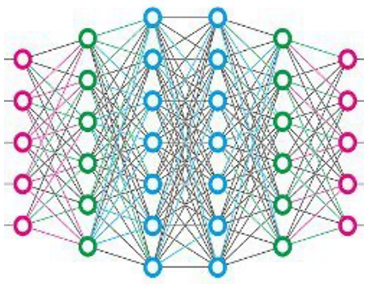
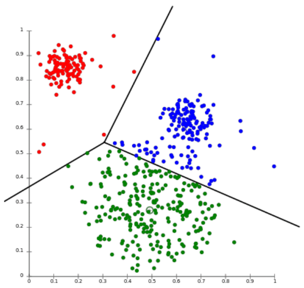
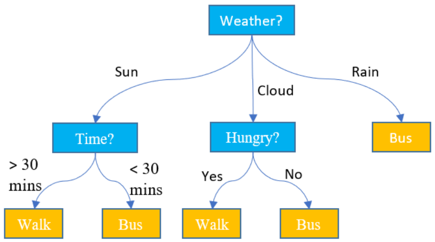
# Welcome to CS 445
# Introduction to Machine Learning

## Instructor: Dr. Kevin Molloy







# Announcements

- Workstation Configuration should be complete
- We will be using Jupyter notebooks on Thursday for class
- PA 0 is due in 1 week to Autolab (multiple submissions allowed).
- Canvas Quiz 1 will be due at 11:59 PM tomorrow (Wednesday).
- PA 1 is posted.

# Learning Objectives for Today

- Define and give an example of nominal and ordinal categorical features
- Define and give an example of interval and ratio numeric features.
- Utilize a decision tree to predict class labels for new data.
- Define and compute **entropy** and utilize it to characterize the impurity of a set
- Define an algorithm to determine split points that can be used to construct a decision tree classifier.

# Plan for Today

- Complete Lab Activities 1 – 3 (groups of 2 to 3 people)
- Discussion

- Complete Lab Activities 4
- Discussion

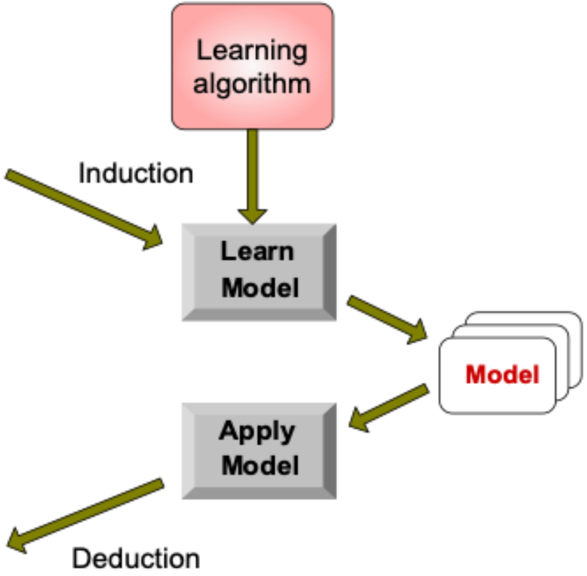
- Complete Lab Activity 5
- Discussion

- Complete Lab Activity 6 and 7
- **Submit** completed PDF to Canvas

# Supervised Learning

| Tid | Attrib1 | Attrib2 | Attrib3 | Class |
|---|---|---|---|---|
| 1 | Yes | Large | 125K | No |
| 2 | No | Medium | 100K | No |
| 3 | No | Small | 70K | No |
| 4 | Yes | Medium | 120K | No |
| 5 | No | Large | 95K | Yes |
| 6 | No | Medium | 60K | No |
| 7 | Yes | Large | 220K | No |
| 8 | No | Small | 85K | Yes |
| 9 | No | Medium | 75K | No |
| 10 | No | Small | 90K | Yes |

Training Set




| Tid | Attrib1 | Attrib2 | Attrib3 | Class |
|---|---|---|---|---|
| 11 | No | Small | 55K | ? |
| 12 | Yes | Medium | 80K | ? |
| 13 | Yes | Large | 110K | ? |
| 14 | No | Small | 95K | ? |
| 15 | No | Large | 67K | ? |

Test Set

**Supervised learning** learns a function that maps an input example to an output. This function/model is inferred from data points with known outcomes (training data).

