CSE/STAT 416 **Name:** ______________________

Spring 2023

Final Exam

6/8/2023 **Student Number:** ______________________

Time Limit: 1 hours 50 minutes

Do not open the exam before the exam begins and close the booklet when time is called. Starting early or working after time is called will lead to a -10 deduction. You may write your name and student number on the front of the exam before the exam starts.

This exam contains 16 pages (including this cover page) and 9 questions. Some questions have sub-parts. True/false, multiple choice, and multiple answer questions will have bubbles for you to fill in. Fill them in clearly to indicate an answer and erase/cross out an answer if you wish for us to not use that choice. **Circles represent problems where you should select one option, while squares represent questions where you should select all that apply.**

You are allowed to have one sheet of paper (both sides) with you as your cheat sheet. All other materials besides writing utensils should be put away before the exam starts. This includes all electronic devices like phones, calculators, and smart watches.

If you need more room to work out your answer to a question, you may ask for scratch paper. If you want to submit work on scratch paper to be graded, you should indicate so on the page of the question in the exam, indicate on the scratch paper which part is the answer, and staple your scratch paper to the **end** of your exam. Failure to do any of these steps may result in your scratch paper not being graded.

Work efficiently. Some questions are easier, some more difficult. Be sure to give yourself time to answer all of the questions. You can move on if you get bogged down in the more difficult ones. Generally, the number of points correlates to how much work is involved in solving a problem.

Initials: ______________________

Initial above to indicate you have read and agreed to these rules. Failure to initial may result in your exam not being accepted for credit.

Good luck!

Question	**Topic**	**Max. score**	**Score**
1	Regularized Linear Regression	13	
2	Classification	16	
3	Fairness	8	
4	Local Methods	6	
5	Deep Learning	20	
6	PCA	5	
7	Recommender Systems	6	
8	Clustering	13	
9	Miscellaneous	13	
	Total	100	
	Extra credit	2	

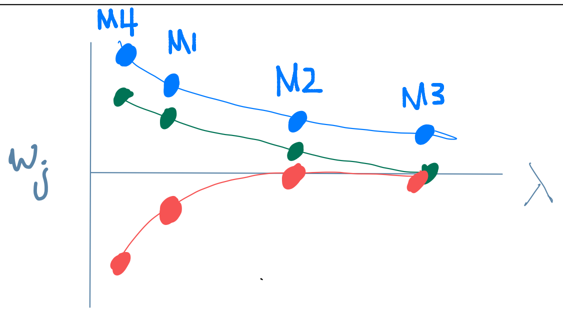
1 Regularized Linear Regression [13 points]

1. [11 points] Four regularization models of the same type, but with varying degrees of regularization strength λ have been trained. Below, the following table shows for each of the models: 1) their coefficients, 2) training error, 3) results from cross validation, and 4) their test error. The four λ strenghts used were $\lambda = 0$, $\lambda = 1$, $\lambda = 5$, $\lambda = 10$. The rows below are shown in no particular order.

Model	Coefficients	Training Error	CV 1 Error	CV 2 Error	CV 3 Error	Average CV Error	Test Error
M1	$0.5h_1(x) + 0.3h_2(x) - 0.1h_3(x)$	0.06	0.09	0.07	0.08	0.08	0.07
M2	$0.1h_1(x) + 0h_2(x) + 0h_3(x)$	0.15	0.16	0.12	0.14	0.14	0.16
M3	$0.35h_1(x) + 0.1h_2(x) + 0h_3(x)$	0.10	0.04	0.08	0.15	0.09	0.05
M4	$0.45h_1(x) + 0.2h_2(x) - 0.05h_3(x)$	0.07	0.08	0.07	0.06	0.07	0.06

(a) [5 pts] In the space below, draw the *coefficient path* for the models as λ increases. Your graph should clearly show the paths for the coefficients for each feature, and you should label which model ($M1 - M4$) exists on each point. Your drawing should be roughly to scale but clearly doesn't need to be measured perfectly.

Solution:



(b) [2 pts] Which type of regularization is most likely being used in this example?

