CalPolyPomona

CS 4210 – Assignment #1

Maximum Points: 100 pts.

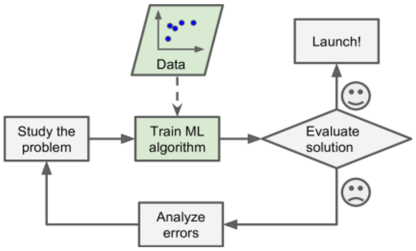
Bronco ID: |__|__|__|__|__|__|__|__|__|

Last Name: ______________________

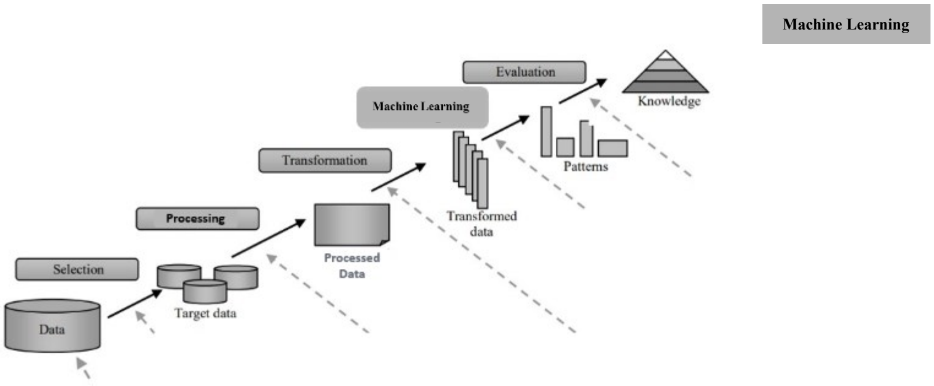
First Name: ______________________

Note 1: Your submission header must have the format as shown in the above-enclosed rounded rectangle.
Note 2: Homework is to be done individually. You may discuss the homework problems with your fellow students, but you are NOT allowed to copy – either in part or in whole – anyone else's answers.
Note 3: Your deliverable should be a .pdf file submitted through Gradescope until the deadline. Do not forget to assign a page to each of your answers when making a submission. In addition, source code (.py files) should be added to an online repository (e.g., github) to be downloaded and executed later.
Note 4: All submitted materials must be legible. Figures/diagrams must have good quality.
Note 5: Please use and check the Canvas discussion for further instructions, questions, answers, and hints. The bold words/sentences provide information for a complete or accurate answer.

1. [6 points] A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E (Mitchell, 1997). Explain this definition of a machine learning system informing in your answer how **E, T, P correlate** with **each component** of the image below.

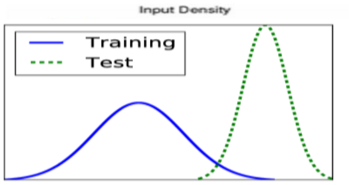


2. [6 points] Some authors present a KDD/Data Mining pipeline process with only 3 main phases instead of those 6 shown in the image below (see the dashed arrows). **Name** those 3 main phases and **explain** their corresponding relevance to building knowledge.

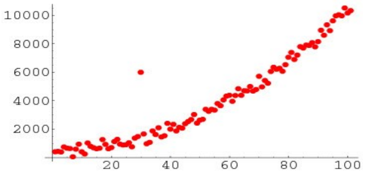


3. [15 points – 3 points each] Machine learning algorithms face multiple challenges while analyzing data such as scalability, distribution, sparsity, resolution, class imbalance, noise, outliers, missing values, and duplicated data. For **each** image below, **name** and **explain** what the corresponding challenge is from this list (you do not need to explain how to solve the challenge).

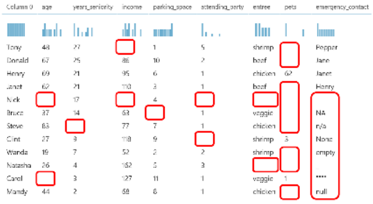
a.



b.

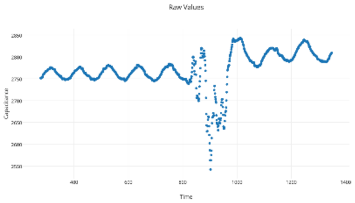


c.

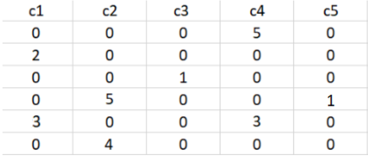


Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
Tony	48	27		1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick		17		4				
Bruce	37	14	63		1	veggie		NA
Steve	63		77	7	1	chicken		n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp		empty
Natasha	26	4	152	5	3			
Carol		3	127	11	1	veggie	1	****
Mandy	44	2	68	8	1	chicken		null

d.

