Identikey: _____

Artificial Intelligence (F19)
Quiz 3 (QZ3)

First Name: _____

Overview
7.5 Point Quiz on Search (7.5% of final grade)
30 minute closed book quiz

Last Name: _____

Learning Objective
This assignment satisfies learning objective 1 (LO1) as specified in the syllabus. You will demonstrate conceptual understanding of the core AI topics.

The quiz is worth 7.5 points total but there are 8 points available. This means that you can lose 0.5 points, but still earn a perfect score of 7.5 points on the quiz. If you score above 7.5 points, your total score will be rounded down to 7.5 points.

1) **Short answer/multiple choice questions**
   i) [Points: 0.25] Consider the problem of determining the price of a house in Boulder based on carpet area, location, number of rooms, and age. What kind of problem is this?
      i) Unsupervised learning problem
      ii) Classification problem
      iii) Regression problem
      iv) None of the above

      Solution: iii) Regression Problem

   ii) [Points: 0.25] What precise accuracy metric would you use to evaluate a model trained to solve the above problem?

      Solution: R-Square measure/coefficient of determination is a good metric to evaluate such models.

   iii) [Points: 0.25] When is accuracy (percent correct) an acceptable metric for a classification problem?

      Solution: When the classes are approximately balanced.

   iv) [Points: 0.25] What is the role of the regularization (i.e., C) parameter in a support vector machine?
      a) It controls the behavior of the kernel
      b) It controls the width of the margin
      c) Both (a) and (b)
      d) Neither (a) nor (b)

      Solution: b) It controls the width of the margin

   v) [Points: 0.25] How does an SVM solve a difficult linearly inseparable problem?

      Solution: By projecting the data to a higher dimensional space where it may be linearly separable. Note: adding slack variables doesn't help with severe linear inseparability

vi) [Points: 0.25] Consider a classifier that randomly assigns instances to a positive or negative class. What is a likely AUROC for this classifier?

i) -0.5
ii) 0.01
iii) 0.52
iv) 0.98

Solution: iii) 0.52

2) [Points: 0.75] A decision tree of depth 6 has achieved 100% accuracy on a data set. Which of the following **two statements** are accurate?

i) The tree at depth 4 will have higher bias than the tree at depth 6
ii) The tree at depth 4 will have lower bias that the tree at depth 6
iii) The tree at depth 4 will have higher variance than the tree at depth 6
iv) The tree at depth 4 will have lower variance than the tree at depth 6

Justify your response:

Solution: i) and iv)

Justification:

i) The tree at depth 6 would consider more features than the tree at depth 4, and hence it would have more information in making decisions and would be less biased.

iv) In this case, it has 0 bias. However, it is likely overfitting to the data due to the fact that accuracy is perfect. Hence, the tree at depth 4 will have lower variance.

3) [Points: 0.75] What is the entropy of a feature with the following eight values: 0,0,0,1,1,1,1,1

$$H(S) = -\sum_{x \in X} p(x) \log_2 p(x)$$

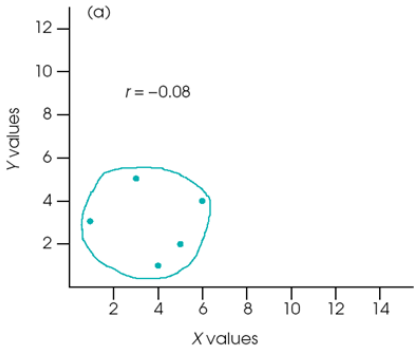
A) -5/8 log(5/8) - 3/8 log(3/8)
B) 7/8 log(7/8) + 3/8 log(3/8)
C) 3/8 log(5/8) + 5/8 log(3/8)
D) -5/8 log(3/8) – 3/8 log(5/8)

Solution: A) -5/8 log(5/8) - 3/8 log(3/8)

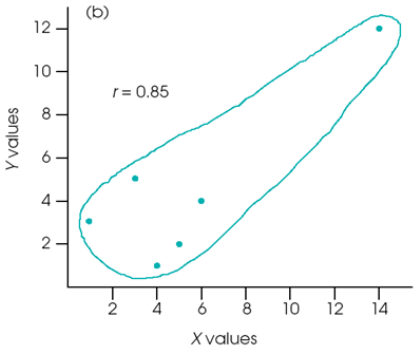
Identikey: _____

4) [Points: 0.75] Linear regression is said to be influenced by outliers. Graphically (draw a sketch) to illustrate this problem.

