Machine Learning for Communication Systems
MALCOM Spring 2021
EURECOM

Final Examination (2 hours)

- This exam is open books and documents. Use of internet resources, web search, and any sort of communication is strictly forbidden during the exam.
- In the short questions/ exercises, partial credit will be given for good explanations even if you have not provided the complete correct answer. Therefore, please explain your idea, reasoning, derivations, calculations, etc., even if you are unsure of your answers.

Examination Honor Code: This online examination, with no physical presence in a classroom, has an Honor Code. You may not consult or collaborate with anyone about the questions. Such collaboration is a violation of the Honor Code. Please sign or type your name below.

I attest on my honor that I have not given or received any unauthorized assistance on this examination.

Signature: ____________________________

The exam contains **8** pages including this cover page.

Part 1 – Multiple Choice Questions (20 points)

For each of the following questions, circle/choose the number of your choice. There may be more than one correct choice (usually explicitly mentioned). No explanation is required. There is no penalty for a wrong answer.
You can send this PDF file marking or highlighting your answer or you can type your answers in the Answer Sheet provided.

Question 1: ML algorithms will have great impact in maximizing the performance of point-to-point wireless communication links.

(i) True
(ii) False

Question 2: Which of the following techniques can be used to reduce model overfitting? Choose all that apply.

(i) Data augmentation
(ii) Batch normalization
(iii) Using AdaGrad optimizer instead of simple SGD
(iv) Dropout

Question 3: Which of the following would you consider to be valid activation functions to train a neural networks?

(i) $f(x) = \min(x, 0.1x)$
(ii) $f(x) = 10x + 1$
(iii) $f(x) = -\min(x, 5)$
(iv) $f(x) = \begin{cases} \max(x, 0.1x) & x \geq 0 \\ \min(x, 0.1x) & x < 0 \end{cases}$

Question 4: Which of the following activation functions can lead to vanishing gradients?

(i) ReLU
(ii) Leaky ReLU
(iii) Tanh
(iv) None of the above

Question 5: We train a neural network and we obtain a training accuracy of 100% and a test accuracy of 40%. Which of the following methods is commonly used to reduce this large gap?

(i) Sigmoid activation function
(ii) Variational autoencoder
(iii) Dropout
(iv) RMSprop optimizer

Question 6: You are thinking of using gradient descent as your optimization method for training your neural network. Which of the following is true? Choose all that apply.

(i) It is possible for Batch Gradient Descent to converge faster than Stochastic Gradient Descent
(ii) It is possible for Mini Batch Gradient Descent to converge faster than Batch Gradient Descent
(iii) It is possible for Mini Batch Gradient Descent to converge faster than Stochastic Gradient Descent
(iv) It is possible for Stochastic Batch Gradient Descent to converge faster than Batch Gradient Descent

Note: faster here is measured in terms of wall clock time

Question 7: SGD in typical ML problems requires fewer parameter updates to converge than full gradient descent.

(i) True
(ii) False

Question 8: Which of the below distributed system architectures does not provide model consistency at each training epoch/iteration. (Choose all that apply)

(i) Parameter server (PS) with synchronous SGD
(ii) PS with asynchronous SGD
(iii) All reduce
(iv) Gossip/decentralized
(v) None of the above

Question 9: In distributed ML system architectures, synchronization is the only thing needed to obtain very high model consistency.

(i) True
(ii) False

Question 10: If a convex function has a minimum, then the minimizer is unique.
(i) True
(ii) False

Question 11: The convergence rate of SGD in the convex case is better than that of SGD in the strongly convex case.
(i) True
(ii) False

Question 12: The convergence rate of batch gradient descent in the convex case is better than that of SGD in the strongly convex case.
(i) True
(ii) False

Question 13: Consider the function $f(x) = -x^2$. For the subgradients of $f(x)$ at $x = 0$, choose which of the following statements are true.
(i) A subgradient exists and is unique.
(ii) A subgradient exists but is not unique.
(iii) A subgradient does not exist as $f(x)$ is differentiable at $x = 0$.
(iv) A subgradient does not exist even though $f(x)$ is differentiable at $x = 0$.

