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

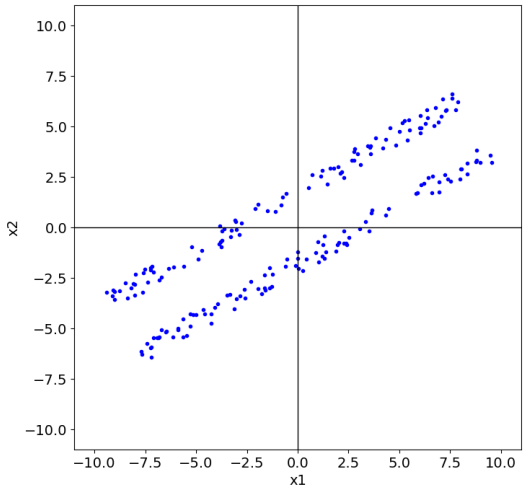
Name: _____

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.

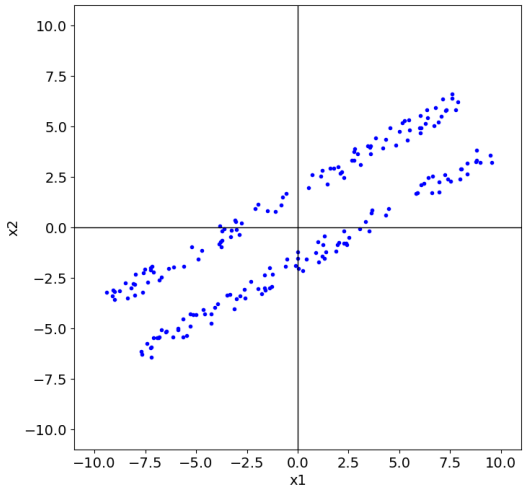
Question 1: Unsupervised Learning in Pictures (16 points)

(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.



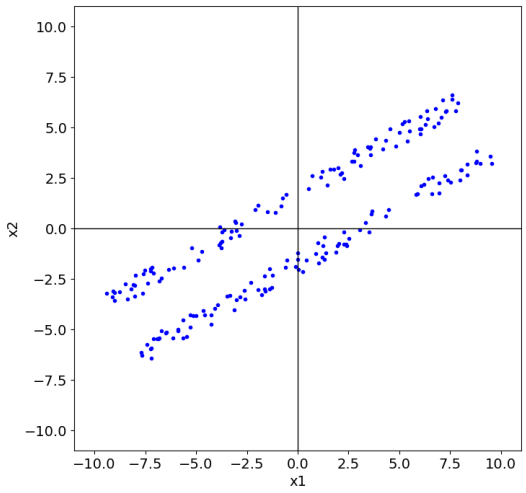
Solution: You will learn two equally-sized circular clusters, one for the left half of the data and one for the right half of the data. This is because k-Means only cares about Euclidean distance–it cannot learn a non-spherical cluster shape.

(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.



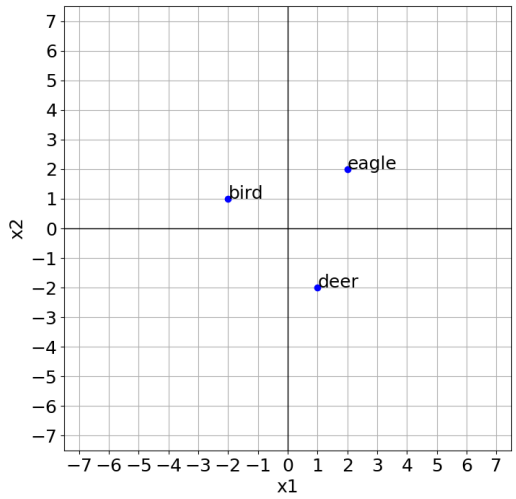
Solution: You will learn one cluster for the top part and one cluster for the bottom part. This is because a GMM can learn a covariance matrix that matches the covariance of the data.

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



Solution: The correct answer is a vector pointing to the right and 30 degrees upwards from the x-axis (in other words, parallel to the "lines" of the data). This is the direction that minimizes squared distance of points to the line, or equivalently projecting onto this keeps the most variance of the data.

(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.



Solution: The correct answer is $(-3, -3)$. This is because mammal is to deer as bird is to eagle, and the difference $bird - eagle = (-4, -1)$, so we should add $(-4, -1)$ to

the deer vector.

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.

Solution: Out of these methods, only logistic regression learns a linear decision boundary for binary classification. MLP and k-Nearest Neighbors do not learn linear decision boundaries, and linear regression is for regression not classification.