Solution:

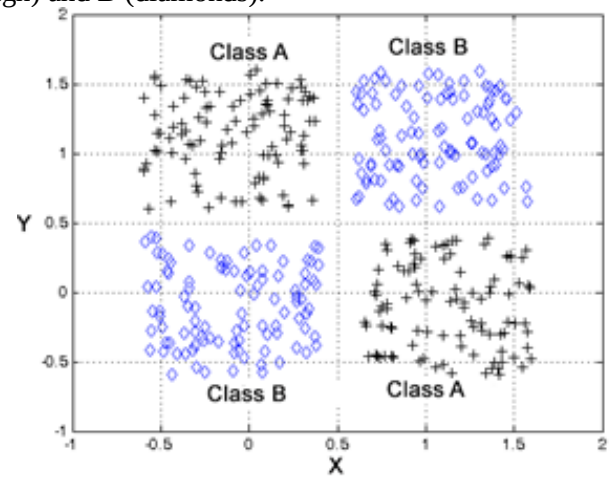


Original Data		
SUBJECT	X	Y
A	1	3
B	3	5
C	6	4
D	4	1
E	5	2



Data with Outlier Included		
SUBJECT	X	Y
A	1	3
B	3	5
C	6	4
D	4	1
E	5	2
F	14	12

5) [Total points: 1.25] Consider the following plot of a dataset with two features X and Y used to classify A (plus sign) and B (diamonds).

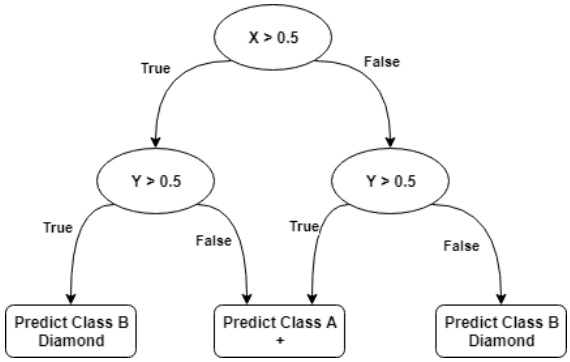


a) [Points: 0.5] Is logistic regression an appropriate classifier for this dataset? Why or why not?

Solution: We cannot use logistic regression for this data because it is not linearly separable.

b) [Points: 0.75] Sketch a decision tree for the dataset above. You need not compute exact values – we are just asking for a rough sketch of the tree.

Solution:



7) [Points: 1] You train a logistic regression classifier to classify an email as spam or not spam and are considering two classification thresholds. Threshold A yields a precision of 80% and recall of 20% whereas Threshold B yields a precision of 20% and recall of 80%. Which one would you prefer in this case? *You need to justify your response to get full credit and a good justification should communicate your assumption, an understanding of precision, recall, and the tradeoff among the two.*

Solution:
Either option is correct as long as it is justified using appropriate technical language.
Example: We would prefer Threshold A since in the case of email spam we want to reduce the number of non-spam emails that are labeled as spam. Higher precision means the fraction of emails classified as spam that were actually spam is higher. A higher recall would indicate that a larger fraction of all possible spam emails were classified as spam, but that's not ideal in this case since that would cause more non-spam emails to be labeled as spam.

8) [Points: 1] The data set below represents balls used for different games. List which instances would be used for testing and training in a 3-fold cross validation. You can assume that instances 1, 2, and 3 are assigned to fold 1, instances 4, 5, and 6 to fold 2, and instances 6, 7, and 9 to fold 3.

Instance No.	**Color**	**Diameter**	**Material**	**Game**
1	orange	12"	leather	basketball
2	black	10"	iron	quiditch
3	red	12"	leather	quiditch
4	cream	10"	leather	football
5	cream	2"	leather	baseball
6	red	10"	leather	basketball
7	brown	10"	leather	football

8	gold	1”	wooden	quiditch
9	red	2”	leather	baseball

Solution:
Training [1-6] Testing [7-9]
Training [1-3, 7-9] Testing [4-6]
Training [4-9] Testing [1-3]

9) [Points: 1] Consider the “House Inspection” dataset, which determines whether a house is acceptable or not (Acceptable column) based on Furniture, # Rooms, and Kitchen. Using a Naïve Bayes classifier, write out the equation for the probability that a house without furniture, 4 rooms, and a new kitchen will be deemed acceptable. *Note – we are only asking for the equation. You do not have to plug in any numbers or compute anything.*

House	**Furniture**	**# Rooms**	**Kitchen**	**Acceptable**
1	Not Included	3	New	Yes
2	Included	3	Old	No
3	Not Included	4	Old	Yes
4	Not Included	3	Old	No
5	Included	4	Old	Yes

The following equation should help you out. *Hint. We are mainly interested in the numerator.*

$$P(h \mid E) = \frac{P(E \mid h)P(h)}{P(E)}$$

Solution:

$$\mathrm{P}(\mathrm{Yes} \mid \{\mathrm{Not\ Included}, 4\ \mathrm{rooms}, \mathrm{New\ Kitchen}\}) = \frac{P\ (Not\ Included \mid Yes)\ *\ P\ (4\ Rooms \mid Yes)\ *\ P\ (New\ Kitchen \mid Yes)\ *\ P(Yes)}{P(Not\ Included,\ 4\ Rooms, New\ Kitchen)\ or\ P(E)}$$