Fall 2023 CS 760 Final Exam Practice Questions

Instructions: Answer the following questions. Space is provided below each question for your solution and should be sufficient for responses. Read all the questions first. Even if you cannot obtain a final answer, make sure to write your setup and explain how you would obtain the answer as we will consider partial credit. Some problems are harder than others so if you get stuck on one then try coming back to it after completing others.

Good luck!

Wisc ID: .....

Name: .....

You get 2 points for writing your name and Wisc ID correctly. (2 pts)

1 General Machine Learning Concepts

State if the following statements are true or false.

(a) Clustering, generative modeling, and dimensionality reduction are all unsupervised learning techniques. [True/False]

True **All of these are unsupervised learning techniques.**

(b) If you see high training error and high testing error on a supervised learning problem then you should try learning with a more expressive hypothesis class.[True/False]

True. High training error and high test error indicates that your chosen hypothesis class cannot represent the true underlying mapping from inputs to outputs.

(c) Increasing the learning rate for q-learning will always result in faster convergence and better final performance [True/False].

False, q-learning with a high learning rate can lead to slow learning because each update is only a noisy approximation of the desired update.

(d) In a classification problem with imbalanced classes, accuracy is a reliable performance metric for evaluating the model's effectiveness. [True / False]

False, accuracy can be misleading because simply guessing the most common class can achieve accuracy better than random guessing.

(e) Increased model complexity will always lead to lower testing or generalization error [True / False].

False, could lead to overfitting and hence worse testing error.

2 Neural Networks and Linear Regression

1. Which is true about recurrent neural networks (RNNs)?

   (a) They are primarily used to process image data.

   (b) By having a hidden state, they are insensitive to exactly where a certain pattern appears in an input image (i.e., they are translation invariant).

   (c) The hidden state gives the RNN the ability to remember older inputs.

   (d) RNNs are less likely to experience the exploding gradient problem compared to feedforward multi-layer perceptrons.

   (e) None of the above.

   (c), This is the main use of RNNs.

2. An RNN has m parameters (i.e., weights and biases) and is trained on input sequences of length k. At test time it is evaluated on input sequences of length $2k$. At test time, how many parameters does the RNN have?

   m, changing sequence length does not change the number of parameters required.

3. You are a data scientist preparing to use linear regression on a supervised regression problem with > 100 input features. Your boss tells you that he only wants the learned θ_* to depend on a small subset of these features. Describe how you would adjust training to satisfy this request. Use $L1$ regularization or Lasso which will bias linear regression towards a sparse solution (i.e., many components of θ_* are zero).

3 Clustering

1. You receive a series of data points and you want to find a way to partition the points into disjoint clusters. Since you're an adept machine learning student you recognize that you have a clustering problem. Do you use K-Means or hierarchical clustering? K-Means which is a form of partitional clustering.

2. **Hierarchical Agglomerative Clustering** You are given the following 1D data points: $1, 2, 4, -1, 5, 7$ and will run hierarchical agglomerative clustering with Euclidean distance as a similarity metric.

   (a) Draw the tree obtained by HAC using single-linkage.

   (b) Draw the tree obtained by HAC using complete-linkage.

4 Principal Component Analysis

You are given a dataset of m d-dimensional feature vectors which we will represent with the matrix $X \in \mathbf{R}^{n \times d}$, where each row of X corresponds to one of the m features vectors and each column represents one of the d dimensions. You may assume that the mean of each dimension across the m feature vectors is 0. Let $v_1, v_2, ..., v_d$ represent the eigenvectors of $X^\top X$ sorted in order of decreasing eigenvalue. Let $\lambda_1, \lambda_2, ..., \lambda_d$ be the corresponding eigenvalues for each v_i.

1. Calculate the percentage of variance explained by the first k eigenvectors where $k \leq d$. $\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{d} \lambda_j}$

2. We would like to represent a feature vector, $x' \in \mathbf{R}^d$, as a linear combination of 2 vectors with minimal reconstruction error. What 2 vectors would you select to accomplish this goal? v_1, v_2.

