Final Exam

600.464/664 Artificial Intelligence
Spring 2024

Name:

Instructions

- Please be sure to write your name in the space above!
- Please be sure to read through the entire exam before you start, and be mindful of the time. If one question is taking too long, it may be worth moving on and coming back to the problem question(s) after the rest of the exam is complete.
- Remember that you are only allowed one sheet (both sides) of notes, everything else besides that and the test itself must be off of your workspace.
- Please show ALL relevant work for your answers and provide explanations when prompted. Failure to do either will result in loss of credit.

Probabilistic Reasoning *20 points*

When surveying adults in the United States, on average we find the following probabilities.

	Prefers Blue	Prefers Pink
Is Tall	.32	.08
Is not Tall	.38	.22

1. (5 points) Show mathematically if color preference and being tall are independent or not independent of each other.

   Not independent:
   $p(b)p(t) = (p(b,t)+p(b,n)) \times (p(b,t)+p(p,t)) = (.32+.38) \times (.32+.08) = .7 \times .4 = .28 \neq .32 = p(b,t)$

2. (5 points) A more detailed survey adds a question about gender. It finds the following probabilities.

	Prefers Blue / Is Tall	Prefers Pink / Is Tall	Prefers Blue / Is not Tall	Prefers Pink / Is not Tall
Female	.05	.05	.20	.20
Male	.27	.03	.18	.02

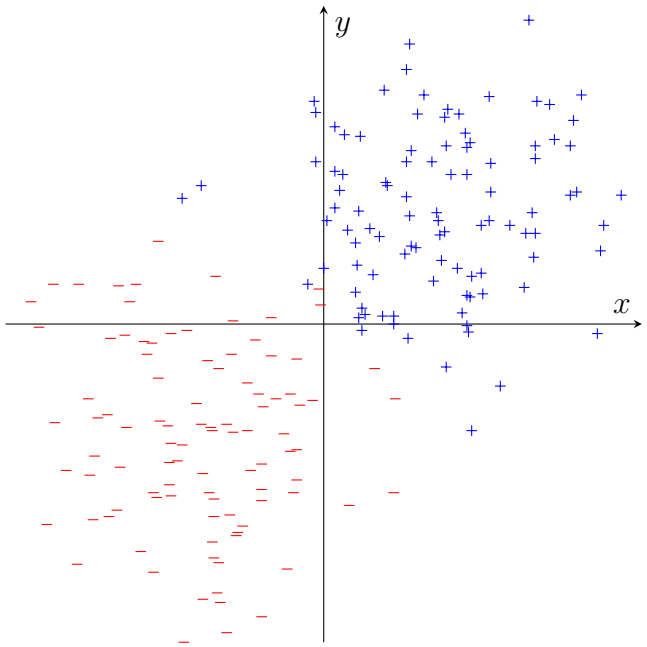
   Show mathematically if color preference and being tall are conditionally independent given gender?

   $p(b,t|m) = \frac{p(b,t,m)}{\sum_{b',t'} p(b',t',m)} = \frac{.27}{.27+.03+.18+.02} = \frac{.27}{.5} = .54$
   $p(b|m) = \frac{\sum_{t'} p(b,t',m)}{\sum_{t',m'} p(b,t',m')} = \frac{.27+.18}{.27+.18+.05+.20} = \frac{.45}{.50} = .9$
   $p(t|m) = \frac{\sum_{b'} p(b',t,m)}{\sum_{b',m} p(b',t',m)} = \frac{.27+.03}{.27+.18+.05+.20} = \frac{.30}{.50} = .6$
   $p(b|m) \times p(t|m) = .9 \times .6 = .54$
   Since both are .54, they are conditionally independent.

3. (10 points) Draw a Bayesian network that models these probabilities. Start the network with gender and have the model reflect any conditional independence you have found.

Machine Learning *20 points*

The following data samples are drawn from two Gaussian distributions. The positive data points from a distribution centered in (1,1), the negative examples from a distribution centered in (-1,-1).



1. (5 points) How would a trained linear classifier divide up the space? Sketch an estimate of the decision boundary into the graph.

   Line more or less diagonal from top left to bottom right across the original

2. (5 points) How would a powerful classifier that is able to classify all training examples correctly divide up the space? Sketch an estimate of the decision boundary into the graph.

   Very detail line that snakes around each outlier data point.

3. (10 point) Which model will likely do better on new unseen test samples? Explain how this example illustrates the concepts: *Overfitting* and *Ockham's razor*.

   The more powerful model is likely to overfit, e.g., memorize the training examples and generalize less. Ockham's razor prefers a simpler solution, in this case the linear classifier.

Deep Learning *20 points*

Consider the deep learning model defined by the following equations.