○ **Ridge** ● **LASSO** ○ **Not enough information to tell**

(c) [2 pts] Which model would you choose? **(Select one)**

○ **M1** ○ **M2** ○ **M3** ● **M4**

(d) [2 pt] What would you estimate the generalization error on future data for the model you chose in the last problem?

Generalization error = **0.06**

2. [2 ptsx] Your friend is working on a regression model for a Kaggle competition and consistently sees a high error for their model on the public leaderboard. Based only on the information above, would you recommend adding regularization? If so, explain why regularization would help. If not, what piece(s) of additional information would you need before deciding whether or not you should recommend adding regularization?

Solution: Only knowing the error is high is not sufficient information for if regularization is helpful or not. We need to know if the model is overfitting, which is characterized by high future error and low training error. Asking for if their training error is low would give you a better sense of if the model was overfitting.

2 Classification [16 points]

1. [5 pts] James is training a logistic regression model on a dataset of cartoon reviews using these features, where x is a text sentence in the dataset In this example, a positive label is a good review for a cartoon. The features extracted from this sentence are bag of words counts with the indices for counts of: 1) amazing, 2) ok, 3) bad, 4) not, 5) very. After training the model on this data, we learn the following coefficients for logistic regression (index 0 is the intercept).

$$\hat{w} = [1, 2, 4, -3, -1, 0]$$

Using this model, what is the probability that the following sentence is predicted to be a **negative** review for a cartoon.

> *"Tom and Jerry is very very amazing. It is not bad at all, it is more than ok, it is amazing!"*

Since you do not have a calculator, you do not need to convert your answer as a decimal. You should write out the formula with numbers plugged in. For example, an answer like $1/(e^4)$ would be sufficiently simplified.

Solution:

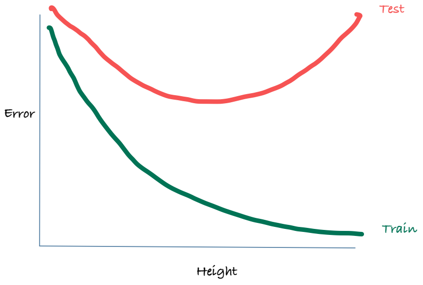
$$Score(x) = 1 + 2 \cdot 2 + 4 \cdot 1 - 3 \cdot 1 - 1 \cdot 1 + 0 \cdot 0 = 5$$

$$P(y = +1|x) = \frac{1}{1 + e^{-Score(x)}} = \frac{1}{1 + e^{-5}}$$

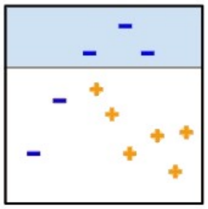
$$P(y = -1|x) = 1 - P(y = +1|x) = 1 - \frac{1}{1 + e^{-5}} = \frac{e^{-5}}{1 + e^{-5}}$$

2. [2 pts] Consider the task of training decision trees of various heights. In the space below, draw a graph that shows the learning curves for training and test error as a function of the height of the tree learned. The x-axis should be max-tree height, and the y-axis should be error.

Solution:

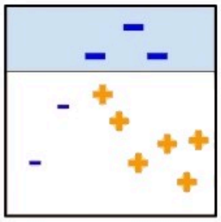


3. [2 pts] At time $t = 1$ of AdaBoost, you learn the following decision boundary (above the line is predicted negative, below the line is predicted positive):

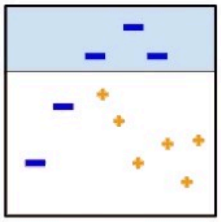


Predicted decision boundary at time *t=1*

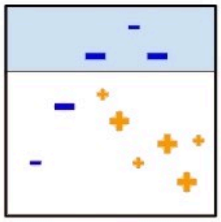
Which of the following represents how AdaBoost would reweight the points as a result of the above decision boundary?