# Types of Data (IDD 2.1)

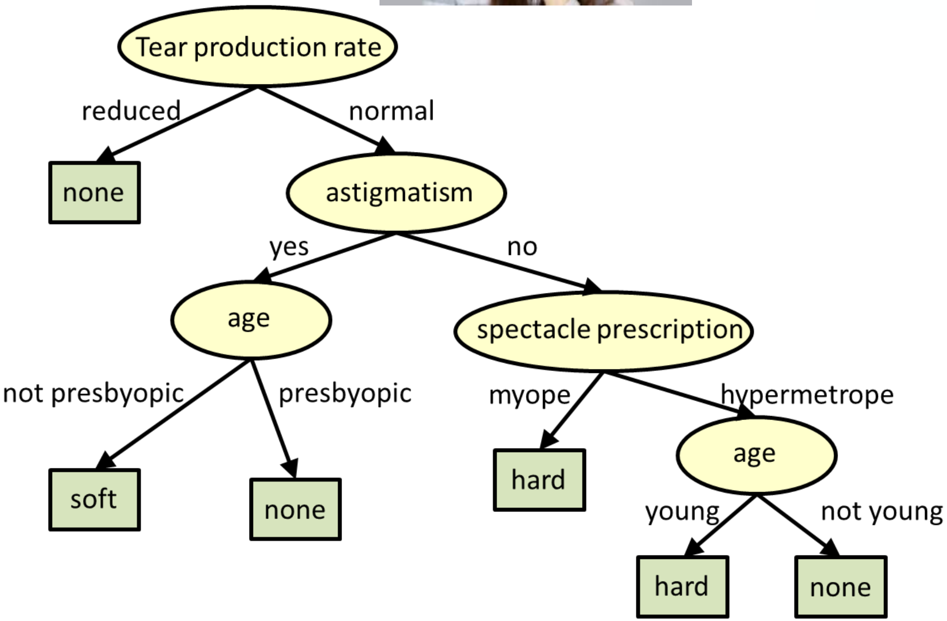
| | Attribute Type | Description | Examples | Operations |
|---|---|---|---|---|
| Categorical Qualitative | Nominal | Nominal attribute values only distinguish. (=, ≠) | zip codes, employee ID numbers, eye color, sex: {*male, female*} | mode, entropy, contingency correlation, $\chi 2$ test |
| | Ordinal | Ordinal attribute values also order objects. (<, >) | hardness of minerals, {*good, better, best*}, grades, street numbers | median, percentiles, rank correlation, run tests, sign tests |
| Numeric Quantitative | Interval | For interval attributes, differences between values are meaningful. (+, - ) | calendar dates, temperature in Celsius or Fahrenheit | mean, standard deviation, Pearson's correlation, *t* and *F* tests |
| | Ratio | For ratio variables, both differences and ratios are meaningful. (*, /) | temperature in Kelvin, monetary quantities, counts, age, mass, length, current | geometric mean, harmonic mean, percent variation |