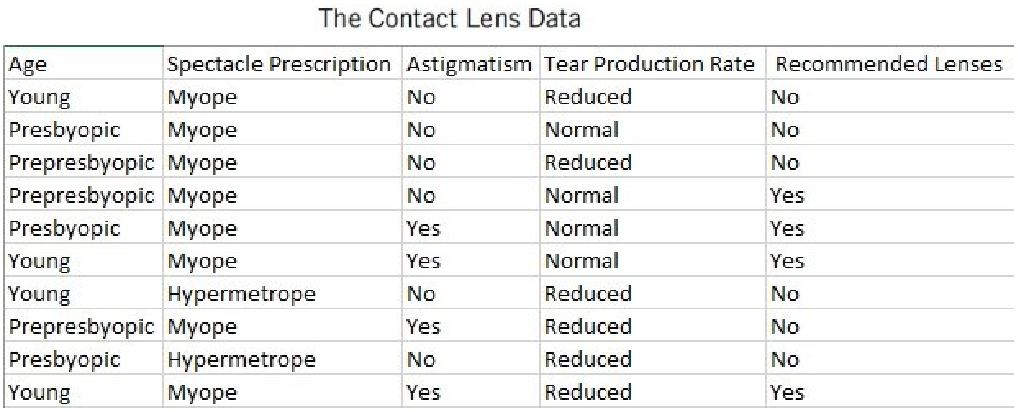


e.



c1	c2	c3	c4	c5
0	0	0	5	0
2	0	0	0	0
0	0	1	0	0
0	5	0	0	1
3	0	0	3	0
0	4	0	0	0

4. [18 points – 3 points each] Analyze the dataset below and answer the proposed questions:



The Contact Lens Data

Age	Spectacle Prescription	Astigmatism	Tear Production Rate	Recommended Lenses
Young	Myope	No	Reduced	No
Presbyopic	Myope	No	Normal	No
Prepresbyopic	Myope	No	Reduced	No
Prepresbyopic	Myope	No	Normal	Yes
Presbyopic	Myope	Yes	Normal	Yes
Young	Myope	Yes	Normal	Yes
Young	Hypermetrope	No	Reduced	No
Prepresbyopic	Myope	Yes	Reduced	No
Presbyopic	Hypermetrope	No	Reduced	No
Young	Myope	Yes	Reduced	Yes

a. What is the most likely task that data scientists are trying to accomplish?

b. **In general**, what is a feature, and how would you **exemplify** it with **this data**?

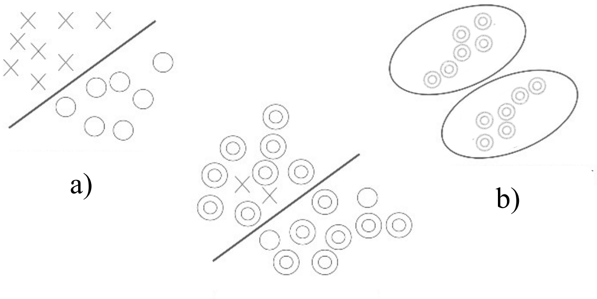
c. **In general**, what is a feature value, and how would you **exemplify** it with **this data**?

d. **In general**, what is dimensionality, and how would you **exemplify** it with **this data**?

e. **In general**, what is an instance, and how would you **exemplify** it with **this data**?

f. **In general**, what is a class, and how would you **exemplify** it with **this data**?

5. [9 points] Identify and explain what **kind of machine learning** (supervised, unsupervised, semi-supervised, reinforcement) **system** should be used for each scenario below including in your answer information about **data labels**. Hint: check the images to figure out which data sample is labelled.

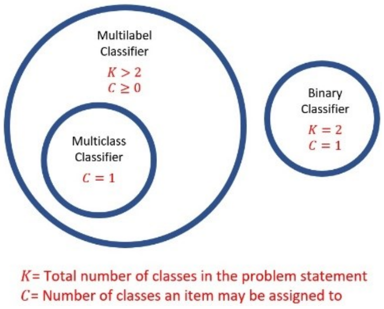


a)

b)

c)

6. [9 points] Explain the **tasks** addressed by each classifier below. K and C **must be present** on your answer.


Multilabel Classifier
$K > 2$
$C \geq 0$
Multiclass Classifier
$C = 1$
Binary Classifier
$K = 2$
$C = 1$
K= Total number of classes in the problem statement
C= Number of classes an item may be assigned to

7. [37 points] Regarding the training data shown in question 4:

   a. [20 points] Derive the decision tree produced by the standard ID3 algorithm. Show your calculations for **entropy** and **information gain** for **all** splits. **Plot** your final tree at the end.

   b. [15 points] Complete the given python program (decision_tree.py) that will read the file contact_lens.csv and output a decision tree. Add the link to the online repository as the answer to this question.

   c. [2 points] The tree you got in part b) should be the same one you got in part a), but there are probably some differences. Try to explain why.

Important Note: Answers to all questions should be written clearly, concisely, and unmistakably delineated. You may resubmit multiple times until the deadline (the last submission will be considered).

NO LATE ASSIGNMENTS WILL BE ACCEPTED. ALWAYS SUBMIT WHATEVER YOU HAVE COMPLETED FOR PARTIAL CREDIT BEFORE THE DEADLINE!