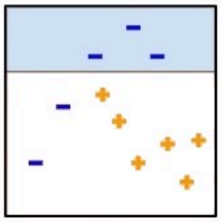
○(A)



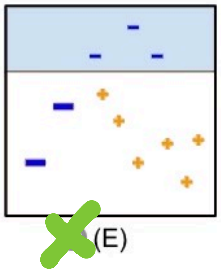
○(B)



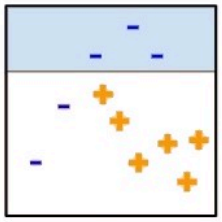
○(C)



○(D)

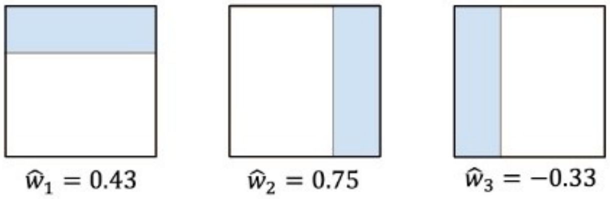


[✓ (E)]

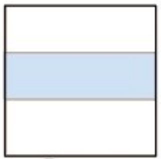


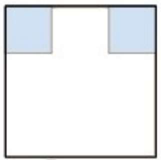
○(F)

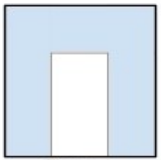
4. [3 pts] After a run of AdaBoost, you learn the below three decision stumps, with their corresponding weights.

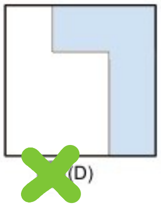


Which of the following represents the final decision boundary of the ensemble?


○ (A)


○ (B)


○ (C)


[✓ (D)]

5. [4 pts] Consider the Random Forest ensemble model. Which of the following statements are generally true about Random Forests.

- □ Combines the power of short trees (decision stumps)
- ■ **Combines the power of tall trees (overfit trees)**
- ■ **Can be trained in parallel**
- ■ **Uses bootstrapping to obtain different trees on each iteration**
- □ Weights misclassified points higher when training the next iteration
- ■ **Can handle non-linear relationships of features and labels**

3 Fairness [8 points]

1. [8 pts] Imagine a company is using a machine learning algorithm to screen resumes and decide who to invite for interviews. The company has noticed that the algorithm is significantly less likely to select female candidates for interviews, even though female interviewed candidates perform just as well in later stages of the hiring process. Below is a confusion matrix of the model's performance where the positive label means the candidate was successful in their interview and then negative label was they were not.

Confusion matrix for male applicants

	Predicted +	**Predicted -**
Actual +	360	40
Actual -	100	500

Confusion matrix for women applicants

	Predicted +	**Predicted -**
Actual +	270	130
Actual -	80	520

(a) [3 pts] Describe a possible reason this discrepancy might be happening. There is not a single correct answer, but outline a reasonable potential reason for this. If you are making any assumptions about what is being included in the model, outline those assumptions when discussing your rationale.

Solution: There are many possible answers here. Some common examples include historical bias from the potential that historically, women might have been interviewed at lower rates or representation bias from the potential that the things asked for in interviews (e.g., hobbies) could disadvantage women.

A common mistake in this question was not explaining a possible reason *why* this issue arose, and pointed out the statistics of the fact there was a problem.

(b) [5 pts] There are many possible fairness metrics that can be used to identify any disparate treatment by the model. Describe a fairness metric we have discussed that you could use to measure the bias present in this model. You do not need to name the fairness metric, but you should clearly describe what it is computing. In particular, use the confusion matrix above to show how to calculate the values of that fairness metric. You should include all of the relevant numbers in your calculation, but you do not need to simplify your answer without a calculator.

Solution: In this example, out of 400 qualified female applicants (Actual Positive), the algorithm correctly selects 270 for interviews and incorrectly rejects 130. Out of 600 unqualified female applicants (Actual Negative), the algorithm correctly rejects 520 and incorrectly selects 80 for interviews.

In both cases, the total number of applicants is the same (1000), and the proportion of qualified applicants is the same (40%). However, the true positive rate (TPR), or ”recall”, is different for men and women. For men, TPR = 360 / 400 = 90%, whereas for women, TPR = 270 / 400 = 67.5%.

This is an example of a fairness issue according to the definition of equal opportunity, which suggests that qualified candidates should have an equal chance of being selected, regardless of their gender.

4 Local Methods [6 points]

1. [4 pts] Consider the task of using 1-Nearest-Neighbor on a dataset with 2 features. Draw a small dataset and query point such the result of finding the 1-NN of the query point will differ if we only changed the distance metric from Euclidean to Manhattan. In other words, make a dataset and query such that the nearest neighbor for the query depends on which distance metric is used.