From S. S. Stevents

# Decision Trees


+

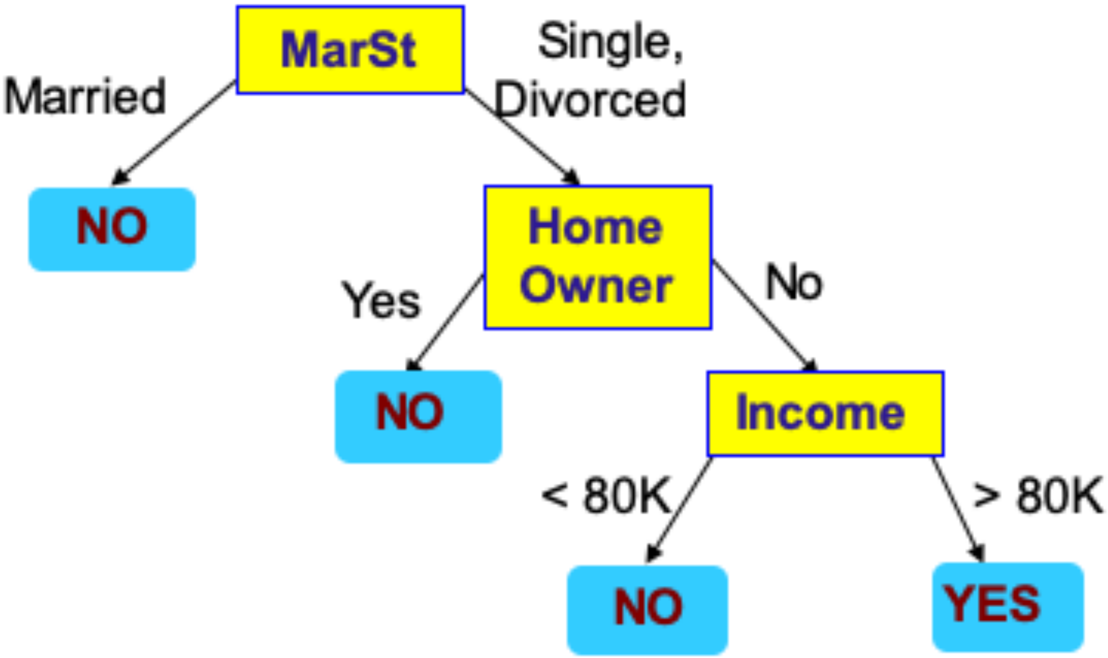
Proceed to our in-class activity today and complete Activities 1, 2, and 3




What type of contact lens a person may wear?

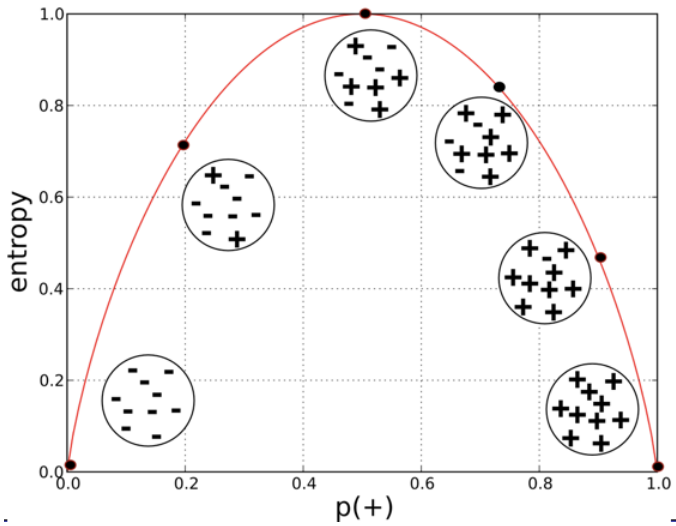
From Bhi ksha Raj, Carnegie Mellon University

# Predicting an Outcome given the Tree




| Homeowner | Marital Status | Income | Class (Loan will default?) |
| --- | --- | --- | --- |
| No | Married | 80,000 | ?? |

# Node Impurity




**Entropy formula** $-\sum_{c=0}^{c-1}\left(p_i(t)\log_2\left(p_i(t)\right)\right)$

Recall that: $\log_2 x = \dfrac{\log_{10} x}{\log_{10} 2}$

And in python **math.log(x,2)** or **np.log2(x)**

**Question**: Given 13 positive examples and 20 negative examples.

What is the entropy?