Question 14: Which of the following function(s) have a unique minimizer?
(i) $f(x) = x^2,\ x \in [-3, 2]$
(ii) $f(x) = \log x\ ,\ x \in (0,\ 10]$
(iii) $f(x) = \sin(x)\ ,\ x \in [-10, 10]$
(iv) $f(x) = e^{3x} + x^4 - 3x,\ x \in [-10, 10]$

Question 15: The maximum likelihood model parameters (α) can be learned using linear regression for the model: $y_i = \log(x_1^{\alpha_1} e^{\alpha_2}) \quad + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is i.i.d. noise and $x_1 > 0$.
(i) True
(ii) False

Question 16: There is a webpage (*www.thiseurecomstudentdoesnotexist.fr*) which generates pictures of Eurecom students who does not exist. Every time we open the webpage, we see a fake picture of a student that was sampled from the distribution learned by a GAN. Which part of the GAN is deployed in the webpage server?

(i) The discriminator
(ii) The generator
(iii) The discriminator and the generator
(iv) The last layer of the discriminator

Question 17: The loss function of GAN quantifies the similarity between the generative data distribution and the real sample distribution by Kullback-Leibler (KL) divergence when the discriminator is optimal.

(i) True
(ii) False

Question 18: We have a reinforcement learning setting with dynamics $p(s_6|s_6, a_1) = 0.5$, $p(s_7|s_6, a_1) = 0.5$, and reward r as follows: for all actions, $+1$ in state s_1, $+10$ in state s_7, 0 otherwise. We assume $V_k = [1 \ \ 0 \ \ 0 \ \ 0 \ \ 0 \ \ 0 \ \ 10]$, $k = 1$, $\gamma = 0.5$. Let $\pi(s) = a_1$, $\forall s$. What is the value of $V_{k+1}(s_6)$?

(i) 0
(ii) 2.5
(iii) 10
(iv) None of the above

Question 19: For a communication-efficient implementation of federated learning, each worker (device) can quantize the gradients sent back to the parameter server.

(i) True
(ii) False

Question 20: Using a Generative Adversarial Network (GAN), we aim at explicitly estimating the density of data distribution.

(i) True
(ii) False

Part 2 – Short Answer Questions & Exercises (50 points)

For the questions in this section, please be concise and provide 2-3 sentences (or more if needed, e.g. question 19) in your responses.

Question 1: We would like to implement a deep wireless transmitter by training a fully-connected neural network with 8 hidden layers, each with 15 hidden units. The input is a 20-dimensional vector and the output is a scalar. What is the total number of trainable parameters in your network?

Question 2: Characterize/state the difference between classification and regression in one sentence.

Question 3: ML is used in communication systems only to reduce complexity. Comment the statement.

Question 4: The best way to validate your supervised ML model is to split your data into two parts, a training dataset and a test dataset. True or False? Why?

Question 5: ML is powerful! "*There always exists an algorithm that can learn any globally optimal target function within an arbitrary small error using a finite number of samples.*" Comment the statement.

Question 6: AI University needs to classify student applications into good or bad categories. It also needs to detect student applicants who lie in their applications using density estimation to detect outliers. You are recruited to help them meet these requirements. Do you recommend using a discriminative or generative classifier? Why?

Question 7: The same above University offered you an internship to automate their admission process using neural networks. Your model automatically classifies CVs in order to send acceptance/rejection letters to students. Which of the following metrics is more important for your model? Explain.

$$\text{Metric 1} = \frac{\text{True positives}}{\text{Total positive samples}}$$

$$\text{Metric 2} = \frac{\text{True positives}}{\text{Total predicted positive samples}}$$

Question 8: During neural network training, the validation error happens to be significantly lower than the training error. Why this could happen?

