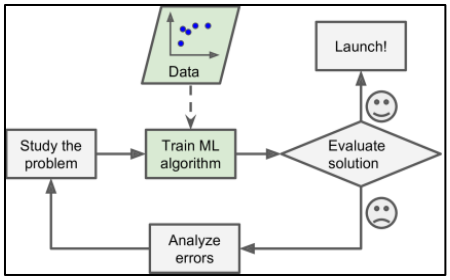
Bronco ID: 014429779
Last Name: Koepke
First Name: Christopher

Assignment #1

CS 4210.01

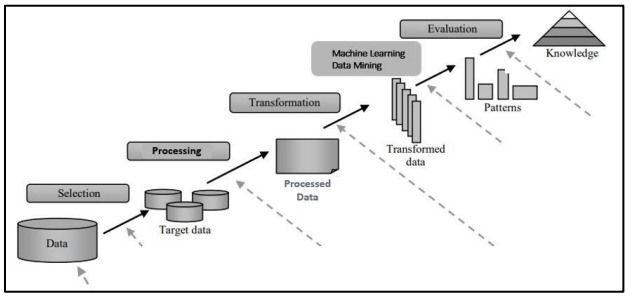
September 17, 2022

1) 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.



- Task T is the answers that we seek, such as predicting the weather or determining if a picture is of a dog or cat. The experience E is the training data that is fed into the ML algorithm that will be used to attain these predictions based on historical data. The performance P is the accuracy of those predictions.

2) 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.

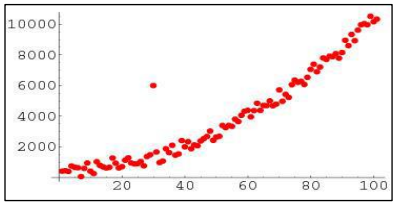


1. **Preprocessing**: Transforming the raw data available into a format that will be used to train the ML algorithm. Some considerations are on how to handle missing values, data sparsity, unnecessary features.
2. **Machine Learning/Data Mining**: This is where the ML algorithms and techniques will be designed and implemented. Many different systems can be used for ML, such as supervised, unsupervised, Online vs Batched learning, instanced vs model-based learning; multiple systems can also be combined to achieve the desired results.
3. **Postprocessing**: The application of various techniques (visualization, filtering, …) to extract knowledge from the output of the ML algorithm to make informed decisions or to simply document the results.

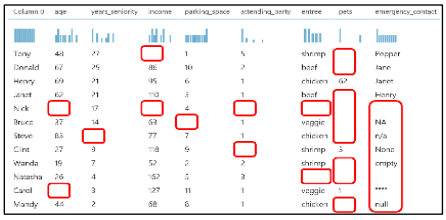
3) Machine learning algorithms face multiple challenges while analyzing data such as scalability, distribution, sparsity, resolution, class imbalance, noise, outliers, missing values, and duplicated data. For each image below, name and explain what the corresponding challenge is from this list (you do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

a.

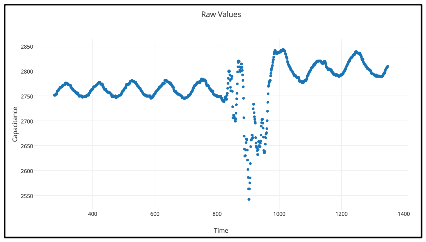
b.



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	2
3	0	0	3	0
0	4	0	0	0

a) **Data Distribution**: training data that is more evenly distributed (normal distribution) should result in more accurate predictions in the testing data.
b) **Outlier**: Valid data points outside the normal range of the data. These outliers can cause inaccuracies in the predictions generated by the ML algorithms. Outlier detection and removal should be considered.
c) **Missing/Incomplete Feature Values**: missing data points will impact learning negatively by introducing bias. Two techniques used to correct this issue is the removal of the record with the missing values or insert some value (mean/most suitable value) into that field to be able to use that record.
d) **Noise**: The collection of unwanted/unintended data (such as background noise collected from a microphone). Use of a noisy dataset in ML algorithms could result in the algorithm generating patterns from it and using that pattern in generating its output. Overfitting is also a problem with noisy datasets.
e) **Data Sparsity**: dataset containing zeros in many fields in most records. This sparsity can increase the complexity of the ML algorithms (both space and time) and cause unintended behaviors with the ML model.