3. PCA aims to find a new representation of a set of feature vectors by projecting data onto a set of basis vectors that minimize variation. [True / False] False, want to maximize variance and minimize reconstruction error.

4. PCA can be performed using an eigendecomposition on the sample covariance matrix True, sample covariance is proportional to $X^\top X$ which we perform eigendecomposition on for PCA. Normalizing $X^\top X$ to make it the sample covariance matrix will not change principal components.

5 Graphical Models

You are given the following graph which gives the probabilistic dependencies between a set of variables {A, B, C, D, E}. The structure of the network is as follows: $A \leftarrow B \rightarrow C \leftarrow D \leftarrow E$.

1. For each of the following, say whether conditional independence is guaranteed to hold or is not guaranteed to hold.

   (a) $A \perp\!\!\!\perp C|B$
   (b) $B \perp\!\!\!\perp D|C$
   (c) $A \perp\!\!\!\perp E|C$
   (d) $A \perp\!\!\!\perp B|C$

   (a) guaranteed, (b) not guaranteed, (c) not guaranteed (d) not guaranteed

2. Suppose all variables are binary. How many parameters are needed for all conditional probability tables in this graphical model?

   10 (partial credit for 20 in which a parameters is given for probability of complement outcome.)

3. Assuming you know the parameters of each conditional probability table, describe a procedure for generating a new sample from this graphical model.

   First, sample the value of B and E, then sample D conditioned on the value of E, then sample A conditioned on B and C conditioned on the value of B and D.

4. If all variables are binary, write an expression that gives the probability of $E = 1$ given that $B = 1$ and $D = 1$. This expression *must* be only in terms of probabilities that are found in a CPT for some node of the graphical model.

   $\Pr(E = 1|B = 1, D = 1) = \frac{\Pr(E=1,B=1,D=1)}{\Pr(B=1,D=1)}$.

   We can evaluate this probability if we compute $\Pr(E = 1, B = 1, D = 1)$ and $\Pr(E = 0, B = 1, D = 1)$.

   $\Pr(E = 1, B = 1, D = 1) = \Pr(E = 1)\Pr(B = 1)\Pr(D = 1|E = 1)\sum_{c\in\{0,1\}}\Pr(C = c|D = 1, B = 1)\sum_a \Pr(A = a|B = 1) = \Pr(E = 1)\Pr(B = 1)\Pr(D = 1|E = 1)$.

   Similarly, $\Pr(E = 0, B = 1, D = 1) = \Pr(E = 0)\Pr(B = 1)\Pr(D = 1|E = 0)$.

   Then plug in to $\frac{\Pr(E=1,B=1,D=1)}{\Pr(E=1,B=1,D=1)+\Pr(E=0,B=1,D=1)}$ to obtain solution. Note that $E \perp\!\!\!\perp B|D$ and so we can effectively ignore all other variables since we know $D = 1$. Using this fact, we obtain:

$$\Pr(E = 1|B = 1, D = 1) = \frac{\Pr(E = 1)\Pr(D = 1|E = 1)}{\Pr(E = 0)\Pr(D = 1|E = 0) + \Pr(E = 1)\Pr(D = 1|E = 1)}$$

6 Learning Theory

Consider the following learning theory result that was discussed in class. Specifically, we have that:

$$R(h_S) \leq \frac{1}{m}(\log|H| + \log\frac{1}{\delta}),$$

where $R(h_S)$ is the generalization error of the hypothesis returned by a learning algorithm, m is the number of training examples, $|H|$ is the size of our hypothesis class, and $1 - \delta$ is the probability that we want the bound to hold with. What does this result imply about the generalization error in terms of m, δ, and $|H|$?

Generalization error increases as δ decreases (i.e., we want to be more sure that the bound holds) and as $|H|$ increases. Generalization error decreases as we obtain more data.

7 Reinforcement Learning

1. In RL, the action-value function gives the expected discounted sum of rewards following action a regardless of the current state. [T/F]

   False, the action-value function gives the expected discounted sum of rewards following action a in a particular state.

2. In RL, the state-value function gives the expected discounted sum of rewards following state s when following a certain policy. [T/F] True, by definition.