Input values: x_1, x_2
State variables: s_i
Trainable parameters: w_i
Computed value: y
Correct output value: t
Loss: $\frac{1}{2}(t-y)^2$

$$
\begin{aligned}
s_1 &= x_1 \times w_1 \\
s_2 &= x_2 \times w_2 \\
s_3 &= s_1 + s_2 \\
s_4 &= s_3 \times s_3 \\
y &= s_4 \times w_3
\end{aligned}
$$

1. (5 points) Draw a computation graph for this model, using similar notation to the one we used in class.

   (no solution - should be obvious)

2. (5 points) If the input is $(x_1, x_2) = (2, -1)$ and all initial weights are 1, and the correct output value is 10, what is the loss?

   $s_1 = 2 \times 1 = 2$
   $s_2 = -1 \times 1 = -1$
   $s_3 = 2 + -1 = 1$
   $s_4 = 1 \times 1 = 1$
   $y = 1 \times 1 = 1$
   $L = \frac{1}{2}(10-1)^2 = 81/2$

3. (5 points) Add formulas to the graph that allow backpropagation to update the weights w_i.

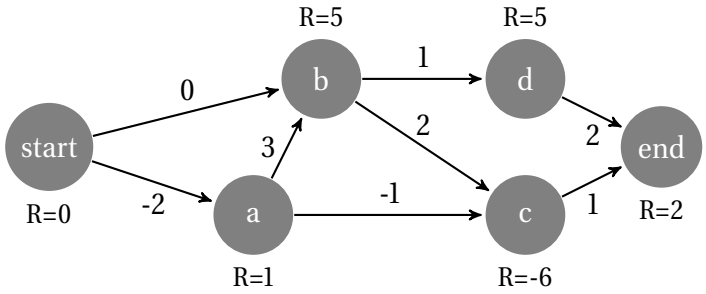
   $\frac{\partial L}{\partial y} = t - y$ $\frac{\partial y}{\partial s_4} = w_3$ $\frac{\partial y}{\partial w_3} = s_4$ $\frac{\partial s4}{\partial s_3} = 2s_3$ $\frac{\partial s3}{\partial s_2} = 1$ $\frac{\partial s3}{\partial s_1} = 1$ $\frac{\partial s2}{\partial w_2} = x_2$ $\frac{\partial s1}{\partial w_1} = x_1$

4. (5 points) With a learning rate of 0.01, what is the value of updated weight w_3 after carrying out gradient descent training with this example?

   $\frac{\partial L}{\partial w_1} = \frac{L(y(w_3))}{\partial w_3} = L'(y)y'(w_3) = (t-y)s_4 = (10-16)16 = 96$
   $\Delta w_3 = -96 \times 0.01 = -0.96$
   $\Delta w_3 \leftarrow 1 - 0.96 = 0.04$

Reinforcement Learning *20 points*

Consider the *deterministic* reinforcement environment drawn below. Let $\gamma = 1$. *Immediate* rewards are indicated at the nodes (the number after R). The current state of the Q table is indicated on the arcs. Once the agent reaches the **end** state the current episode ends.



Recall the update formula for Q-Learning (simplified from the class lecture).

$$\Delta Q(s,a) = \alpha(R(s') + \gamma \max_{a'} Q(s',a') - Q(s,a))$$

1. (4 points) Assuming our RL agent exploits its policy, what is the path it will take from start to end?

   start, b, c, end

2. (12 points) Following a random path selection, the agent RL takes the path (start, a, c, end). Assume a learning rate of $\alpha = 1$. How are the Q values updated?

   (a) start→a

   Update: $1 + -1 - -2 = 2$, *new value:* $-2 + 2 = 0$

   (b) a→c

   Update: $-6 + 1 - -1 = -4$, *new value:* $-1 + -4 = -5$

   (c) c→end

   Update: $2 + 0 - 1 = 1$, *new value:* $1 + 1 = 2$

3. (4 points) With sufficient number of random paths, will the policy learned with Q-learning convergence to optimal values?

   Yes.

Generative Artificial Intelligence *20 points*

You want to apply the idea of diffusion to language models. For the purpose of this question, we treat the problem of generative language models to generate a sentence of English, e.g., *My cat likes to jump on the table.*

1. (5 points) Define the process the generates training data for this approach. What would this process do with the sentence *My cat likes to jump on the table*? Give an example training item generated from it.

   Randomly change each word, flip the position of two words, add or drop words with some probability. For instance, My cat likes to jump on the table. → My cat likes to moon on the table.

2. (10 point) We want to use a decoder-only Transformer model that learns from the generated training data. Draw a diagram and briefly describe the architecture of this model (no need for formulas or detailed steps such as layer normalization but describe inputs, outputs and intermediate representations).

   Solution should include: noised sequence of words as input, original sequence of words as output. Mapping words to embeddings at input, softmax predictionsa as output. Each transformer layer uses self-attention. Representations in intermediate steps are contextualized word representations.

3. (5 points) What is the key difference between an autoencoder and a variational autoencoder?

   In a variational autoencoder we predict the mean and variance of the distribution that a data item is mapped to. A subsequent random step generates the representation. In a regular autoencoder, the representation is directly predicted.