   Your answer does not need to be complex, just show a dataset with a few points and a query point where the 1-NN differs for a query when using the Euclidean vs the Manhattan distance. You should label the coordinates of the points and the query as well as including the Manhattan and Euclidean distances between the points and the query.

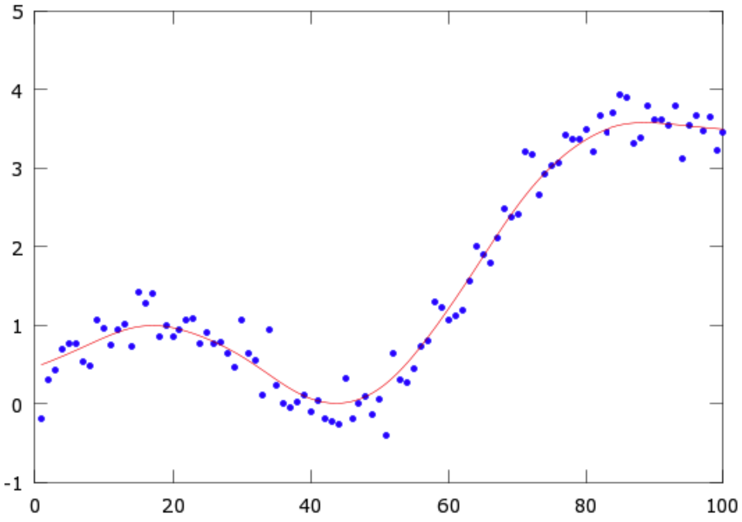
Solution: There are many possible answers to this question. The simplest answer is the query point $x_q = (0, 0)$, $x_1 = (1, 1)$, and $x_2 = (1.5, 0)$. With these points x_1 is the nearest neighbor according to Euclidian distance

$\left(d_{euclid}(x_q, x_1) = \sqrt{2} < 1.5 = d_{euclid}(x_q, x_2)\right)$

but x_2 is the nearest neighbor according to Manhattan distance

$(d_{manhattan}(x_q, x_1) = 2 > 1.5 = d_{manhattan}(x_q, x_2))$

2. [2 pts] The following image shows a predictor for a regression model. Which model is this predictor most likely from?

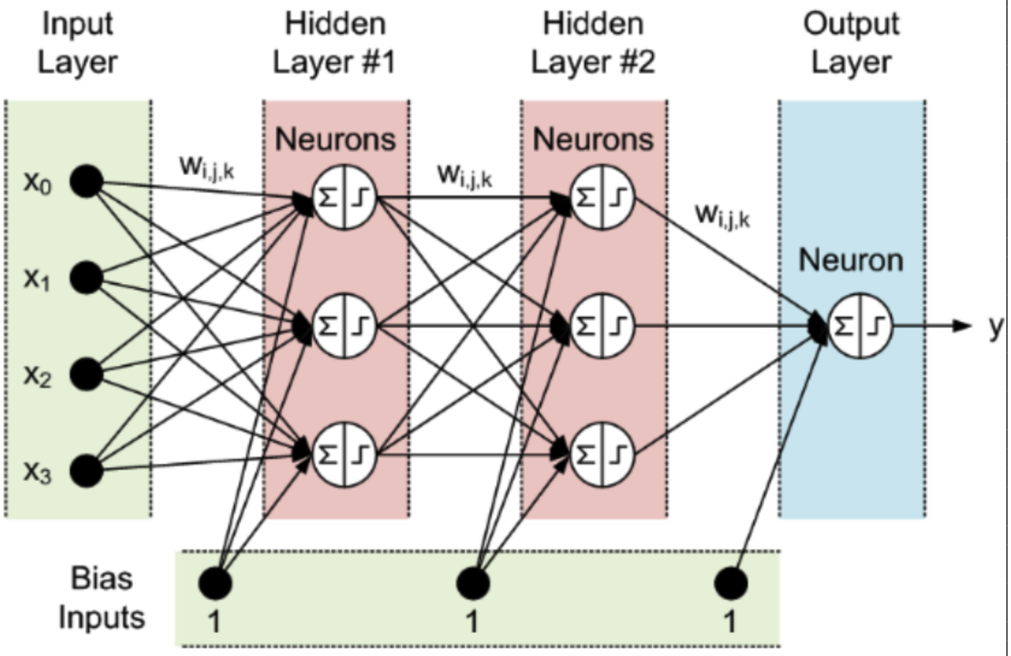


○ 1-NN Regression
○ 5-NN Regression
○ 100-NN Regression
○ Kernel Regression with Boxcar Kernel
● **Kernel Regression with Gaussian Kernel**