Question 9: High-tech company Backrub proposes the following non-linear function for neural networks.

$$f(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Are you going to use this non-linear function into your new ML project? Why or why not?

Question 10: A binary classification problem could be solved with the two approaches described below:

Approach 1: Simple Logistic Regression (one neuron)
Your output will be $\hat{y} = \sigma(W_1 x + b_1)$
Classify as 0 if $\hat{y} \leq 0.5$ and 1 otherwise.

Approach 2: simple softmax regression (two neurons)
Your output will be $\hat{y} = softmax(W_2 x + b_2) = [\hat{y}_1, \hat{y}_2]^T$
Classify as 0 if $\hat{y}_1 \geq \hat{y}_2$ and 1 otherwise.

Approach 2 involves twice as many parameters as approach 1.
Can approach 2 learn more complex models than approach 1?
If yes, give the parameters (W_2, b_2) of a function that can be modeled by approach 2 but not by approach 1. If no, show that (W_2, b_2) can always be written in terms of (W_1, b_1).

Question 11: In a neural network, the non-linear function is
$f(x, y, z, u, v, w) = 3xyzuvw + x^2 y^2 w^2 - 7xz^5 + 3yvw^4$.
What is the value of $\left[\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} + \frac{\partial f}{\partial z} + \frac{\partial f}{\partial u} + \frac{\partial f}{\partial v} + \frac{\partial f}{\partial w}\right]\Bigg|_{x=y=z=u=v=w=1}$?

Question 12: In federated learning, it is beneficial to perform shuffling when delivering datasets to workers to perform training. True or False?

Question 13: Gradient diversity is an important measure of dissimilarity. Can this metric/concept be used in federate learning? If so, how you will use it? Would minimization or maximization of gradient diversity be better? Comment.

Question 14: We are training a neural network for classification. We first train the network for 100 samples and although training converges, the training loss is very high. We then train the network on 10.000 examples. Is this approach to fixing the problem correct? If yes, explain the most likely results of training with 10.000 examples. If not, give a solution to this problem.

Question 15: For three distributions P, Q, and R show the following equality

$$D_{\mathrm{KL}}(P,Q) = \mathbb{E}_{x\sim P}\left\{\log\frac{R(x)}{Q(x)}\right\} + D_{\mathrm{KL}}(P,R)$$

Question 16: You want to give a solution to the straggler problem. For that, you want to achieve the same expected time scaling performance (neglecting the incurance of loss) of a scheme waiting the 10 fastest codes with an MDS(Y,n) code. How will you choose Y? How many redudant nodes you will need if you use 100 nodes containing information? (*MDS: maximum distance separable*)

Question 17: In straggler mitigation, one wants to have the same order performance in terms of expected time between replication r times and waiting for $(1-\varepsilon)n$ nodes? How much r should be for $n = 20$ and waiting for 50% of the nodes?
(Assume: $\log_2 10 \approx 3$).

Question 18: *The Bet.* The class project is a regression task with square loss function. Your best friend uses linear regression and least squares. You are using a neural network with 10 layers and activation functions $f(x) = 2x$. You put a bet: whoever has substantially better scores will win a meal at an expensive restaurant. Who will pay and why?

Question 19: We assume an arbitrary discrete set $\mathcal{H}$ of classifiers and $P(h)$ and $Q(h)$ denote probability masses at h.

a) Calculate the Kullback-Leibler (KL) divergence between P and Q, when P,Q are two Bernoulli distributions with probability of success p and probability of success q, respectively.
b) Under which condition, the KL divergence is symmetric?
c) If you have to guess the Jensen-Shannon divergence (JSD) between P and Q, will you choose that JSD = 0.5 or JSD = 2? (*Note: base 2 logarithms are used*)
d) If $W = (P+Q)/2$ is approximated by a Bernoulli distribution with probability of success w, find values of w that minimizes and maximizes the JSD when $p = q = 0.5$.

Note: base 2 logarithms are used

Note: Each question in Part 2 counts for 2 points except Question 19, which counts for 14 points.