4) 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

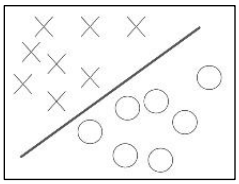
a. What is the most likely task that data scientists are trying to accomplish?
   - To determine if lenses are recommended for a given combination of attributes.
b. In general, what is a feature and how would you exemplify it with this data?
   - A feature is an attribute of an object or instance of a dataset. In the example above, astigmatism is one feature of each instance of the dataset.
c. In general, what is a feature value and how would you exemplify it with this data?
   - A feature value is the measurable property of a feature. For the case of the feature Astigmatism, its feature value is either No or Yes.
d. In general, what is dimensionality and how would you exemplify it with this data?
   - Dimensionality is the number of features that an instance has relative to the number of instances in a dataset. A dataset with many instances (let's say ~1000) with only a few features (let's say ~4) might have many cases of feature value combinations being repeated. A dataset with many instances (~1000) and many features (~100) might have very few feature value combinations repeated. The first example will allow an ML algorithm to denote patterns in the data whereas the other example might not see any patterns. For the dataset given, I would consider it as being low in dimensionality,

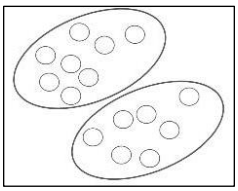
thus removing itself from the curse of dimensionality; where a dataset with a relatively large number of features would increasingly lead to sparsity in the dataset.

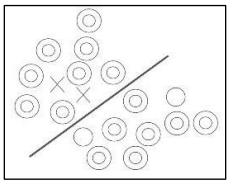
e. In general, what is an instance and how would you exemplify it with this data?
   - An instance is a collection of fields that describe an object. Each row of the dataset is an instance.

f. In general, what is a class and how would you exemplify it with this data
   - A class (or label) is the name of the output of the ML algorithm. Our target in this example is whether an instance is recommended lenses or not and the class/label is either No or Yes.

5) 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.

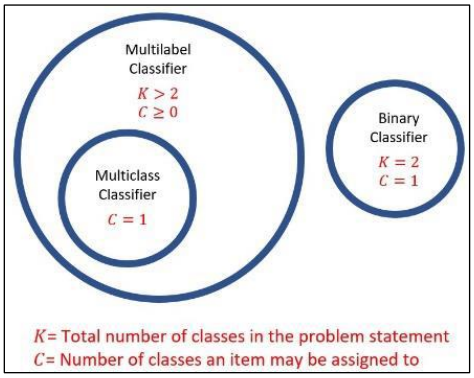
a) 

b) 

c) 

a) Supervised Learning; linear regression, all data is labeled.
b) Unsupervised Learning; clustering, no data is labeled.
c) Semi-Supervised Learning; few labeled data

6) Explain the tasks addressed by each classifier to the right.
   - Multiclass Classifier; Only one label can be assigned to each class, but the number of possible labels is greater than two.
   - Multilabel Classifier; Zero or more labels can be assigned to each class and the number of possible labels is greater than two.
   - Binary Classifier: Only one label can be assigned to each class and the number of possible labels is exactly two.



7) Regarding the training data shown in question 4:
   a. 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.

$$Values(Recommended\ Lenses\ as\ S) = Yes, No$$

$$Values(Age\ of\ Eyes) = Young, Presbyopic, Prepresbyopic$$

$$Values(Spectacle) = Myope, Hypermetrope$$

$$Values(Astigmatism) = Yes, No$$

$$Values(Tear\ Production) = Normal, Reduced$$

$$S = [4+, 6-]$$

$$Entropy(S) = -\left(\frac{4}{10}\right)\log_2\left(\frac{4}{10}\right) - \left(\frac{6}{10}\right)\log_2\left(\frac{6}{10}\right) \approx 0.970951$$