5 Deep Learning [20 points]

1. [5 pts] You've been asked to help a friend understand how a simple neural network functions. Create a diagram that includes inputs, weights, a bias, two hidden layers, an activation function, and outputs. Briefly describe each of these components and explain their role in the network.

Solution: Showing the diagram of the network described above. Correct solution should also briefly define each of the components (omitted for space).



2. [4 pts] Select all of the statements that are generally true about deep learning and neural networks

■ **Neural networks are often powerful models, but are very prone to overfitting without careful attention paid to how to train them.**

□ Increasing the number of hidden layers in a neural network always results in a better model.

□ The process of training a neural network always guarantees finding the global minimum of the cost function.

■ **A neural network with no hidden layers can represent linearly separable functions or decisions.**

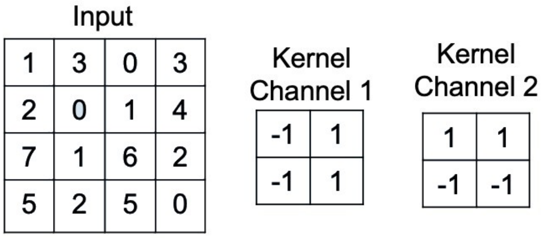
□ A larger learning rate in the training of a neural network always leads to faster and better convergence.

□ A generally used because they are interpretable models.

3. [10 pts] Suppose that you have convolutional neural network with the following components:

- One **2D-convolutional** layer with depth two (e.g. two kernels) with **stride 2** and **1x1 0-padding** (i.e., padding with zeros)
- **Max-pooling** layer of size 2x2 with stride 2
- A fully connected layer (with an intercept term)
- A **ReLU** activation function on the output (note: not usually done in practice).

Suppose you propagate the input below (left) through the CNN with the following kernel weights in the convolutional layer:



Input

1	3	0	3
2	0	1	4
7	1	6	2
5	2	5	0

Kernel Channel 1

-1	1
-1	1

Kernel Channel 2

1	1
-1	-1

(a) [6 pts] What is the output **after** the max-pooling layer? Show your work, but clearly indicate your answer for each channel.

Solution: With the specifics of the problem above, this was not solvable due to a size mismatch. We clarified in the exam to either remove the 0 padding or use max-pool with stride 1. We show both options below. After convolution

C1

0	6
-9	-9

C2

2	-2
1	3

After max pool

C1: 6, C2: 3

(2nd option) After convolution

C1

1	-3	-3
9	6	-6
5	3	0

C2

-1	-3	-3
-5	-6	2
5	7	0

After max pool

C1

9	6
9	6

C2

-1	2
7	7

(b) [3 pts] Suppose the fully connected layer has coefficients:

- Intercept: -1
- First channel: 2
- Second channel: 3

What is the value of the **after** applying the final activation:

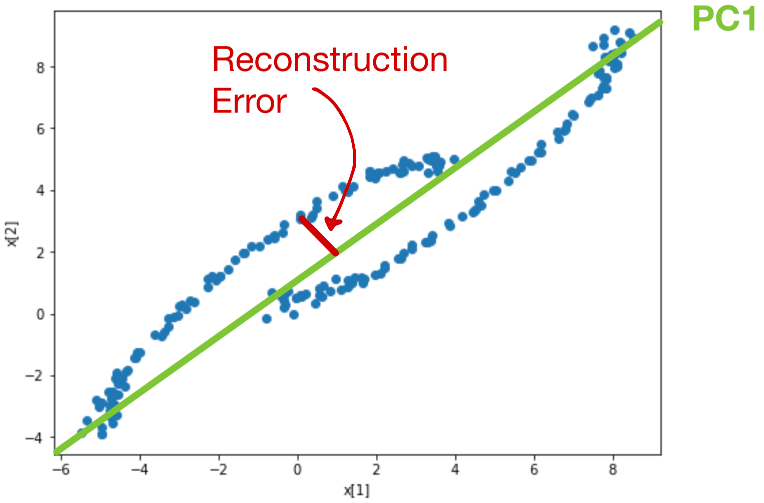
Solution: With the shapes in problem (a) and the clarification during the exam, this problem was not solvable if you did the max-stride 1. We gave everyone credit for this problem.

