USC CSCI 467 Intro to Machine Learning	**Final Exam** May 4, 2023, 2:00-4:00pm	Spring 2023 Instructor: Robin Jia

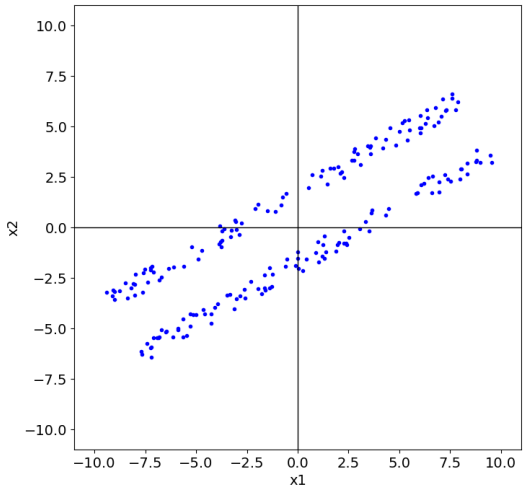
Name: _____

USC e-mail: _____@usc.edu

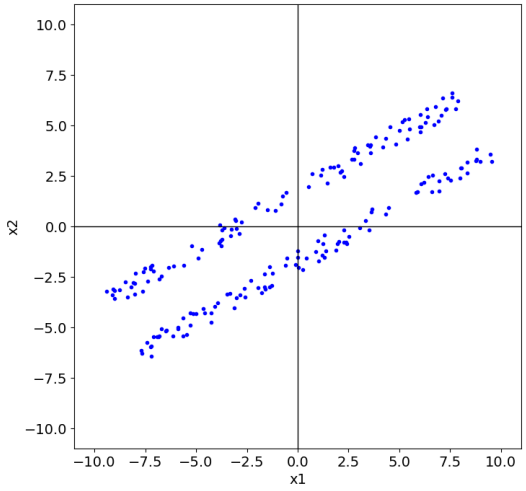
Answer the questions in the spaces provided. **If you write solutions on the back of the pages, indicate this on the front of the pages so we know to look there, but please try to avoid this if possible.** You may use the backs of pages for scratch work. This exam has 6 questions, for a total of 150 points. Note that the questions are not ordered by difficulty; we recommend that you try every problem before spending too much time on one problem.

Question 1: Unsupervised Learning in Pictures (16 points)

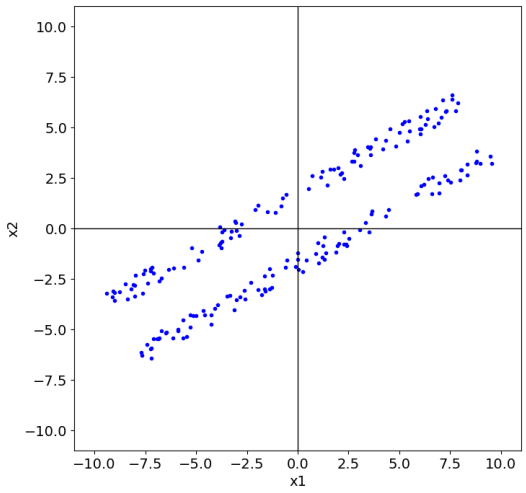
(a) (4 points) For the dataset below, group the examples into two clusters based on how k-Means clustering with $k = 2$ would cluster the data. Explain your reasoning.



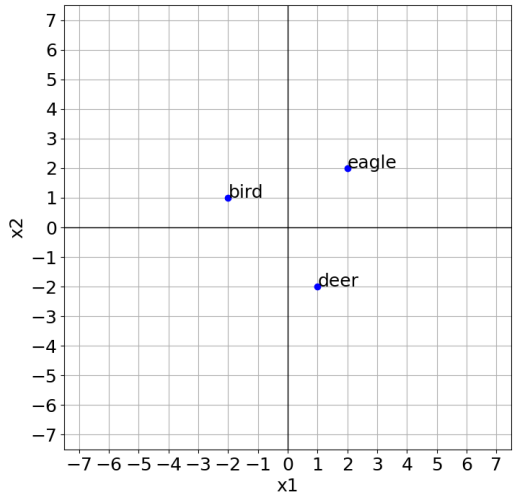
(b) (4 points) For the dataset below, group the examples into two clusters based on how a Gaussian Mixture Model with two clusters would cluster the data. Explain your reasoning.



