SAMPLE FINAL EXAM

CS 178, SPRING 2023

June 7, 2023

YOUR STUDENT ID:

YOUR NAME:

YOUR SEAT NUMBER:

Instructions

- DO NOT GO BEYOND THIS PAGE TO START UNTIL INSTRUCTED TO DO SO.
- Please write your student ID, full name, and seat number in the spaces provided above.
- Put away everything (including any electronic devices, written materials, etc) except for a pen or pencil.
- If you think there is a typo in a question, please read the question carefully at least twice. If you still think there is a typo please raise your hand and we will respond.
- The exam will last for 2 hours. There are 5 questions on the exam, all worth equal points.
- If you need to use a rest-room please raise your hand and let one of the TAs know. You will need to drop off your cellphone at the front of the class before you leave: you can pick it up after the exam is over.
- When instructed that the exam is over, put down your pen/pencil and stop writing. Remain in your seat and we will collect exams row by row.
- If you finish with more than 10 minutes of the exam remaining you can leave your seat and drop your exam off at the front of the lecture hall. Otherwise wait in your seat until we collect the exams.

Feel free to use this page for your rough work: it will be ignored during grading

Feel free to use this page for your rough work: it will be ignored during grading

Problem 1

State TRUE or FALSE to indicate if each of the following statements are true or false:

1. The kNN classifier can learn non-linear decision boundaries:
   False: kNN has piece-wise linear boundaries.

2. The sigmoid function should be used in the final layer of a neural network for multi-class classification problems:
   False: sigmoid is used for binary classification problems; softmax is used for multi-class problems.

3. Regression models can be trained by minimizing the mean squared error loss function:
   True

4. Random forest models are fit by only considering a subset of the features at each split:
   True

5. For any given classification problem, the optimal decision tree (achieving zero error rate) is unique:
   False: there are often many different decision trees which will achieve perfect accuracy on the training data.

6. A drawback of the decision tree model is that it can only learn decision boundaries which are parallel to the axes in data space:
   True

7. A major benefit of decision trees is that the predictions they make are often easy to interpret:
   True

8. Ensemble methods can only be used if each classifier in the ensemble is of the same type, e.g. all neural networks:
   False: you can ensemble models of different types, e.g. stacking a neural network and a logistic regression model.

9. Principal components are always directions that are perpendicular to one another.:
   True

10. Reinforcement learning is an example of a supervised learning problem, where the targets are the rewards:
    False

11. If the transition probabilities in a Markov process are unknown, one strategy for estimating these probabilities is to simulate episodes from the Markov process: **True**

Problem 2

1. For each classifier listed in each row below, indicate what type of decision boundary the classifier produces in the input space $\mathbf{x}$ for the two class case of $K = 2$. Your answer for each classifier should be one of { *linear, piecewise-linear, non-linear* }. For kNN, k is the number of neighbors used in the classifier.

   (a) kNN, $k = 1$: piecewise-linear

   (b) kNN, $k > 1$: piecewise-linear

   (c) Nearest Centroid: linear

   (d) Logistic Classifier: linear

   (e) Feedforward Neural Network with sigmoid hidden units: non-linear

   (f) Classification Tree, depth $= 1$: linear

   (g) Classification Tree, depth > 1: piecewise-linear

   (h) Random Forest: piecewise-linear

2. State ”yes” or ”no” to indicate which of the following are convex optimization problems. For the neural networks assume that the networks have a single hidden layer and are using a sigmoid (logistic) activation function.

   (a) Log-loss with a logistic classifier: yes

   (b) Log-loss with a feedforward neural network classifier: no

   (c) MSE-loss with a linear regression model: yes

   (d) MSE-loss with a polynomial regression model: yes

   (e) MSE-loss with a feedforward neural network regression model: no

Problem 2: continued

3. For each of the classifiers below assume that the classifier has already been trained. Now assume that we need to store the classifier in memory to make predictions, e.g., store an image classifier on a mobile phone. For each classifier below, state how many numbers (real-valued or integer) need to be stored in memory for the classifier, as a function of K classes, d features, and n training examples, and/or any other variables mentioned in the description of the classifiers below.

   You can assume for simplicity that there are no bias/intercept parameters in the logistic and neural network models.

   As an example, for a logistic classifier with $K = 2$ classes the correct answer is d (the d weights of the classifier).

