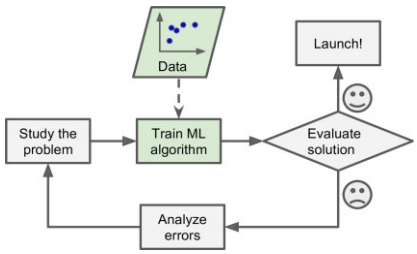
CalPolyPomona

CS 4210 – Assignment #1 Completed
Maximum Points: 100 pts.

Bronco ID: 013141414
Last Name: Mehrani
First Name: Arsham

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.

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 including in your answer details about **E, T, P** and how you **correlate** them with the **components** of the image below.



Machine Learning is defined as a programs ability to learn without being explicitly programmed. This is done through experience/training and is measured by a performance test designed specific to the algorithm.

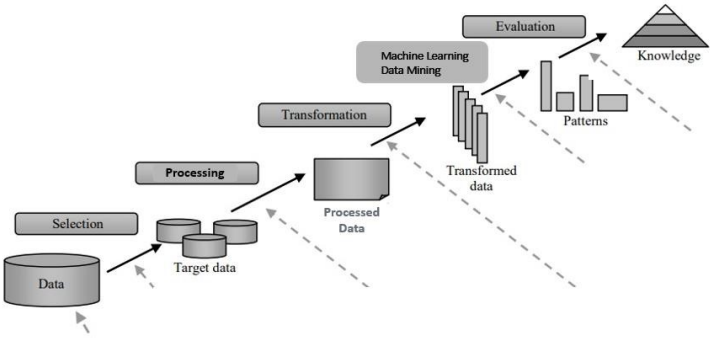
T = Task that needs to be done using Machine Learning

E = The training/experience data that the algorithm is going to use to improve

P = Performance testing using a portion of the same data set.

Study the problem: in this step the task T is analyzed to build solution. *Train ML algorithm* in this step the experience E is introduced to the model for training purposes. *Evaluate solution,* in this step the performance P is measured using the testing data set. *Analyze errors* This is part of the Task T stage, since in this step The algorithm is analyzed for errors to find the root of the problem. Once the problems are identified in last step the cycle continues to study these errors and training the model further. Once the performance P is satisfactory the algorithm is ready to launch.

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). **Explain** the reasons why they probably decided to do that including in your answer **what** are those 3 main phases and their **corresponding relevance** to build knowledge.



The pipeline process is summarized into 3 main steps as described below. This is because steps 1-4 all have to do with the preparation of data to well suit the algorithm.

Steps:

1. *Preprocessing(1-4):*

   This is probably the most time-consuming part of the ML process. The data selection is key to a good result. If the data is not optimized (i.e. features selection, dimensionality reduction, etc.) then the results that come out are prone to errors. Therefore, a clear data set that is optimized for a certain ML algorithm is essential to accurate results.

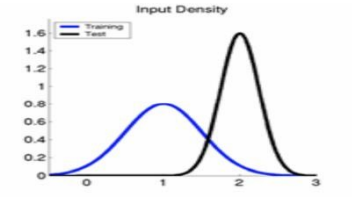
2. *Machine Learning(5):*

   This step is where the ML algorithm is written to take in the training data. This algorithm needs to be able to use the processed data and make improvements to its output over time. The data produced by the end of this step will be used to build knowledge.

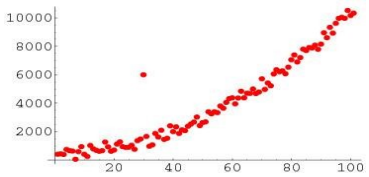
3. *Postprocessing(6):*
   This step is where the data gathered from the ML algorithm is visualized and optimized for better understanding of the results. i.e. making graphs and models that easily describe the experiment. Other devices of learning such as logic (ex. If A then B, C is direct cause of D) are used as well to better paint the picture.

3. [10 points] Machine learning algorithms face multiple challenges while analyzing data, such as scalability, dimensionality, data distribution, sparsity, resolution, 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).

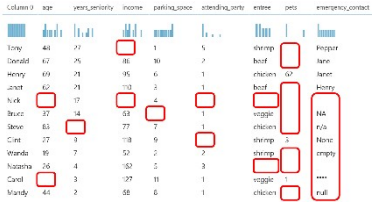
a.



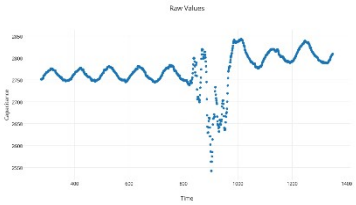
b.



