how to explain the training data.
   3. It helps in decoding prior knowledge.
5. What is few shot domain adaptation? How is it different from semi-supervised learning? [5]
6. Explain the efficient forward architecture search (Petridish) approach. [5]
7. What are the advantages and disadvantages of performing neural network compression using knowledge distillation. [4]
8. 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 [6]

possible reasons of the poor performance? Propose a solution to the problem. Are there any other ways of solving the problem, justify your choice?

9. Consider a coin with head probability equal to r. Let X be a random variable such that $X = -2$, if head appears and $X = 2$, if tail appears. Consider the probability mass function (PMF) p_X of X when $r = 0.1$ and the PMF q_X of X when $r = 0.9$. Find the KL divergence between p_X and q_X. [6]

10. Given a dataset $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ 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 $\boldsymbol{\theta}$. To maximize this log-likelihood function, we used VAE as one of the approach where we use latent variables $\mathbf{z}$ that semantically represent the input dataset. Then, we learn the conditional distribution $p(\mathbf{x}|\mathbf{z})$ and use it to sample new points from the distribution $p(\mathbf{x})$. Now, consider a distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ parametrized by the parameters $\boldsymbol{\phi}$. Then, show that the log-likelihood $\log p_{\boldsymbol{\theta}}(\mathbf{x})$ is equal to [6]

$$-D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})|p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\right) + D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})|p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})\right).$$

11. 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 $p_{\mathrm{data}}(\mathbf{v})$. 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]

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

Now, assume that we already know the optimal parameters $\mathbf{W}$ and $\mathbf{c}$. Write the down the contrastive-divergence algorithm to find the optimal $\mathbf{b}$.

End of exam