   (a) Nearest Centroid classifier: Kd

   (b) kNN classifier with $k = 1$: $n(d + 1)$

   (c) kNN classifier with $k > 1$: $n(d + 1)$

   (d) Logistic classifier with $K > 2$ classes: Kd

   (e) Neural network with with $K > 2$ classes and a single hidden layer with H hidden units: $dH + HK$

   (f) Voting ensemble of M neural networks, where each network has p parameters: Mp

   (g) Stacked ensemble of M neural networks, where each network has p parameters: $Mp + M$

Problem 3: Logistic Classifier

Assume in the problem below (unless otherwise stated) that we are working with a binary classification problem with class labels $y = 0$ and $y = 1$ and a d-dimensional feature vector $\mathbf{x}$.

The logistic classifier is defined by the following equations, as discussed in class:

$$z(\mathbf{x}; \boldsymbol{\theta}) = \theta_0 + \sum_{j=1}^{d} \theta_j x_j$$

and

$$f(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{1 + e^{-z(\mathbf{x}; \boldsymbol{\theta})}}$$

1. **3 points:** What range of values can $f(\mathbf{x}; \boldsymbol{\theta})$ take? Your answer should be in the form of (a, b) where a and b are numbers.

   $(a, b) \quad = (0, 1)$

2. **3 points:** In the logistic classifier, for binary classification, we predict class 1 or 0 depending on whether or not $f(\mathbf{x}; \boldsymbol{\theta}) > t$ where t is some threshold: what value of t is used with the logistic classifier?

   $t \quad = 0.5$

Problem 3 (continued)

3. **6 points:** For the case of two features ($d = 2$), we have $z(\mathbf{x}; \boldsymbol{\theta}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$. Show an equation that defines the decision boundary in two dimensions: this should be an equation involving x_1 and x_2.

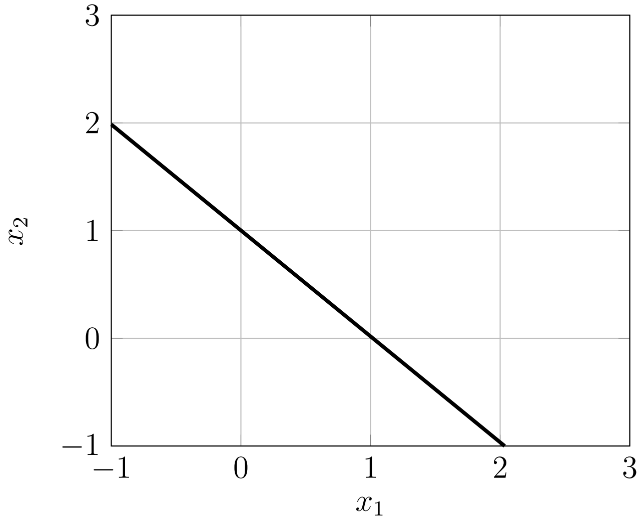
   The decision boundary is defined by

   $$\theta_0 + \theta_1 x_1 + \theta_2 x_2 = 0.$$

   Solve for x_2:

   $$x_2 = -\frac{1}{\theta_2}(\theta_1 x_1 + \theta_0).$$

4. **4 points:** For $\theta_0 = -1, \theta_1 = 1, \theta_2 = 1$, sketch on the given 2d plot (where x_1 is the horizontal axis and x_2 is the vertical axis) where the decision boundary is. The horizontal and vertical axes should each range from -1 to 3.



5. **4 points:** As discussed in class, we can extend the logistic classifier from 2 classes to K classes. How many parameters in total are in the K-class logistic model?

   There are d weights per class, and K bias terms. Hence, there are $dK + K = (d+1)K$ parameters in total.

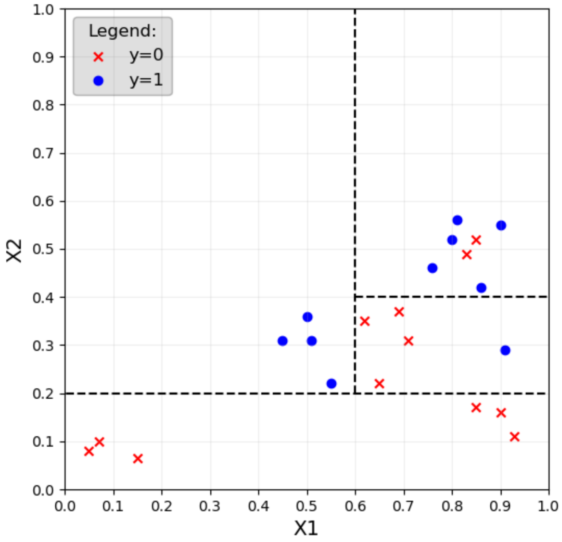
   Alternatively, the weights for the last class are redundant, so

   $$d(K-1) + (K-1) = (d+1)(K-1)$$

   is also a valid answer.

