Spring 2022 Final EN.520.650

Total points 60

Q1. (10 points) In this MDP, available actions are left, right, and stay. Stay always results in the agent staying in its current square. Left and right are successful in moving in the intended direction half of the time. The other half of the time, the agent stays in its current square. An agent cannot try to move left at the leftmost square and cannot try to move right on the rightmost square. Staying still on a square gives a reward equivalent to the number on that square (see figure below) and all other transitions give zero reward (meaning any transitions in which the agent moves to a different square give zero reward).

4	0	0	36

(i) [2 pts] $V_0(s)$ is 0 for all s. Perform one step of value iteration with $\gamma = 0.5$ and write $V_1(s)$ in each corresponding square.

(ii) [2 pts] Perform another step of value iteration with $\gamma = 0.5$ and write $V_2(s)$ in each corresponding square.

(iii) [6 pts] Using $\gamma = 0.5$ again, write $V^*(s)$ in each corresponding square (Hint: think about states where the optimal action is obvious and work from there)

Q2. (10 points)

(a) (2 points.) Write down the expression for (forward) inference using a deep feedforward fully connected neural network. (That is, if I have an already trained neural network, how do I use it to get a prediction given a training example x?)

(b) (2 points) What is the computational cost of running inference on a deep neural network? Suppose that all the nonlinearities operate element-wise using the ReLU function ReLU(a) = max(a, 0) and the sizes of the layers are d (where $x \in R^d$), $d_1, d_2, \ldots, d_{L-1}$. How many numerical operations would computing this forward pass require? Which operation dominates: the matrix multiplies or the nonlinearities?

(c) (4 points) Now write down the expression/algorithm for backpropagation on this same network, using ReLU activation functions. (That is, how do I compute the gradient of this neural network with respect to the weights for a training example x?)

(d) (2 points) What is the computational cost of running backpropagation on a deep neural network? That is, how many numerical operations would backpropagation require? Which operation dominates: the matrix multiplies or the nonlinearities?

Q3. (15 points)

The questions in this section can be answered in 2-4 sentences. Please be concise in your responses.

(a) (2 points) The gradient estimated during a step of mini-batch gradient descent has on average a lower bias when the data is i.i.d. (independent and identically distributed). True or False? Explain why.

(b) (2 points) You have two data sets of similar size for a binary classification task. However, one contains almost entirely positive examples, and the other contains only negative examples. You would

like to use both sets to train your model. Describe a scenario in which combining these two data sets could lead to a failure of the model to learn.

(c) (4 points) Ann recommends the use of convolutional neural networks instead of fully-connected networks for image recognition tasks since convolutions can capture the spatial relationship between nearby image pixels. Bill points out that fully-connected layers can capture spatial information since each neuron is connected to all of the neurons in the previous layer. Both are correct but describe two reasons we should prefer Ann's approach to Bill's.

(d) (2 points) You're solving a binary classification task. The final two layers in your network are a ReLU activation followed by a sigmoid activation. What will happen?

(e) (2 points) You are searching the best learning rate for your model. You decide to test the following values between 0.01 and 1:
– learning rate = 0.01
– learning rate = 0.16
– learning rate = 0.21
– learning rate = 0.84
– learning rate = 0.94
Is that a good method? Explain why.

(f) (3 points) (i) (1 point) Describe one advantage of using mini-batch gradient descent instead of full-batch gradient descent. (ii) (1 point) Describe one advantage of using mini-batch gradient descent instead of stochastic gradient descent with batch size 1. (iii) (1 point) Describe one advantage of using Adam optimizer instead of vanilla gradient descent.

Q4. [5 points]

Describe the mode collapse problem in GANs and suggest a solution.

Q5. (10 points)

(a) (2 points) Which of the following statements are true regarding adversarial attacks? (Circle all that apply.)

(i) If you generate an adversarial example to fool a cat classifier A, there's a chance it will fool another cat classifier B.
(ii) The Fast Gradient Sign Method is an iterative method that can generate adversarial examples.
(iii) Using dropout is an effective defense against adversarial attacks.
(iv) You can create an adversarial attack against a neural network that has been encrypted on a device, where you can access neither its architecture nor its parameters.

(b) (2 points) Recall the Fast Gradient Sign Method for generating adversarial examples: $x^* = x + \varepsilon \cdot \text{sign}(\partial J / \partial x)$. Given $x = [1\ 2\ 3]^T$, $\partial J/\partial x = [0.5\ -0.5\ 1]^T$, and $\varepsilon = 0.01$, what would the resulting adversarial example be? Show your work.

© (2 points) The magnitude of ε needed to create the adversarial example increases with the dimension of x. Do you agree with this statement? Explain your reasoning (1-2 sentences).

(c) (2 points) Given the two options of (A) saturating cost and (B) non-saturating cost, which cost function would you choose to train a GAN? Explain your reasoning. (1-2 sentences).

(d) (2 points) You are training a standard GAN, and at the end of the first epoch you take note of the values of the generator and discriminator losses. At the end of epoch 100, the values of the loss functions are approximately the same as they were at the end of the first epoch. Why are the quality of generated images at epoch 1 and epoch 100 not necessarily similar? (1-2 sentences).

Q6 (10 points)

(a) (2 points) Describe an example of perceived AI bias and suggest a solution.

(b) (1 points) Describe an example of interpretable AI before deep learning arrived.

© (2 points) Prove that GRAD-CAM is a generalization of CAM