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. *Data Distribution.* In this graph we see two distributions for test and training that have zero to no correlation.
b. *Outlier.* This graph demonstrates all the points aligned in a quadratic fashion while one point is way outside the acceptable boundaries. This is an example of an outlier or anomaly. Could happen because of an error or because of that certain test case.
c. *Missing values.* This table clearly is lacking some values in almost all its samples.
d. *Noise.* In this graph we see that during a certain period the experiment returned bad results that do not follow the logic of the rest of the experiment. This is likely because of a certain error that happened during that time that tampered with the experiment.
e. *Sparsity.* which is are many zero values that do not contribute to the experiment

4. [12 points – 2 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?*
The most likely task is to try and predict a nearsightedness in people with underlying conditions or based on prescription.

b. ***In general**, what is a feature and how would you **exemplify** it with **this data**?*
Feature is a quality or a variable that are used to perform the task. In this case features include "Age", "Spectacle Prescription", "astigmatism" , etc. features are all the columns in a data set.

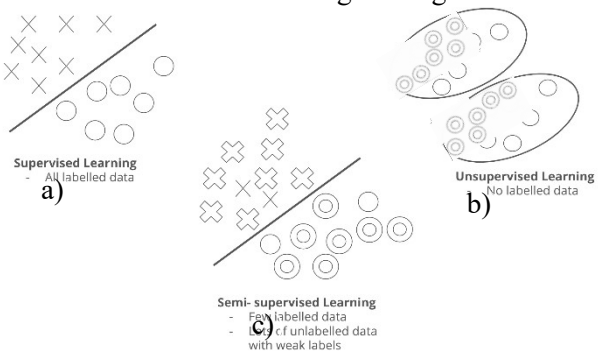
c. ***In general**, what is a feature value and how would you **exemplify** it with **this data**?*
Each feature has a measurable quality. And in every sample these qualities are measured and noted in their designated area. For example here first row third column has a feature value of "No" which is essentially the measurement of that particular feature for this particular sample.

d. ***In general**, what is dimensionality and how would you **exemplify** it with **this data**?*
Dimensionality is referred to the number of features each sample in a data set contains. Basically this is the number of columns a data set has. In this case the dimensionality is 5, because it has 5 attributes or features.

e. ***In general**, what is an instance and how would you **exemplify** it with **this data**?*
A sample or an instance is each row of the data set. Each row signifies a particular case that was examined. In this case each of these samples happen to be a person. For each sample all the features are measured, and the results are recorded.

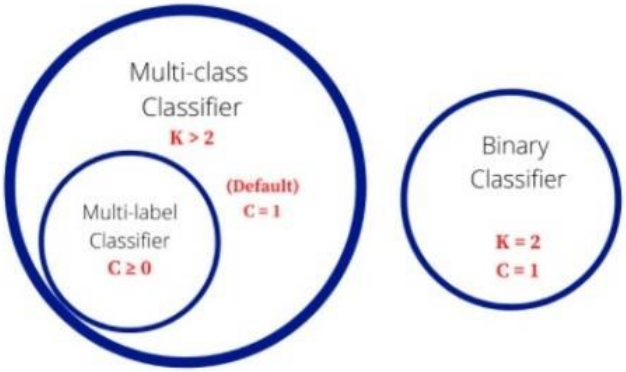
f. ***In general**, what is a class and how would you **exemplify** it with **this data**?*
Class/labels are used for classification of each sample. In this example we don't see the class vector.

5. [6 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 **data labels** information. Hint: check the images to figure out which data sample is labelled.



a. *supervised learning,* because some of the elements are known and others are undefined and could be used as test cases
b. *unsupervised,* elements are not labeled, and algorithm is working to identify clusters
c. *semi-supervised* because some of the figures are defined and others are not (X marks)

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



K = Total number of classes in the problem statement
C = Number of classes an item maybe assigned to

Binary Classifier:

This classifier has 2 possible outcomes/classes, and each item can only be assigned to one of these not both nor 0 classes.

Multi-class Classifier:

This classifier has **strictly more** than 2 possible outcomes/classes and by default each item is assigned **one**.

Multi-Label Classifier:

This classifier has **strictly more** than two outcomes/classes, but the items can be assigned 0 or more of these classes.

7. [54 points] Regarding the training data shown in question 4:
   a. [12 points] Find a **maximally specific hypothesis** that fits the data **following** the strategy of **Find-S algorithm**. Hint: start from the hypothesis h ← (∅, ∅, ∅, ∅).

      My solution is on next page.

   b. [12 points] Complete the given python program (find_s.py) that will read the file contact_lens.csv and output the hypothesis of **Find-S algorithm** (the hypothesis you got in part a). The output should be in this format: ['Sunny', '?', 'Strong', '?']. Add the link to the online repository as the answer to this question.