(c) (4 points) For the dataset below, draw a vector indicating the direction of the first principal component. Explain your reasoning.



(d) (4 points) Below are word vectors for "eagle", "deer", and "bird." Draw the exact location on the grid where you think the word vector for "mammal" should be. Explain your reasoning.



Question 2: k-Means and Linear Classifiers (25 points)

Charlotte has written some code for binary classification and wants to try it out. However, she only has access to an unlabeled dataset $\{x^{(1)}, \ldots, x^{(n)}\}$, where each $x^{(i)} \in \mathbb{R}^d$. She decides to run k-Means clustering on this dataset with $k = 2$. This yields cluster centroids $\mu_1, \mu_2 \in \mathbb{R}^d$, as well as an assigned cluster $z_i \in \{1, 2\}$ for each example $x^{(i)}$. She then defines $y^{(i)}$ to be 1 if $z_i = 1$ and -1 if $z_i = 2$, and creates the supervised binary classification dataset $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}$.

Throughout this problem, you may assume that there are no examples $x^{(i)}$ that are equally close to μ_1 and μ_2.

(a) (3 points) In terms of μ_1, μ_2, and $x^{(i)}$, write a formula for when $y^{(i)} = 1$ and when $y^{(i)} = -1$. Fill in the blank in the equation below:

$$y^{(i)} = \begin{cases} 1 & \text{if } \underline{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad} \\ -1 & \text{otherwise} \end{cases}$$

(b) (5 points) The code Charlotte has written learns a linear decision boundary for binary classification. Is it most likely that she has implemented a multi-layer perceptron with tanh activation function, linear regression, logistic regression, or k-Nearest Neighbors? Explain why your answer is correct and why the other three answers are wrong.

(c) (10 points) Prove that the binary classification task that Charlotte has created is linearly separable. To do this, you should find a vector $w \in \mathbb{R}^d$ and bias $b \in \mathbb{R}$ such that $w^\top x^{(i)} + b > 0$ if $y^{(i)} = 1$ and $w^\top x^{(i)} + b < 0$ if $y^{(i)} = -1$ for all $i = 1, \ldots, n$. Please circle your answer for the values of w and b (they will be some expressions in terms of μ_1 and μ_2). Hint: You should start with your expression from part (a).

(d) (7 points) Charlotte randomly splits the labeled dataset she created into a training set and test set, then trains her linear binary classifier on the training dataset. Draw a possible dataset and train/test split where Charlotte could achieve 100% accuracy on the training set but less than 100% accuracy on the test set. Explain your reasoning. Assume that her code has no bugs.

Question 3: EM for a One-dimensional GMM (23 points)

In this problem, you will do one step of the EM algorithm for a Gaussian Mixture Model. We have a dataset with 5 examples $\{x^{(1)}, \ldots, x^{(5)}\}$, where each $x^{(i)}$ is a scalar. X_i is the random variable denoting the i-th example (whose observed value is $x^{(i)}$), and Z_i is the latent random variable denoting the cluster that the i-th example came from.

We will start with the E-step. Our current guess of π is $[\mathbf{0.4}, \mathbf{0.6}]$ (recall that π_c is the prior probability of an example coming from cluster c). Based on our current guesses for the means $\mu^{(1)}, \mu^{(2)}$ and standard deviations $\sigma^{(1)}, \sigma^{(2)}$ for each of the two clusters, we have already computed the probability density for each datapoint conditioned on being on each cluster. This information, as well as the values of all the $x^{(i)}$'s, is shown in the table below:

i	$x^{(i)}$	$P(X_i = x^{(i)} \mid Z_i = 1; \mu^{(1)}, \sigma^{(1)})$	$P(X_i = x^{(i)} \mid Z_i = 2; \mu^{(2)}, \sigma^{(2)})$
1	4	0.05	0.3
2	3	≈ 0	0.3
3	10	0.5	≈ 0
4	6	0.3	0.2
5	1	≈ 0	0.2

Where values are ≈ 0, you may treat them as being equal to 0 in your calculations (even though technically, these probabilities will never be exactly 0).[1]

(a) (10 points) Do the E-step, which computes the value $r_{ic} = P(Z_i = c \mid X_i = x^{(i)})$ for each $i \in \{1, \ldots, 5\}$ and for each cluster $c = \{1, 2\}$. Fill out the table below, showing your work in the space on the next page. Each answer in the table should be a single number (not an unsimplified expression).

i	r_{i1}	r_{i2}
1		
2		
3		
4		
5		

[1] The values in the table are also not from an actual Gaussian pdf, but were chosen to make the arithmetic nice.