4. [2 points] You are training a deep neural network model and the accuracy on both the training and validation set is very low. List 2 strategies to improve it and explain why you think they will improve the model

Solution: There are many possible explanations for this problem. The key thing is, as described, an underfit model is the most likely explanation. That means any reasonable explanation to increase model complexity would be acceptable. Some common examples include 1) Increase the number of training epochs 2) increase the number of nodes/layers 3) add transfer learning 4) reduce/remove regularization 5) etc.

6 PCA [5 points]

1. [5 pts] Consider the following dataset with two inputs $x[1]$ and $x[2]$. Suppose we are interested in using PCA to reduce the dimensionality of this dataset to one dimension.



(a) [3 pts] On the graph above, draw the direction of the first principle component. Additionally, annotate your answer by visually showing the concept of "reconstruction error" for a single example. Label your answers so it is clear where the principle component is and where your example of reconstruction error is.

(a) [2 pts] In this example, we are hoping that using PCA to reduce to one dimension will maintain the fact that this data clearly falls into two distinct clusterings. Does PCA preserve these clusters present in the original data? Why or why not?

Solution: No, PCA does not preserve these clusters because there are overlapping datapoints in the two curves. In other words, when we project onto the first PC line data points from the two clusters will collide and they will no longer be in differentiable clusters

7 Recommender Systems [6 points]

1. [5 points] Suppose you are building a recommendation system to predict TikTok engagement. You have data for two users and two videos. A higher number means the user is more likely to engage with that video.

	Video 1	**Video 2**
User 1	3	?
User 2	?	2

You are running Matrix Factorization and you have obtained the following decomposition for this data in the current iteration of coordinate descent.

	User Factors
User 1	[1, 2, 1]
User 2	[3, 1, 0]

	Video Factors
Video 1	[1, 2, 0]
Video 2	[1, 0, 4]

(a) [1 pt] What is the predicted rating for User 2's engagement with Video 1?

$\hat{\boldsymbol{r}}_{\mathbf{2,1}}$ = $\mathbf{3 \cdot 1 + 1 \cdot 2 + 0 \cdot 0 = 5}$

(b) [2 pts] What is the current mean squared error loss?

MSE = $\mathbf{1/2 \cdot ((3-6)^2 + (2-3)^2) = 5}$

(c) [2 pts] If the next iteration of coordinate descent updates the user factors, which user factor(s) would change? Select all that apply.

■ **User 1**

■ **User 2**

2. [1 pt] Adding more dimensions to the user/video factors in Matrix Factorization can help solve the cold start problem.

○ True

● **False**

8 Clustering [13 points]

1. [4 pts] In the following sentence, there are options to fill in the sentence surrounded by parentheses where each option is separated by a /. For each place, choose the best option to complete the sentence by circling or underlining the option that best completes the sentence. For example: Hunter likes ([✓ dogs] / cats)!

When running k-means++ for a fixed value of k, we use a

(fixed / uniformly random / [✓ biased random]) initialization to start the algorithm

and then run repeated steps of assignment and updating until the algorithm converges.

We generally run this algorithm **(once / [✓ multiple times])** and choose the clustering

with the **([✓ lowest] / highest)** heterogeneity. The result found will

(be a local optima / [✓ be a global optima] / have no convergence guarantees)

---

2. [9 pts] Consider the following 1-dimensional dataset. Each point is labeled with a number.

- A: $x_A = 1$
- B: $x_B = 4$
- C: $x_C = 4.5$
- D: $x_D = 6$
- E: $x_E = 10$

Suppose you perform Agglomerative Hierarchical Clustering with **single linkage**.

