CalPolyPomona

CS 4210 – Assignment #1

Maximum Points: 100 pts.

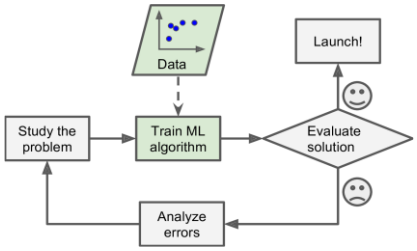
Bronco ID: |0|1|3|8|9|0|8|5|1|

Last Name: Manam

First Name: Viswadeep

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.

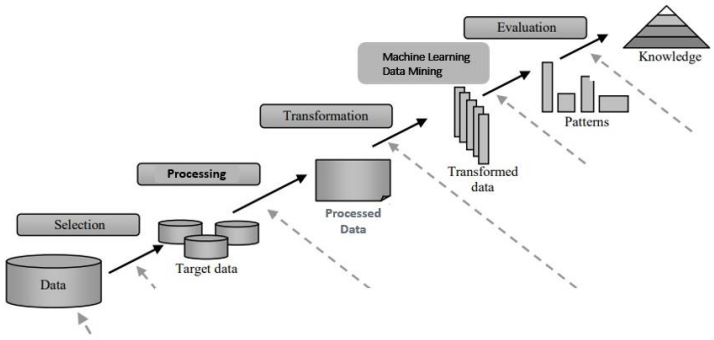


T is the actual problem the program is trying to solve, so in this case it will relate to the "Study the Problem" component.

E relates to how the program uses data to recognize patterns in that data in an attempt to potentially predict or make assumptions on new instances. As a result, this corresponds to the "Data" and "Trian ML algorithm" components.

P is related to how the program is performing or how accurately it is making these assumptions. This relates to the "Evaluate solution" stage as it is in this stage we measure this accuracy of our program. If we are unhappy with its accuracy as measured by P, then we restart this process of analyzing our T, acquiring new data, etc. If not, we launch our program.

2. [6 points] Some authors present a machine learning/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 build knowledge.

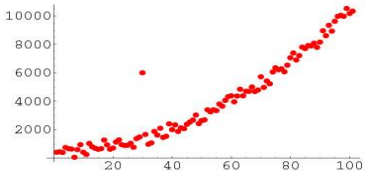


The first phase is preprocessing. In this stage, we clean and potentially reduce the dimensions of our dataset so that our model can much easily recognize patterns in our dataset. The second phase is actually building our model or using data mining techniques to get information on our model. The last main phase is postprocessing. In this stage, we might visualize our results and interpret our gleaned information.

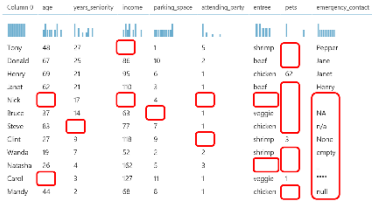
3. [15 points – 3 points each] Machine learning algorithms face multiple challenges while analyzing data such as scalability, distribution, sparsity, resolution, class imbalance, noise, outliers, missing values, and duplicated data. For **each** image below, **name** and **explain** what the corresponding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

a.

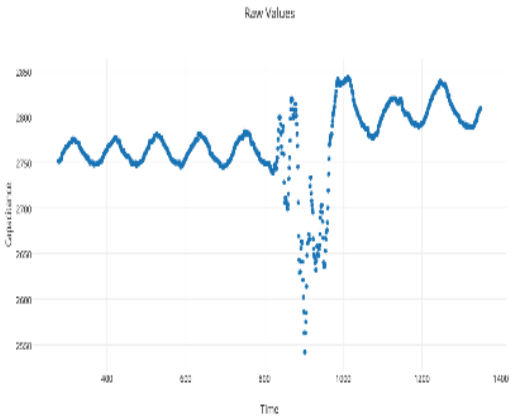
b.



c.



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

a. Distribution. The training data instances have a different distribution than the testing data instances, in the sense that there are two different patterns in the two data sets.

b. Outliers. There is a measured datapoint that is outside the variation of the overall population.
c. Missing Values. There are some missing values in the dataset.
d. Resolution. From time t=0 to 820, the data is measured properly in that there is continuous data. However, from t=820 to t=1000, the data is not measured continuously reducing the resolution during that time period.
e. Sparsity. There are a lot of zeroes in the dataset.

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?

Whether or not to recommend contact lenses to people based on their physical attributes.

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

A feature is an attribute of a data instance. In this case, it would be Age, Spectacle Prescription, Astigmatism, and Tear Production Rate.

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

A feature value are all the possible values an attribute in a data instance can have. For example in this dataset, the possible values for feature Age are {Young, Presbyopic, Prepresbyopic}, Spectacle Prescription are {Myope, Hypermetrope}, Astigmatisim are {No, Yes}, etc.

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

The dimensionality is equal to the cardinality of our feature set. In this case, it would be 4.

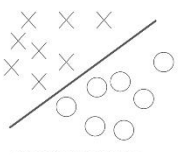
e. **In general**, what is an instance and how would you **exemplify** it with **this data**?