3. You want to apply reinforcement learning to enable a robot to learn to move around Union South, bringing items from one room to another. The robot is equipped with a camera to enable it to see its surroundings and has two legs to enable walking. How would you define the states of the decision process? What are examples of actions available to the agent? Describe a reward function such that the optimal policy for this reward function will enable the robot to successfully bring items from one room to another.

   The state could include the location of the robot and positions of other objects and agents around it, whether the robot has an item or not, and what room the robot should take the item to. Full credit would be given as long as the state can reasonably be seen as Markov, i.e., saying "the robot's current camera view" would not receive credit.

   The actions could be controls sent to the legs to make it move.

   The reward is +1 when the robot arrives in the correct room with the correct item.

4. Consider the MDP with 3 states $\{A, B, C\}$ and a special terminal state s_∞. In each of the states $\{A, B, C\}$ the agent can take either action a_1 or action a_2. Action a_1 takes the agent to the next state (i.e., A leads to B, B leads to C, and C leads to s_∞) whereas action a_2 keeps the agent in the same state. In s_∞ there is only a single action that results in a transition back to s_∞. The reward function always gives zero reward *except* for when the agent takes action a_1 in state C where it will receive reward 1. Let $\gamma = 0.8$. What is the state-value function for this MDP and the policy that selects action a_1 with probability 0.5? Give the value for all states.

   Multiple ways to solve this. One approach is to use the Bellman equation for state-values to obtain a system of equations with 4 equations and 4 unknowns:

$$\begin{aligned} v_\pi(A) =& 0.5\gamma v_\pi(A) + 0.5\gamma v_\pi(B) && (1)\\ v_\pi(B) =& 0.5\gamma v_\pi(B) + 0.5\gamma v_\pi(C) && (2)\\ v_\pi(C) =& 0.5\gamma v_\pi(C) + 0.5(1 + \gamma v_\pi(s_\infty)) && (3)\\ v_\pi(s_\infty) =& \gamma v_\pi(s_\infty). && (4) \end{aligned}$$

   You can solve this system of equations to find the value for each state. Alternatively, you can note that the value for each state depends upon it's own value and the value of the state after it. This allows you to work backwards and compute $v_\pi(s_\infty) = 0$, $v_\pi(C) = 5/6$, $v_\pi(B) = \frac{0.5\gamma v_\pi(C)}{1-0.5\gamma} = \frac{(2/5)(5/6)}{6/10} = \frac{20}{36} = \frac{5}{9}$, $v_\pi(A) = \frac{0.5\gamma v_\pi(B)}{1-0.5\gamma} = \frac{(2/5)(5/9)}{(6/10)} = \frac{20}{54} = \frac{10}{27}$.

5. Prove the statement $v_\pi(s) = \mathbf{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1})]$ where $\mathbf{E}_\pi$ denotes the expected value when actions are taken by following policy π and states and rewards are sampled from the MDP's transition and reward function.

$$\begin{aligned} v_\pi(s) =& \mathbf{E}_\pi[R_{t+1}\gamma v_\pi(S_{t+1})]\\ =& \mathbf{E}_\pi[R_{t+1}\gamma \mathbf{E}_\pi[q_\pi(S_{t+1}, A_{t+1})]]\\ =& \mathbf{E}_\pi[R_{t+1}\gamma q_\pi(S_{t+1}, A_{t+1})] \end{aligned}$$

   First equation is the Bellman equation, second uses the fact that the state-value for state s is the expected action-value under policy π at state s, final equation uses the definition of expected value when following policy π.

6. Policy-based RL algorithms only learn policies and never value functions. [True / False] False, actor-critic algorithms are policy-based methods that learn both policies and value functions.

8 Societal Implications of Machine Learning

1. Adversarial examples can make neural network classifiers output incorrect classifications, though these examples usually distort the input so that people can detect them [True / False].

   False, adversarial examples are usually imperceptible to people and make neural networks misclassify.

2. Removing names from sensitive data is sufficient to prevent personal data being recovered from the anonymous data [True / False].

   False, it is often possible to find other data sources *with* names and match that data with anonymized data so as to determine identifies.