NOTE: Number of Positives = Number of Negatives ➔ Entropy = 1.000000

NOTE: All instances is either positive or negative ➔ Entropy = 0.000000

- **Age of Eyes**

$$S_{Young} = [2+, 2-]$$
$$S_{Presbyopic} = [1+, 2-]$$
$$S_{Prepresbyopic} = [1+, 2-]$$

$$Entropy(S_{Young}) = 1.000000$$
$$Entropy(S_{Presbyopic}) = -\left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right) - \left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right) \approx 0.918296$$
$$Entropy(S_{Prepresbyopic}) = -\left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right) - \left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right) \approx 0.918296$$

$$Gain(S, Age\ of\ Eyes) = Entropy(S) - \sum_{v \in \{Young, Presbyopic, Prepresbyopic\}} \frac{|S_v|}{|S|} Entropy(S_v)$$
$$= 0.970951 - \left(\frac{4}{10}\right)(1.000000) - \left(\frac{3}{10}\right)(0.918296) - \left(\frac{3}{10}\right)$$
$$= 0.019973$$

- **Spectacle Prescription**

$$S_{Myope} = [4+, 4-]$$
$$S_{Hypermetrope} = [0+, 2-]$$

$$Entropy(S_{Myope}) = 1.000000$$
$$Entropy(S_{Myope}) = 0.000000$$

$$Gain(S, Spectacle\ Prescription) = Entropy(S) - \sum_{v \in \{Myope, Hypermetrope\}} \frac{|S_v|}{|S|} Entropy(S_v)$$
$$= 0.970951 - \left(\frac{8}{10}\right)(1.000000) - \left(\frac{2}{10}\right)(0.000000)$$
$$= 0.170951$$

- **Astigmatism**

$$S_{No} = [1+, 5-]$$
$$S_{Yes} = [3+, 1-]$$

$$Entropy(S_{No}) = -\left(\frac{1}{6}\right)\log_2\left(\frac{1}{6}\right) - \left(\frac{5}{6}\right)\log_2\left(\frac{5}{6}\right) \approx 0.650022$$
$$Entropy(S_{Yes}) = -\left(\frac{3}{4}\right)\log_2\left(\frac{3}{4}\right) - \left(\frac{1}{4}\right)\log_2\left(\frac{1}{4}\right) \approx 0.811278$$

$$Gain(S, Astigmatism) = Entropy(S) - \sum_{v\in\{No,Yes\}} \frac{|S_v|}{|S|} Entropy(S_v)$$
$$= 0.970951 - \left(\frac{6}{10}\right)(0.650022) - \left(\frac{4}{10}\right)(0.811278)$$
$$= 0.256426$$

- **Tear Production**

$$S_{Normal} = [3+, 1-]$$
$$S_{Reduced} = [1+, 5-]$$

$$Entropy(S_{Normal}) = -\left(\frac{3}{4}\right)\log_2\left(\frac{3}{4}\right) - \left(\frac{1}{4}\right)\log_2\left(\frac{1}{4}\right) \approx 0.811278$$
$$Entropy(S_{Reduced}) = -\left(\frac{1}{6}\right)\log_2\left(\frac{1}{6}\right) - \left(\frac{5}{6}\right)\log_2\left(\frac{5}{6}\right) \approx 0.650022$$

$$Gain(S, Astigmatism) = Entropy(S) - \sum_{v\in\{Normal,Reduced\}} \frac{|S_v|}{|S|} Entropy(S_v)$$
$$= 0.970951 - \left(\frac{4}{10}\right)(0.811278) - \left(\frac{6}{10}\right)(0.650022)$$
$$= 0.256426$$

- ROOT OF DECISION TREE
  - Both Astigmatism and Tear Production have the highest value for Information Gain, thus I choose Astigmatism since it was calculated first.

$$S_{Astigmatism,NO\ Branch} = [1+, 5-]$$
$$Entropy(S_{Astigmatism,NO\ Branch}) = -\left(\frac{1}{6}\right)\log_2\left(\frac{1}{6}\right) - \left(\frac{5}{6}\right)\log_2\left(\frac{5}{6}\right) \approx 0.650022$$

- **Age of Eyes (Astigmatism No Branch)**

$$S_{No\ Branch,Young} = [0+, 2-]$$
$$S_{No\ Branch,Presbyopic} = [0+, 2-]$$
$$S_{No\ Branch,Prepresbyopic} = [1+, 1-]$$

$$Entropy(S_{No\ Branch,Young}) = 0.000000$$
$$Entropy(S_{No\ Branch,Presbyopic}) = 0.000000$$
$$Entropy(S_{No\ Branch,Prepresbyopic}) = 1.000000$$

$$Gain(S_{No\ Branch}, Age\ of\ Eyes) = Entropy(S_{No\ Branch}) - \sum_{v\in\{Young,Presbyopic,Prepresbyopic\}} \frac{|S_v|}{|S|} Entropy(S_v)$$
$$= 0.650022 - \left(\frac{2}{6}\right)(0.000000) - \left(\frac{2}{6}\right)(0.000000) - \left(\frac{2}{6}\right)(1.000000)$$
$$= 0.316689$$