An instance is essentially a recorded observation of each feature and class. In this case, an instance would correspond to each row in the dataset.

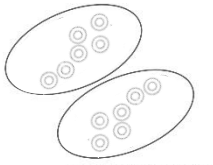
f. **In general**, what is a class and how would you **exemplify** it with **this data**?

A class corresponds to the outputs. So, in this case it would be equal to Recommended Lenses. Or, more specifically, the class of a data instance could correspond to either Yes or No.

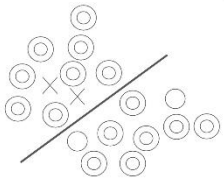
5. [9 points] Identify and explain what **kind of machine learning** (supervised, unsupervised, semi-supervised, reinforcement) **system** should be used for each scenario below including in your answer information about **data labels**. Hint: check the images to figure out which data sample is labelled.



a)



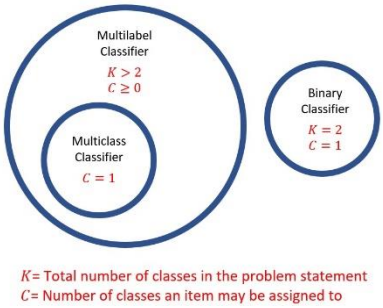
b)



c)

   a. Supervised. When a new data instance is introduced, it is easy to determine which group it belongs to by observing its features as there's a clear separation between the two output classes.
   b. Semi-Subervised. In the two groups, we know the data labels for some of them which allows us to make assumptions of the two groups these labels each belong to.
   c. Unsupervised. We do not know the labels of the data instances. As a result, all we could do is group them based on their features.

6. [9 points] Explain the **tasks** addressed by each classifier below.



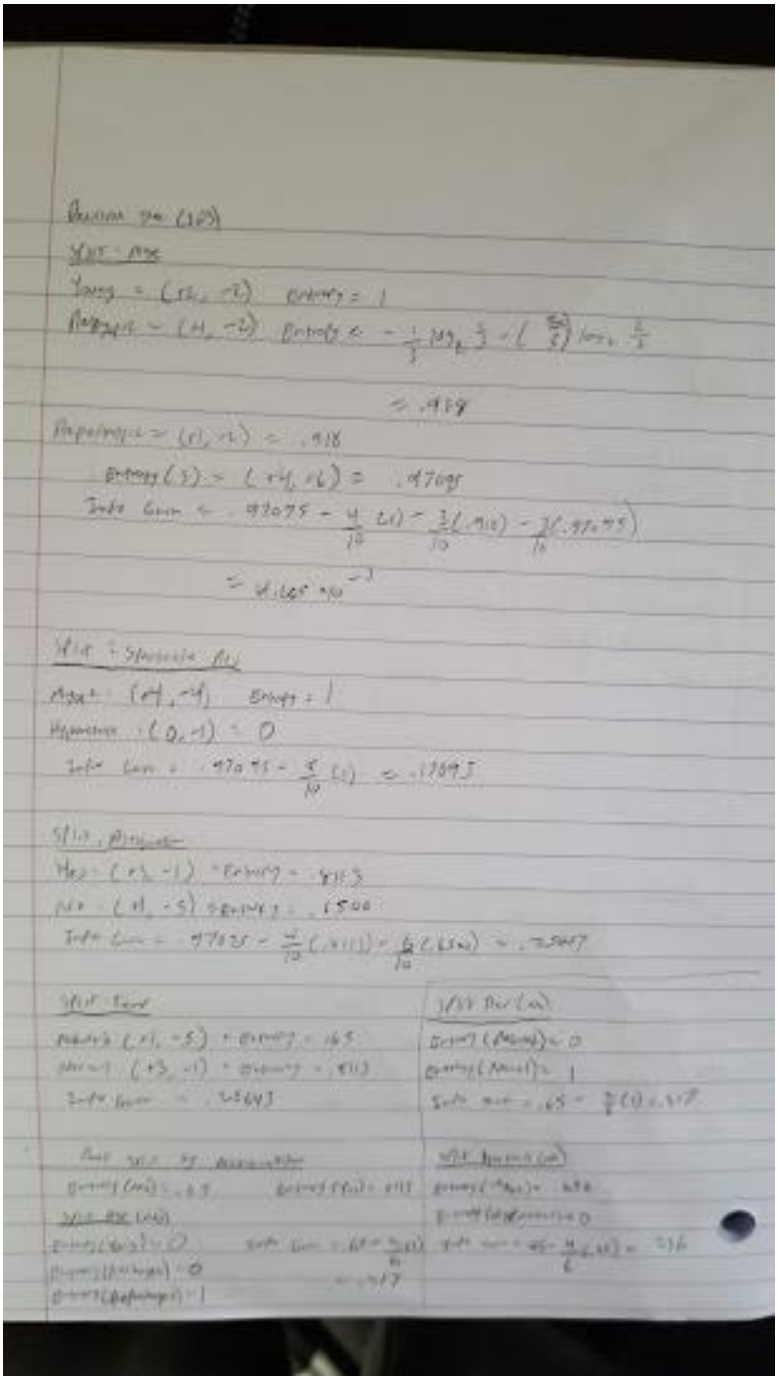
Binary classifier. In this classifier, we have two classes and each data instance could belong to only one of them. For example, a class could be is it Hot or Cold, and each of our data instances could correspond to only one of these two classes.