[Scratch space for problem 3a]

(b) (5 points) Using your E step calculations, do the M step update for π. Circle your final answer, which should be a new value for π. Show your work.

(c) (8 points) Using your E step calculations, do the M step update for $\mu^{(1)}$ and $\mu^{(2)}$. Circle your final answers, which should be values of $\mu^{(1)}$ and $\mu^{(2)}$. Show your work. You may write your answers as the quotient of two floating point numbers.

Question 4: Explaining Reinforcement Learning (28 points)

In this problem, you must explain reinforcement learning concepts **in English**. **Do not use any equations.** You may use the letters s and a to denote a state and action, respectively.

(a) (6 points) Provide a definition of the Q function. What are its inputs and what is its output? Be as specific as possible.

(b) (3 points) What objective function is optimized by policy gradient? Does policy gradient minimize or maximize this objective?

(c) (4 points) What is exploration and why is it important during training?

(d) (3 points) What is one strategy that can be used to promote exploration during Q-learning? You should both give the name of the method and describe what it does.

(e) (4 points) Suppose we have a continuous state space and do not discretize the state space. Describe one problem that would occur if we tried to use tabular Q-learning.

(f) (8 points) Your friend wants to use deep Q-learning to play a text-based video game. In this game, the agent takes an action by choosing from a fixed list of commands, after which it receives a text-based description of the new state. Suggest a deep learning architecture that would be suitable for this task. You must specify both how the input will be encoded, as well as how the model will predict the Q value. **For this part only, you may use mathematical notation if desired (not required).**

Question 5: Reweighting Subgroups (23 points)

In class, we discussed how problems can arise when certain types of examples are underrepresented in the data. One natural solution is to *reweight* the data. Suppose we have a training dataset D for linear regression that is the union of two (disjoint) datasets A and B, where $|A| >> |B|$. For example, A and B might represent data from two different groups of individuals. Let $(x_a^{(i)}, y_a^{(i)})$ denote the i-th example in A and $(x_b^{(i)}, y_b^{(i)})$ denote the i-th example in B. The reweighted training loss is defined as

$$L(w) = \frac{1}{|A|} \sum_{i=1}^{|A|} (w^\top x_a^{(i)} - y_a^{(i)})^2 + \frac{1}{|B|} \sum_{i=1}^{|B|} (w^\top x_b^{(i)} - y_b^{(i)})^2,$$

where $w \in \mathbb{R}^d$ is the weight vector parameter for linear regression (in this problem, we will omit the bias term).

(a) (5 points) Explain why this loss function is more likely to promote more equal treatment of individuals in dataset A and individuals in dataset B, compared with running normal linear regression on D.

(b) (8 points) This loss function can be optimized by gradient descent. Compute the gradient of $L(w)$ with respect to w.

(c) Suppose you run gradient descent on this loss function. You achieve 0 error on training data from group A as well as on training data from group B. However, test error is much higher on individuals from group B than from group A.

i. (5 points) Explain why it would make sense that test error is higher on group B than group A, when both training errors are 0.

ii. (5 points) Suggest a change to the loss function that could help improve test error on group B. Explain your reasoning.

Question 6: Short Answer (35 points)

In the following questions, circle the correct answer(s).

(a) Consider a multi-headed attention layer in a Transformer. Let q_t, k_t, and v_t denote the query, key, and value vectors, respectively, for the t-th word, and let o_t denote the output of the multi-headed attention layer for the t-th word. We have an input sentence that is 3 words long. The output for the third word, o_3, can be written in the following way:

$$p = \text{softmax}(\_\_\_\_, \_\_\_\_, \_\_\_\_)$$
$$o_3 = \_\_\_\_\_$$

Answer the following questions:

i. (2 points) What list of three expressions goes in the three blanks in the first line?

A. $q_3^\top k_1$, $q_3^\top k_2$, $q_3^\top k_3$

B. $q_1^\top k_3$, $q_2^\top k_3$, $q_3^\top k_3$

C. $e^{q_3^\top k_1}$, $e^{q_3^\top k_2}$, $e^{q_3^\top k_3}$

D. $e^{q_1^\top k_3}$, $e^{q_2^\top k_3}$, $e^{q_3^\top k_3}$

ii. (2 points) What is the definition of $\text{softmax}(x_1, \ldots, x_n)$? Recall that this function takes in a list of n real numbers $x_1, \ldots, x_n$ and outputs a list of n real numbers.