Problem 4: Decision Trees

1. Consider the following decision boundaries produced by a decision tree fit to a binary classification problem with features X_1 and X_2 and labels $y = 0$ and $y = 1$. Dashed black lines depict the splits this tree made after training on the given data.



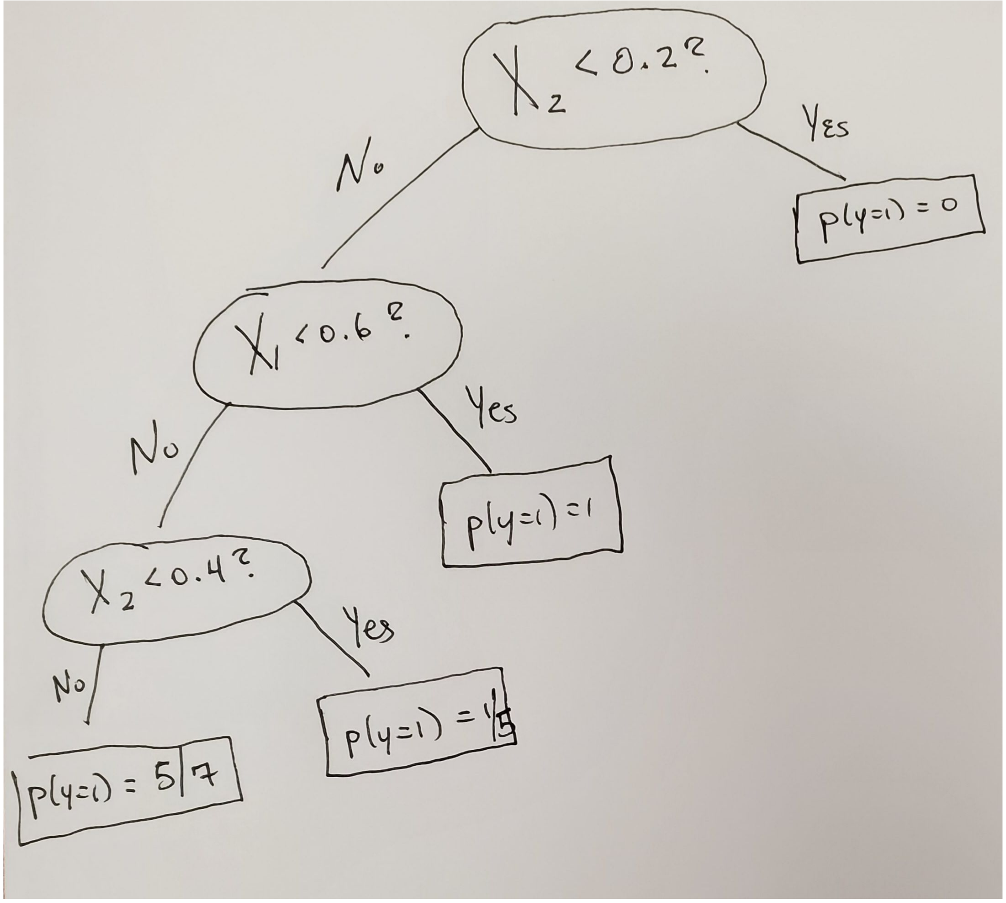
   (a) What is the error rate for the tree shown evaluated on the presented data?

   The error rate is $3/22$.

   (b) If you could allow the tree to introduce one more split, which sector would be the best candidate? Furthermore, which variable would be split and at roughly what threshold value?

   You should split on $X1$ at a threshold of around 0.8. This is because this is the only place you could further split the data (with just a single split) that would reduce the Gini index.

   (c) Draw the decision tree corresponding to the splits in the image. For every leaf node, you should also include $p(y = 1 \mid \text{path})$.


X₂ < 0.2?
No
Yes
p(y=1) = 0
X₁ < 0.6?
No
Yes
p(y=1) = 1
X₂ < 0.4?
No
Yes
p(y=1) = 5/7
p(y=1) = 1/5