      The link: https://github.com/Arsham1024/Introduction_ML

   c. [15 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.

      My solution is on next page.

   d. [15 points] Complete the given python program (decision_tree.py) that will read the file contact_lens.csv and output the decision tree of **ID3** (the tree you got in part c). PS: if this tree is different from yours shown in letter c), try to explain why. Add the link to the online repository as the answer to this question.

      The link : https://github.com/Arsham1024/Introduction_ML

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!

Question 7

	1	2	3	4	Class
	Age	Spectacle Prescription	Astigmatism	Tear Production Rate	Recommended Lenses
1	Young	Myope	No	Reduced	No
2	Presbyopic	Myope	No	Normal	No
3	Prepresbyopic	Myope	No	Reduced	No
4	Prepresbyopic	Myope	No	Normal	Yes
5	Presbyopic	Myope	Yes	Normal	Yes
6	Young	Myope	Yes	Normal	Yes
7	Young	Hypermetrope	No	Reduced	No
8	Prepresbyopic	Myope	Yes	Reduced	No
9	Presbyopic	Hypermetrope	No	Reduced	No
10	Young	Myope	Yes	Reduced	Yes

Question 7:

a) Strategy of finding-S algorithm h = (∅, ∅, ∅, ∅)

b) - Complete python program find-S.py ↳ will need Context-Lenses.csv
- output ['sunny', '?', 'strong', '?'] ↳ will output same as @

c) - Derive the Decision tree
- ID3 algorithm ↳ for all splits!
- show calculation for Entropy & info Gain
- plot the final tree!

D) - Complete Decision-tree.py ↳ Read same Context-Lenses.csv
- output the tree
- if Different Explain why? from c)

Part A) ① hypothesis (∅, ∅, ∅, ∅)

② Loop through all (4) samples: Samples: 4, 5, 6, 10

Sample 4 h (prepresbyopic, Myope, No, Normal)

Sample 5 h (? , Myope, ? , Normal)

Sample 6 h (? , Myope, ? , Normal)

Sample 10 h (? , Myope, ? , ?)

③ Return h (? , myope, ? , ?)

Part C)

$$Entropy(s) = -p_{\oplus}\log_2 P_{\oplus} - p_{\ominus}\log_2 P_{\ominus}$$

find E(S):

$$S = 10 \Rightarrow E(S) = -\frac{4}{10}\frac{\log 0.4}{\log 2} - \frac{6}{10}\frac{\log 0.6}{\log 2}$$

$$\boxed{E(S) = 0.971}$$

Age:

$$Gain(S, A) = Entropy(s) - \sum \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Age) = 0.971 + \left[-S_{young} - S_{pres} - S_{prepre} \right]$$

$$= 0.971 + \left[-0.4(1) - 0.3(0.918) - 0.3(0.918) \right]$$

$$\boxed{Gain(S, Age) = 0.0202}$$

S_{young} [2+, 2−]

$S_{presbyopic}$ [1+, 2−] → $\frac{3}{10} - \frac{1}{3}\left(\frac{\log \frac{1}{3}}{\log 2}\right) - \frac{2}{3}\left(\frac{\log \frac{2}{3}}{\log 2}\right)$

$S_{prepresbyopic}$ [1+, 2−] → $\frac{3}{10}$

Spectacle Prescription (SP)

$$Gain(S, SP) = 0.971 + \left[-S_{myope} - S_{Hyper} \right]$$

$$= 0.971 + \left[-\frac{8}{10}(1) - 0 \right]$$

$$\boxed{Gain(S, SP) = 0.171}$$

S_{myope} [4+, 4−] → E = 1

S_{Hyper} [0+, 2−] → E = 0

Astigmatism:

$$Gain(S, As) = 0.971 + \left[-S_{Yes} - S_{NO} \right]$$

$$= 0.971 + \left[-(0.4)(0.811) - (0.6)(0.650) \right]$$

$$Gain(S, As) = 0.2566$$

S_{Yes} [3+, 1−] → 0.4

S_{NO} [1+, 5−] → 0.6

$$-\frac{1}{6}\frac{\log\frac{1}{6}}{\log 2} - \frac{5}{6}\frac{\log\frac{5}{6}}{\log 2}$$