A. $\text{softmax}(x_1, \ldots, x_n) = \left[\frac{x_1}{\sum_{i=1}^n x_i}, \ldots, \frac{x_n}{\sum_{i=1}^n x_i}\right]$

B. $\text{softmax}(x_1, \ldots, x_n) = \left[\frac{e^{x_1}}{\sum_{i=1}^n e^{x_i}}, \ldots, \frac{e^{x_n}}{\sum_{i=1}^n e^{x_i}}\right]$

C. $\text{softmax}(x_1, \ldots, x_n) = \left[\frac{e^{x_1}}{e^{\sum_{i=1}^n x_i}}, \ldots, \frac{e^{x_n}}{e^{\sum_{i=1}^n x_i}}\right]$

D. $\text{softmax}(x_1, \ldots, x_n) = [e^{x_1}, \ldots, e^{x_n}]$

iii. (2 points) What expression goes in the blank on the second line?

A. $p_3 \cdot v_3$

B. $p_3 \cdot q_t^\top v_t$

C. $\sum_{t=1}^{3} p_t \cdot v_t$

D. $\sum_{t=1}^{3} p_t \cdot q_t^\top v_t$

(b) Recall that the Upper Confidence Bound (UCB) algorithm for bandits uses the formula

$$UCB_t(a) = \hat{\mu}(a) + \sqrt{\frac{2 \log t}{n_t(a)}}.$$

Answer the following questions:

i. (2 points) Which of the following is the most accurate definition of $\hat{\mu}(a)$?

A. The probability of choosing action a.

B. The expected reward for action a.

C. The amount of uncertainty we have about the expected reward for action a.

D. The average reward from times when the agent chose action a.

ii. (2 points) Of the two terms in the UCB formula, which one(s) cause the UCB algorithm to try many different arms early on?

A. First term only

B. Second term only

C. Both terms

D. Neither term

(c) (5 points) In a Hidden Markov Model, which of the following assumptions are made? Choose all that apply. x_t denotes the observation at time t and z_t denotes the hidden state at time t.

A. x_t depends only on z_t.

B. x_t is independent of x_{t-1} (without conditioning on any other random variables).

C. z_t depends only on the previous hidden state z_{t-1}.

D. The probability distribution $p(z_t \mid z_{t-1})$ is the same for all timesteps t.

E. The first hidden state z_1 is chosen uniformly at random from all possible states.

(d) Circle True or False for each statement below.

i. (2 points) **True** or **False**: k-means clustering minimizes a convex loss function, so it will converge to the same clusters no matter how you initialize the cluster centroids.

ii. (2 points) **True** or **False**: word2vec is based on the idea that words with similar meanings tend to appear in similar contexts.

iii. (2 points) **True** or **False**: In PCA, we project the mean-centered data matrix X to a lower dimensional space by choosing the eigenvectors of X corresponding to the largest eigenvalues.

iv. (2 points) **True** or **False**: In an HMM, the Viterbi algorithm can be used to find $p(z_2 \mid x_{1:T})$. Again, x_t denotes the observation at time t and z_t denotes the hidden state at time t.

v. (2 points) **True** or **False**: The Fast Gradient Sign Method (FGSM) takes the gradient of the loss with respect to the model's parameters to create an adversarial example.

vi. (2 points) **True** or **False**: Different fairness metrics can be fundamentally at odds with one another.

(e) Each question below describes a scenario. From the options below, choose the machine learning setting that best matches the given scenario. Just write the single letter of your answer; no explanation required.

A. Regression
B. Classification
C. Clustering
D. Dimensionality Reduction
E. Bandit Problem
F. Reinforcement Learning

i. (2 points) Aman has a collection of video games. He wants to find groups of games that are similar to each other.

i. _____

ii. (2 points) Zhihan is taking care of a plant. Every day, he needs to decide how much water to give it.

ii. _____

iii. (2 points) Phakawat wants to find the perfect chocolate chip cookie recipe. He bakes many batches of cookies on different days, varying the recipe over time. He doesn't own a scale to weigh ingredients precisely, so the amounts of every ingredient are slightly different every time he makes the same recipe.

iii. _____

iv. (2 points) Qinyuan is playing golf. She wants to estimate by eyesight how many feet her ball is from the hole.

iv. _____