- **Spectacle Prescription (Astigmatism No Branch)**

$$S_{No\ Branch,Myope} = [1+, 3-]$$
$$S_{No\ Branch,Hypermetrope} = [0+, 2-]$$

$$Entropy(S_{No\ Branch,Myope}) = -\left(\frac{1}{4}\right)\log_2\left(\frac{1}{4}\right) - \left(\frac{3}{4}\right)\log_2\left(\frac{3}{4}\right) \approx 0.811278$$
$$Entropy(S_{No\ Branch,Hypermetrope}) = 0.000000$$

$$Gain(S_{No\ Branch}, Spectacle\ Prescrip) = Entropy(S_{No\ Branch}) - \sum_{v\in\{Myope,Hypermetrop\}} \frac{|S_v|}{|S|} Entropy(S_v)$$
$$= 0.650022 - \left(\frac{4}{6}\right)(0.811278) - \left(\frac{2}{6}\right)(0.000000)$$
$$= 0.109170$$

- **Tear Production (Astigmatism No Branch)**

$$S_{No\ Branch,Normal} = [1+, 1-]$$
$$S_{No\ Branch,Reduced} = [0+, 4-]$$

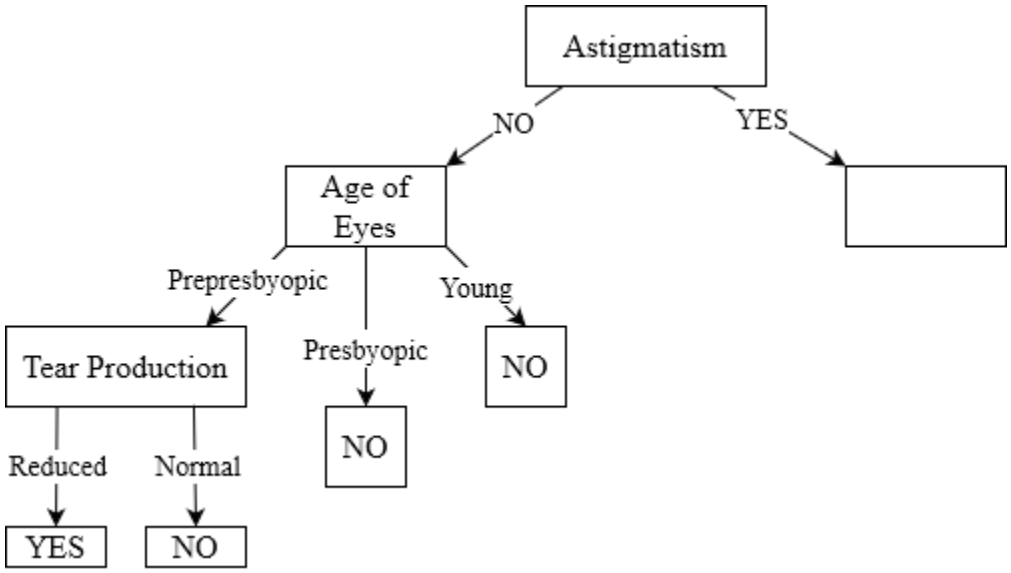
$$Entropy(S_{No\ Branch,Normal}) = 1.000000$$
$$Entropy(S_{No\ Branch,Reduced}) = 0.000000$$

$$Gain(S_{No\ Branch}, Tear\ Production) = Entropy(S_{No\ Branch}) - \sum_{v \in \{Normal, Reduced\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$= 0.650022 - \left(\frac{2}{6}\right)(1.000000) - \left(\frac{4}{6}\right)(0.000000)$$

$$= 0.316689$$

- ASTIGMATISM NO BRANCH NODE
  - Both Age of Eyes and Tear Production have the highest value for Information Gain, thus I choose Age of Eyes since it was calculated first.