(a) [5 pts] What is the sequence at which the points will be joined? If two points are at the same level, select them based on alphabetical order. An example answer is formatted like **ABCDE**.

BCDAE

(b) [2 pts] Suppose we decided to cut the dendrogram for the hiearchical clustering at height of 3.5 units. How many clusters will be used in that case?

2

(c) [2 pts] Name two reasons why we would use a hierarchical clustering method like Agglomerative Clustering instead of k-means.

Solution: Two possible reasons include

- Hierarchical clustering can learn non-spherical shapes
- Can use a dendrogram to investigate different deterimine of k number of clusters

9 Miscellaneous [13 points]

1. [10 pts] You have learned many different methods. We discussed how to compare the methods and evaluate the performance but sometimes we may have a good intuition which model will be a good fit. **From all the methods you have learned, which you would think is a good match for the problems below. Why?**

   This question is designed to be open-ended, but you should justify your modeling choices with specific reasons why they best fit the described task. If you think using two or more of these techniques in combination would be effective you may do so, then provide justification why using both would be good. Note that you should only list more than one choice if you have a good reason to believe that using both is better than just one. In other words, something like "I would use X and Y because I could look at both results and decide which is better" is not an appropriate answer.

   Note the you must include both a model, as well as an explanation, in your answer!

   - Linear Regression
   - Ridge Regression
   - LASSO Regression
   - Logistic Regression
   - Decision Tree
   - Random Forest
   - AdaBoost
   - k-Nearest Neighbors
   - k-means
   - Hierarchical Clustering (Agglomerative, Divisive)
   - Locality Sensitive Hashing
   - Principal Component Analysis
   - Fully Connected Neural Network
   - Convolutional Neural Network
   - Transfer Learning
   - Matrix Factorization

   (a) [2 pts] You have a weather dataset that has very few rows and many data inputs (columns). You want to predict the day's temperature (in Celsius) given other features about the weather.

Solution: We accepted any model that can do regression that was not a neural network since they tend to require much more training data.

(b) [2 pts] You have a high-dimensional dataset of the proportion of different proteins in the body of people with and without long COVID. Before developing any predictive model, you want to visualize this data (in 2D) to see if there are any obvious visual patterns in the data.

Solution: PCA allows you to to find 2-dimensions that can best represent the dataset that minimizes the reconstruction error.

(c) [2 pts] You want to predict the probability a student passes a class given how many hours they studied for an exam. Based on your domain expertise, you suspect there is a linear relationship between study time and likelihood of passing.

Solution: Logistic Regression is the model that we learned in class that estimates probabilities and assumes the likelihood is controlled by a linear function.

(d) [2 pts] You have an unlabeled image dataset containing pictures of dogs and cats, and are interested in grouping all the dog and cat images into separate groups.

Solution: k-means is the unsupervised algorithm that we learned to help cluster examples in a dataset based on similarity. Many students answered PCA which might help get better results, but this does not solve the problem described since you would still need to do something like k-means afterwards.

(e) [2 pts] You only have a little bit of image data (100 images) on chrysanthemums and roses, but you're trying to make a model that differentiates between the two. Your boss doesn't have the budget to get more flower pictures, but she does have a lot of pictures of animals and their labels.

Solution: Since you don't have enough image data to train a whole model on chrysanthemums, you'll want to use Transfer Learning. Train a CNN object detection model on the animal data, and use its convolutional layers as an input into a simpler model that you train on the flower data.

2. [3 pts] Consider the datasets used in programming assignments and sections this quarter. Write 3 ethical considerations to take into account developing machine learning models using those datasets.

Solution: Many possible answers to this question. Note that this problem was asking specifically about our homework and section problems, so an answer was complete only if it made some tie-in to the homeworks or sections from this quarter.

Extra Credit [2 pts] When discussing the advancements of generative AI, we outlined two major ideas: 1) The innovations of the introduction of Transformer models (position embedding and attention) and 2) the two phase training for models like ChatGPT.

Briefly summarize the most important ideas of these two concepts in the box below. Answers should be about 4-5 sentences long, but there is no firm limit.

Solution: All solutions were accepted for this problem as a bonus for the mistake in the Deep Learning sectionl