Tear Production Rate

$$Gain(S, TPR) = 0.971 + \left[-S_{Reduced} - S_{Normal} \right]$$

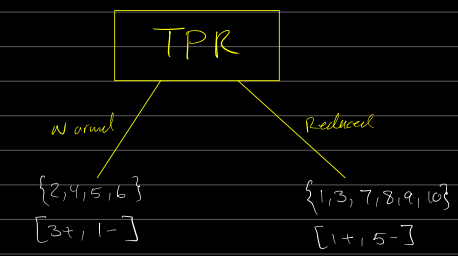
$$= 0.971 + \left[-0.6(0.650) - 0.4(0.811) \right]$$

$$Gain(S, TPR) = 0.2566$$

$S_{Reduced}$ [1+, 5−] ← Done this before

S_{Normal} [3+, 1−]

Tree so far:



Gain(S, Age) = 0.0202 ④

Gain(S, Sp) = 0.171 ③

Gain(S, As) = 0.2566 ②

Gain(S, TPR) = 0.2566 ① → same but I choose Tear Production Rate!

Moving to second level:

Normal:

$$S = 4 \Rightarrow E(S) = -\frac{3}{4}\frac{\log\frac{3}{4}}{\log 2} - \frac{1}{4}\frac{\log\frac{1}{4}}{\log 2}$$

$$E(S) = 0.811$$

Age:

$$Gain(S, Age) = 0.811 + \left[-\cancel{S_{young}}^{0} - \cancel{S_{pre}}^{(1)} - \cancel{S_{prepre}}^{0} \right]$$

$$= 0.811 - (0.5)$$

$$Gain(S, Age) = 0.311$$

S_{young} [1+, 0−] → E=0

S_{pre} [1+, 1−] → E=1

S_{prepre} [1+, 0−] → 0

Spectacle Prescription:

$$Gain(S, SP) = 0.811 + \left[-S_{Myope} - \cancel{S_{Hyper}}^{0}\right]$$

$$= 0.811 - (1)(0.811)$$

$$\boxed{Gain(S, SP) = 0}$$

$S_{Myope}\ [3+, 1-] = \frac{4}{4} = 1$

$S_{Hyper} = 0$

$$\frac{3}{4}\frac{\log\frac{3}{4}}{\log 2} - \frac{1}{4}\frac{\log\frac{1}{4}}{\log 2}$$

Astigmatism:

$$Gain(S, AS) = 0.811 + \left[\cancel{S_{Yes}}^{0} - S_{NO}\right]$$

$$= 0.811 - (0.5)(1)$$

$$\boxed{Gain(S, AS) = 0.311}$$

$S_{Yes} = [2+, 0-] = 0$

$S_{NO} = [1+, 1-] = 1$

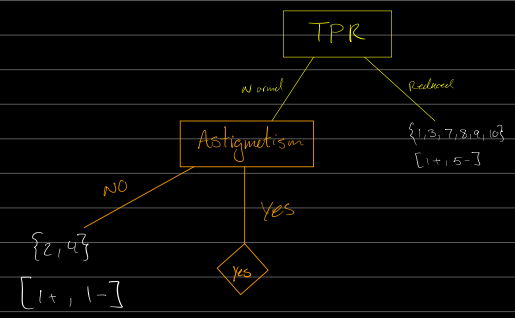
$$-\frac{1}{4}\left(\frac{\log\frac{1}{4}}{\log 2}\right) - \frac{1}{4}\left(\frac{\log\frac{1}{4}}{\log 2}\right)$$

$\boxed{Gain(S, Age) = 0.311}$ (2)

$\boxed{Gain(S, SP) = 0}$ (3)

$\boxed{Gain(S, AS) = 0.311}$ (1)

Tree so far:



TPR → Reduced

$$S = 6 \Rightarrow E(S) = -\frac{1}{6}\frac{\log\frac{1}{6}}{\log 2} - \frac{5}{6}\frac{\log\frac{5}{6}}{\log 2}$$

$$\boxed{E(S) = 0.694}$$

Age:

$$Gain(S, Age) = 0.694 + [-S_{young} - \cancel{S_{pre}}^{0} - \cancel{S_{prepre}}^{0}]$$

$$Gain(S, Age) = 0.694 - 0.459$$

$$\boxed{Gain(S, Age) = 0.235}$$

$S_{young}\ [1+, 2-] =$

$S_{pre}\ [0+, 1-] = 0$

$S_{prepre}\ [0+, 2-] = 0$

$$-\frac{1}{3}\frac{\log\frac{1}{3}}{\log 2} - \frac{2}{3}\frac{\log\frac{2}{3}}{\log 2}$$

$$\frac{3}{6}(0.918)$$

SP:

$$Gain(S, SP) = 0.694 + [-S_{myopre} - S_{Hyper}]$$

$$\boxed{Gain(S, Sp) = 0.361}$$

$S_{Myopre}\ [1+, 3-]$

$S_{Hyper}\ [0+, 2-] = 0$

$$\frac{4}{6}(0.5)$$

AS:

$$Gain(S, AS) = 0.694 + [-S_{Yes} - \cancel{S_{No}}^{0}]$$

$$Gain(S, As) = 0.694 - \left(\frac{1}{3}\right)(1)$$

$$\boxed{Gain(S, As) = 0.361}$$

$S_{Yes}\ [1+, 1-] = 1$

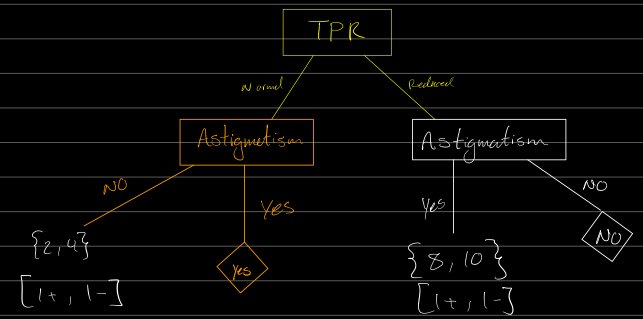
$S_{No}\ [0, 4-] = 0$

$$\boxed{Gain(S, Age) = 0.235} \quad (3)$$

$$\boxed{Gain(S, Sp) = 0.361} \quad (2)$$

$$\boxed{Gain(S, As) = 0.361} \quad (1)$$

tree so far:



TPR = Normal , Astigmatism = NO

S = 2 $\Rightarrow$ E(S) = 1

[1+ , 1-]

Age

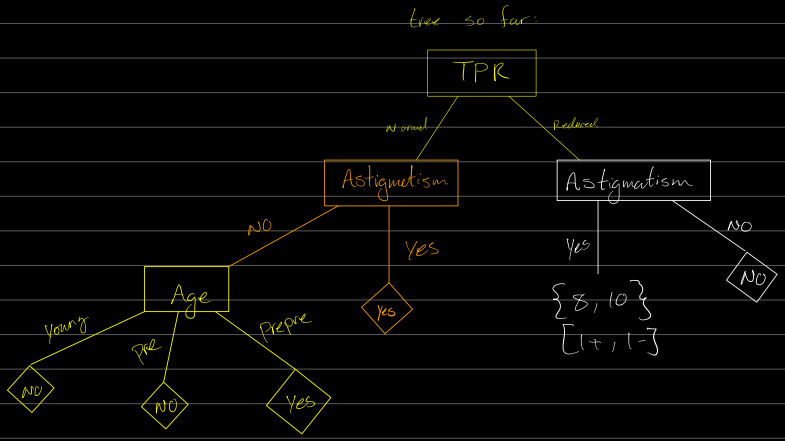
$$Gain(S, Age) = 1 + [\ 0 - S_{pre} - S_{prepre}\ ]$$

(0)

$$\boxed{Gain(S, Age) = 1}$$

SP $\quad Gain(S, SP) = 1 + [\ -1 \quad -0\ ]$

$$\boxed{Gain(S, SP) = 0}$$



TPR = Reduced , Astigmatism = Yes

S = 2 $\Rightarrow$ E(S) = 1

[1+ , 1-]

Age

$$Gain(S, Age) = 1 + [\ -0.5 \quad -0 \quad -0.5\ ]$$

$$= 0$$

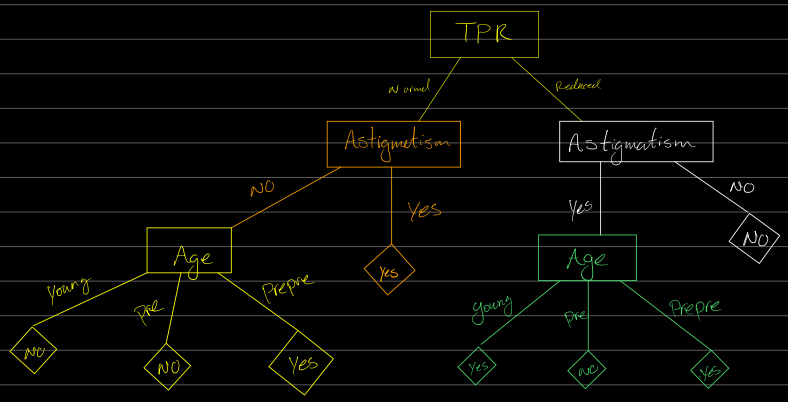
SP

(1)

$$Gain(S, SP) = 1 + [\ -S_{Myope} - \cancel{S_{Hyper}}^{0}\ ]$$

$$= 0$$

Final Tree



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