  - The rest of the No branch from Astigmatism can be inferred from the remaining data in this subset.

$$S_{Astigmatism,Yes\ Branch} = [3+, 1-]$$

$$Entropy(S_{Astigmatism,Yes\ Branch}) = -\left(\frac{3}{4}\right)\log_2\left(\frac{3}{4}\right) - \left(\frac{1}{4}\right)\log_2\left(\frac{1}{4}\right) \approx 0.811278$$

- Spectacle Prescription (Astigmatism Yes Branch)

$$S_{Yes\ Branch,Myope} = [3+, 1-]$$

$$Entropy(S_{Yes\ Branch,Myope}) = -\left(\frac{3}{4}\right)\log_2\left(\frac{3}{4}\right) - \left(\frac{1}{4}\right)\log_2\left(\frac{1}{4}\right) \approx 0.811278$$

$$Gain(S_{Yes\ Branch}, Spectacle\ Prescription) = Entropy(S_{Yes\ Branch}) - \sum_{v \in \{Myope\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$= 0.811278 - \left(\frac{4}{4}\right)(0.811278)$$

$$= 0.000000$$

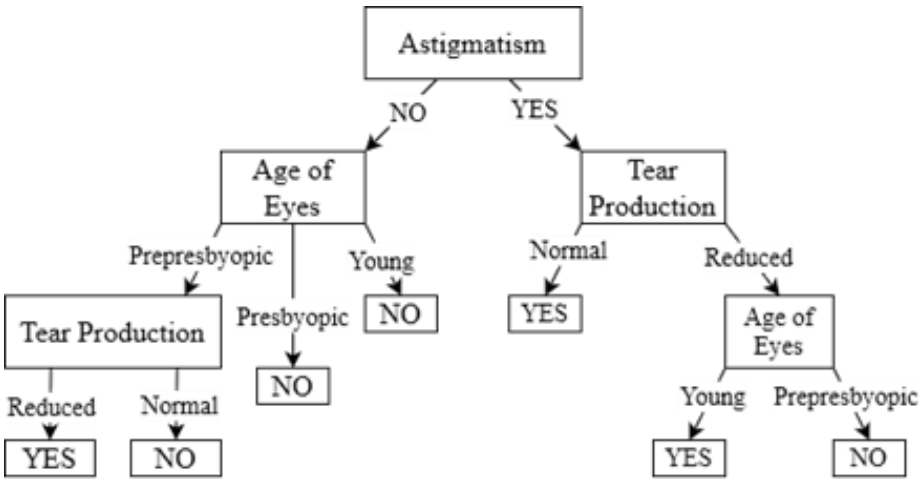
- Tear Production (Astigmatism Yes Branch)

$$S_{Yes\ Branch,Normal} = [2+, 0-]$$
$$S_{Yes\ Branch,Reduced} = [1+, 1-]$$

$$Entropy(S_{Yes\ Branch,Normal}) = 0.000000$$
$$Entropy(S_{Yes\ Branch,Reduced}) = 1.000000$$

$$Gain(S_{Yes\ Branch}, Tear\ Production) = Entropy(S_{Yes\ Branch}) - \sum_{v \in \{Normal,Reduced\}} \frac{|S_v|}{|S|} Entropy(S_v)$$
$$= 0.811278 - \left(\frac{2}{4}\right)(0.000000) - \left(\frac{2}{4}\right)(1.000000)$$
$$= 0.311278$$

- ASTIGMATISM YES BRANCH NODE
  - The only choice available based on information gain would be Tear Production.
  - From the chosen node, it can be inferred that a person with astigmatism and normal tear production will lead to a yes.
  - From the two remaining records on the reduced branch of the tear production node, the only field that could determine the class label would be the age of the eyes, thus completing the decision tree for the given data.



b. 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/chris-k87/CS_4210.01/tree/main/Assignment_1

c. 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.
- The obvious differences are the choices between fields with the same value for information gained. In my tree I chose astigmatism as the root, whereas the python program chose tear production.
- A major difference with the decision tree created by the python program is that each proceeding level of the tree has the same nodes. So at depth 1, the two nodes are astigmatism and at depth 2, it is age of the eyes.
- It was my understanding that a field that was chosen as a node cannot be used for a different branch on the same depth level.