Many students confused k-NN with k-Means. k-NN is a supervised learning algorithm while k-means is unsupervised. k-NN can be used for classification.

Another common mistake was to say that MLP is for multi-class classification. An MLP is a type of neural network, and can be combined with any type of output layer to do binary classification, multi-class classification, or regression.

(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).

Solution: k-Means assigns a cluster based on whether $x^{(i)}$ is closer to μ_1 or μ_2. We

know that

$$\begin{aligned} y^{(i)} = 1 &\Leftrightarrow \|x^{(i)} - \mu_1\|^2 < \|x^{(i)} - \mu_2\|^2 \\ &\Leftrightarrow (x^{(i)} - \mu_1)^\top (x^{(i)} - \mu_1) < (x^{(i)} - \mu_2)^\top (x^{(i)} - \mu_2) \\ &\Leftrightarrow \|x^{(i)}\|^2 - 2x^{(i)\top}\mu_1 + \|\mu_1\|^2 < \|x^{(i)}\|^2 - 2x^{(i)\top}\mu_2 + \|\mu_2\|^2 \\ &\Leftrightarrow 2(\mu_1 - \mu_2)^\top x^{(i)} + (\|\mu_2\|^2 - \|\mu_1\|^2) > 0. \end{aligned}$$

So, we choose $w = 2(\mu_1 - \mu_2)$ and $b = \|\mu_2\|^2 - \|\mu_1\|^2$.

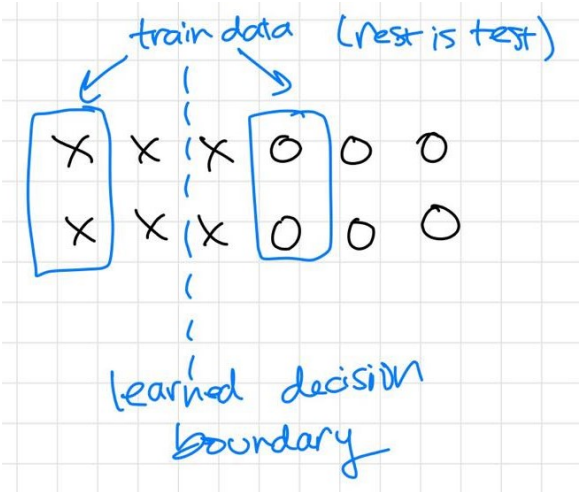
Note that you can multiply both w and b by a positive constant and still be correct. So for example, $w = \mu_1 - \mu_2$ and $b = \frac{1}{2}(\|\mu_2\|^2 - \|\mu_1\|^2)$ is also a correct pair.

Some students got the correct answer but did not use the identity $\|x\|^2 = x^\top x$; they instead wrote out the squared L2 norm as a sum over the dimension. This is still correct, although knowing $\|x\|^2 = x^\top x$ makes the proof easier.

One common problem was to pretend that the x's and μ's are scalars. This results in something that is almost correct, but involves terms like μ_1^2. Squaring a vector is not a valid mathematical operation, so this lost some points.

(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.

Solution: There are many pictures you could draw. Here's one example:



The basic idea is that even though the dataset is linearly separable, with a small training dataset the learning algorithm may choose a different decision boundary that makes mistakes on the test set. In other words, the model will overfit on the small training dataset.

For full credit, you needed to use a linearly separable dataset as an example.

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		

Solution: The ones with ≈ 0 are easy: we can infer that Z_i must be equal to the only cluster with non-zero probability of generating this $x^{(i)}$. For the other two cases, we

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

need to use Bayes Rule. For $i = 1$:

$$P(X_i = x^{(1)}, Z_1 = 1) = 0.4 \cdot 0.05 = 0.02$$
$$P(X_i = x^{(1)}, Z_1 = 2) = 0.6 \cdot 0.3 = 0.18$$
$$r_{11} = 0.1$$
$$r_{11} = 0.9$$

For $i = 4$:

$$P(X_i = x^{(1)}, Z_1 = 1) = 0.4 \cdot 0.3 = 0.12$$
$$P(X_i = x^{(1)}, Z_1 = 2) = 0.6 \cdot 0.2 = 0.12$$
$$r_{11} = 0.5$$
$$r_{11} = 0.5$$

Putting this all together, we have:

i	r_{i1}	r_{i2}
1	0.1	0.9
2	0	1
3	1	0
4	0.5	0.5
5	0	1

The most common error was to make errors when normalizing to compute r_{11}. Remember that we are using Bayes Rule, so you can always write out all the possible numerators (in this case, $P(X_i = x^{(i)}, Z_i = 1)$ and $P(X_i = x^{(i)}, Z_i = 2)$) and divide by the sum to ensure your resulting distribution sums to 1.