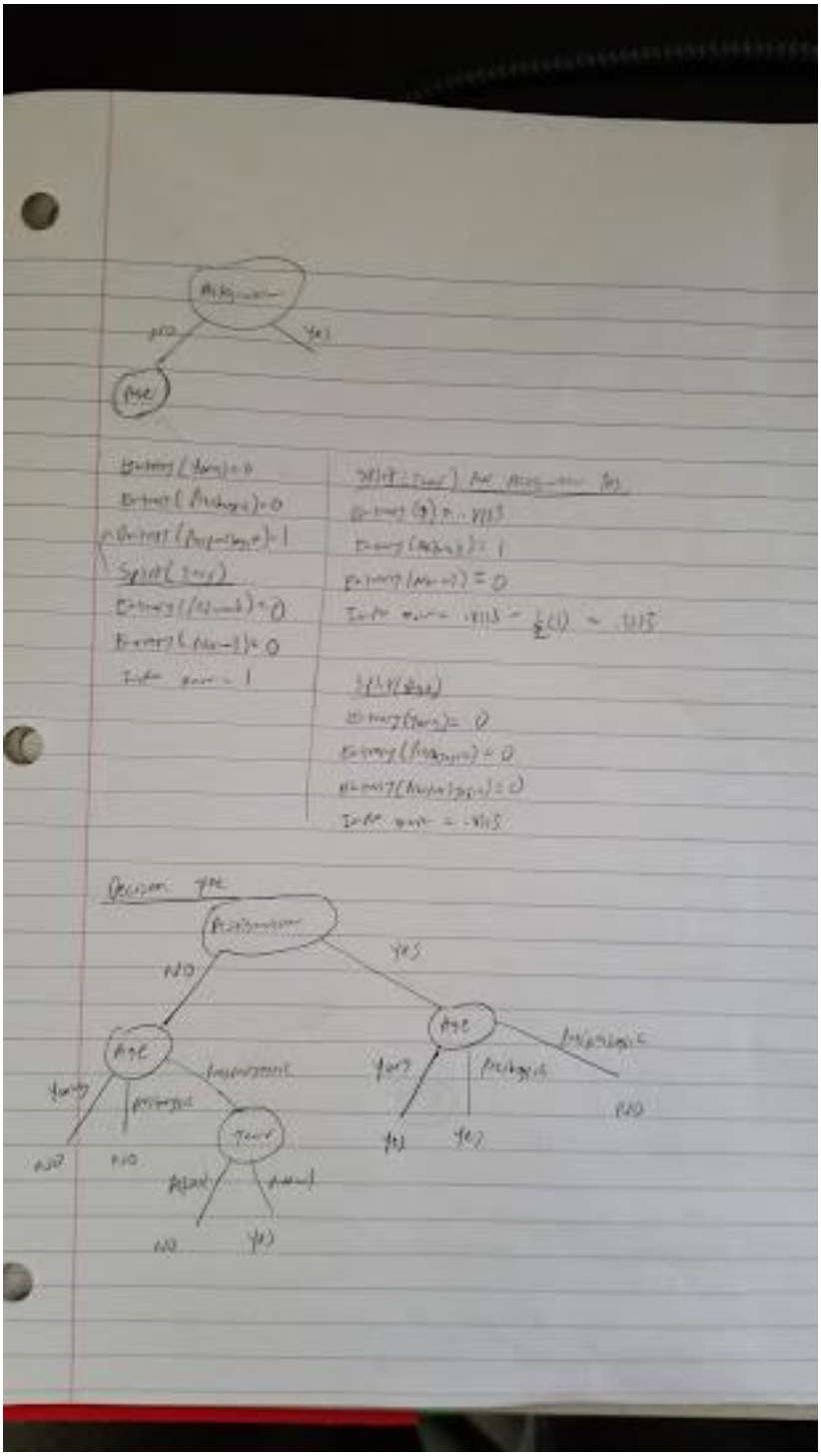
Multilabel classifier. In this classifier, we have more than two classes and each data instance could belong to any number of these classes. For example, we can use a classifier predict the fruits and vegetables used inside a smoothie. In this case, we have many classes {Apple, Grapes, Carrots, etc.} and our data instances could belong to any number of these classes. A juice could be made using {Apple, Grapes} or {Apple} or {Oranges, Carrots}.

Multiclass classifier. In this classifier, we have more than two classes and each data instance could belong to only class. If we assume a juice could be made using only one fruit or vegetable, then we would be using this classifier.

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.

https://github.com/SeveralCube22/CS4210_Assignments/tree/master/Assignment%201

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.

One of the reasons as to why the decision trees could be different is that there multiple times where the information gain between two attributes are the same. As a result, I chose one attribute to made root where the program chose another leading to different trees.

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!