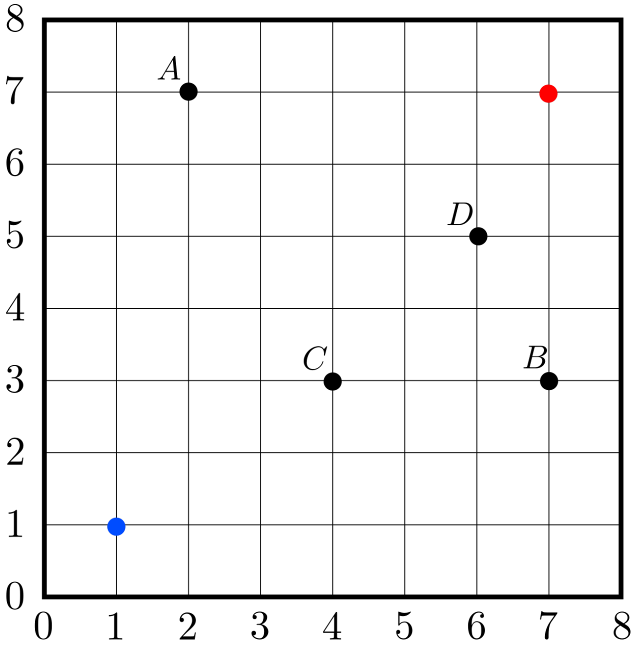
Problem 5: k-Means Clustering

Below are three different 2-dimensional datasets in the process of being clustered using the k-means algorithm with $k = 2$. In each scenario, the dataset consists of four points, $\{A, B, C, D\}$, along with two centroids shown in red and blue. For each scenario, your job is to determine three things:

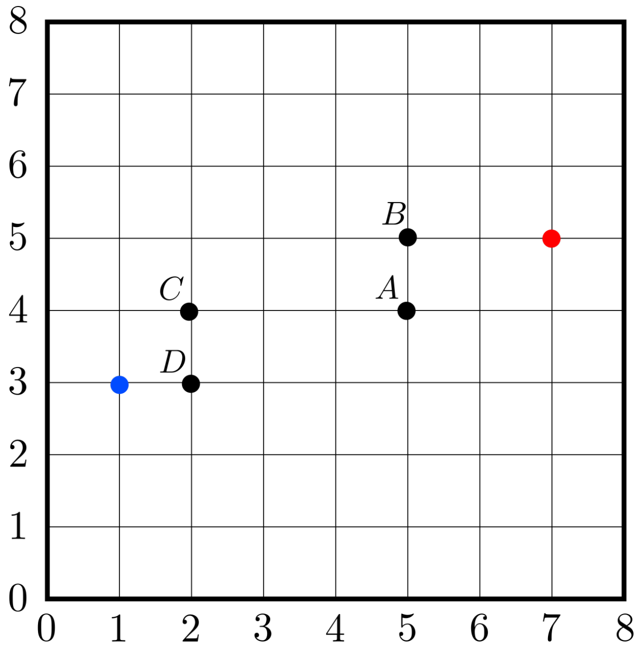
1. For the next iteration of k-means, what cluster will each point be assigned to, red or blue?

2. Draw on the plot where the next red and blue centroids will be using the previously determined assignments.

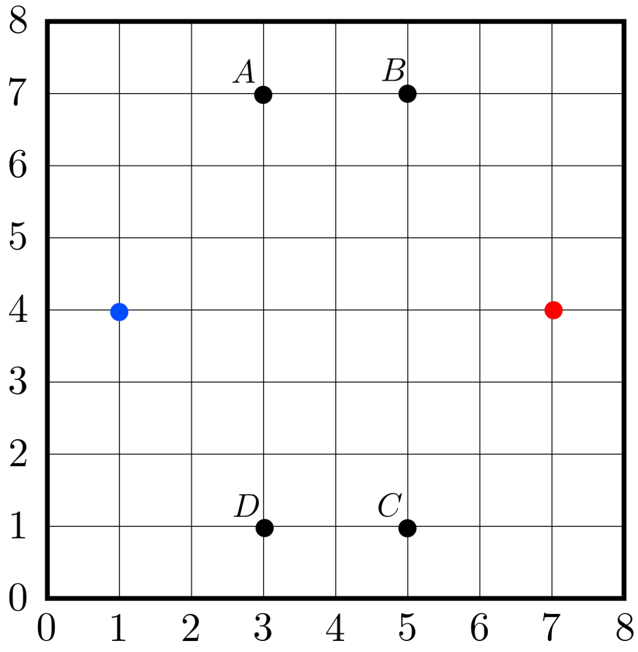
3. Finally, will the algorithm have converged with these new centroids?



New cluster assignments, red or blue $A =$ Red $B =$ Red $C =$ Blue $D =$ Red
New centroid locations: red $= (5, 5)$ blue $= (4, 3)$
Converged? Yes



New cluster assignments, red or blue A = Red B = Red C = Blue D = Blue
New centroid locations: red = $(5, 4.5)$ blue = $(2, 3.5)$
Converged? Yes



New cluster assignments, red or blue A = Blue B = Red C = Red D = Blue
New centroid locations: red = $(5, 4)$ blue = $(3, 4)$
Converged? Yes