(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.

Solution: The M-step formula is to sum up the "pseudo-counts" for each cluster, then divide by the total number of examples So we have

$$\pi = \left[\frac{0.1 + 1 + 0.5}{5}, \frac{0.9 + 1 + 0.5 + 1}{5}\right] = [1.6/5, 3.4/5] = [0.32, 0.68].$$

(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.

Solution: The M-step formula here is to take a weighted average of the datapoints within each cluster, weighted by the probability that each datapoint belongs to that

cluster. So we have

$$\mu^{(1)} = \frac{0.1 \cdot 4 + 1 \cdot 10 + 0.5 \cdot 6}{1.6}$$
$$= \boxed{\frac{13.4}{1.6}} \approx 8.4$$
$$\mu^{(2)} = \frac{0.9 \cdot 4 + 1 \cdot 3 + 0.5 \cdot 6 + 1 \cdot 1}{3.4}$$
$$= \boxed{\frac{10.6}{3.4}} \approx 3.1$$

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.

Solution: The Q function takes as input a state s and an action a. It outputs the expected discounted reward of the optimal policy for the trajectory after starting at state s and taking action a.

For full credit, you had to say "expected" and "discounted" (or equivalent terms), and specify that we are talking about the optimal policy.

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

Solution: Policy gradient maximizes the expected reward of the policy.

For full credit, you have to say the word "expected" (or some equivalent phrase) and also specify what you are taking the expected value of. (It is optional here to say the word "discounted," as we only described the case without discounting during lecture.)

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

Solution: Exploration is the practice of trying new actions to gain information about how good they are, even if we currently believe these are not the optimal action. This is important to gain information about which actions are the best.

(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.

Solution: One strategy is to sometimes choose an action randomly instead of always choosing the action that we believe is best based on observations made so far.

(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.

Solution: There is an infinite state space, so we may run out of memory to store all the Q values, Even if we don't, at test time we will encounter states we have never seen during training, so the agent will not know what action to choose.

(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).**

Solution: A reasonable deep learning architecture would be to encode the text description of the state s with either an RNN or Transformer. We would also learn a parameter vector for each possible command a. We would predict the Q function for a state s and action s by dot producting the representation of s with the vector for a. Some common mistakes here included not describing how you predict the Q-value for each action, or only saying the loss function and update rule but not the neural architecture that would be used. It was also important to specify that the model outputs Q-values, not probabilities of actions.

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.

Solution: Since $1/|A|$ is much smaller than $1/|B|$, this objective function gives more importance to the loss on examples in group B. Thus, the model is less likely to have much worse accuracy on group B than A. In contrast, standard linear regression puts equal importance on each example, so group A as a whole is weighted more heavily.

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

Solution:

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

(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.

Solution: Since dataset B is smaller, overfitting is more of a problem for dataset B than dataset A.

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

Solution: The problem with group B is that we are overfitting. Therefore, a natural improvement is to add a regularization penalty, such as L1 or L2, to reduce overfitting.
Note that it would not be helpful in this case to increase the weight on dataset B during training, because the training error is already at 0 for both datasets. Even if we increase the weight on dataset B, we would still get 0 training error, so it would not make a difference.

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

Solution: A, B, C.

(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

Solution: D, B.

(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.

Solution: A, C, D.
B is false—they are not independent because x_{t-1} depends on z_{t-1}, and x_t depends on z_t which also depends on z_{t-1}.

(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.

Solution: False. k-means is not a convex problem and different initializations can lead to different results.

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

Solution: True. This is the distributional hypothesis.

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.

Solution: False. X is not even guaranteed to be a square matrix. We take the eigendecomposition of $X^\top X$, the covariance matrix.

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.

Solution: False. The forward-backward algorithm does this type of inference.

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.

Solution: False. FGSM takes the gradient of the loss with respect to the input, not the model's parameters.

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

Solution: True. This was discussed during lecture with the 3 fairness metrics we saw.

(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. _____

Solution:

i. C: This is a standard clustering problem.

ii. F: This is a reinforcement learning problem because Zhihan has to decide what action to take every day, and there is a state that persists from day to day (the health of the plant, the amount of water in the soil, etc.).

iii. E: This is a bandit problem because Phakawat is also taking a new action every time, but there is no persistent state from one batch of cookies to the next. Note that there is randomness in how each batch turns out because he does not measure the ingredients super accurately.

iv. A: This is a regression problem because Qinyuan is taking in some input (i.e., her visual field) and trying to predict a real number.