# Decision Tree Algorithm

```
if stopping_conf(E, F) == true
    leaf = CreateNode()
    leaf.label = FindMajorityClass(E)
    return leaf
else
    root = CreateNode()
    root.test_cond = find_best_split(E, F)
    E_left = E_right = {}
    for each e ∈ E:
        if root.test_cond would split e left:
            E_left = E_left ∪ e
        else
            E_right = E_right ∪ e
    root.left = TreeGrowth(Eleft, F)
    root.right = TreeGrowth(Eright, F)
    return root
```

**E** is the set of training examples (including their labels).

**F** is the attribute set (metadata) to describe the features/attributes of E.

# Decision Tree Algorithm (Binary Splits Only)

```
if stopping_conf(E, F) == true
    leaf = CreateNode()
    leaf.label = FindMajorityClass(E)
    return leaf
else
    root = CreateNode()
    root.test_cond = find_best_split(E, F)
    E_left = E_right = {}
    for each e ∈ E:
        if root.test_cond would split e left:
            E_left = E_left ∪ e
        else
            E_right = E_right ∪ e
    root.left = TreeGrowth(Eleft, F)
    root.right = TreeGrowth(Eright, F)
    return root
```

**E** is the set of training examples (including their labels).

**F** is the attribute set (metadata) to describe the features/attributes of E.

# How to Select a Split?

```
root.test_cond = find_best_split(E, F)
```

**Goal:** Select a feature to split and a split point that divides the data into two groups (left branch and right branch) that, when perform recursively, will result in the minimal impurity in the leaf nodes.

**Naïve Solution:** Attempt every possible decision tree that can be constructed.

**Problem:** The search space of possible trees is exponential in the size of the number of features and the number of splits within each feature. Thus, it is computationally intractable to evaluate all trees. This problem is known to be **NP-Complete.**

# A Greedy Approximation

```
root.test_cond = find_best_split(E, F)
```

**Approximation:** At each node, select the feature and split within that feature that provides the largest information gain. This is a **greedy approximation algorithm**, since it picks the best option at a given time (greedy).

$$\textbf{Info Gain} = \boldsymbol{Entropy(Parent) - \sum_{v \in (Left,right)} \frac{N(v)}{N} Entropy(v)}$$

where *N(v)* is the number of instances assign to node *v* (left or right subnode) and *N* is the total number of instances in the parent node.

(See IDD section 3.3.3 Splitting on Qualitative attributes).

# Information Gain: An Example for a Split Candidate

| Home Owner | Martial Status | Annual Income | Defaulted Borrower |
|---|---|---|---|
| Yes | Single | 120,000 | No |
| No | Married | 100,000 | No |
| Yes | Single | 70,000 | No |
| No | Single | 150,000 | Yes |
| Yes | Divorced | 85,000 | No |
| No | Married | 80,000 | Yes |
| No | Single | 75,000 | Yes |

Consider Martial Status (3 possible splits):

Entropy(parent) =

-(3/7 * log2(3/7) + 4/7 * log2(4/7) ≈ 0.99

# Information Gain: An Example for a Split Candidate

| Home Owner | Martial Status | Annual Income | Defaulted Borrower |
|---|---|---|---|
| Yes | Single | 120,000 | No |
| No | Married | 100,000 | No |
| Yes | Single | 70,000 | No |
| No | Single | 150,000 | Yes |
| Yes | Divorced | 85,000 | No |
| No | Married | 80,000 | Yes |
| No | Single | 75,000 | Yes |

Consider Martial Status (3 possible splits):

Entropy(parent) =

$-(4/7 \log_2(4/7) + 3/7 \log_2(3/7) \approx 0.99$

1 of 3 possible splits:

- (single) to the left
- (married/divorced) right

# Information Gain: An Example for a Split Candidate

| Home Owner | Martial Status | Annual Income | Defaulted Borrower |
|---|---|---|---|
| Yes | Single | 120,000 | No |
| No | Married | 100,000 | No |
| Yes | Single | 70,000 | No |
| No | Single | 150,000 | Yes |
| Yes | Divorced | 85,000 | No |
| No | Married | 80,000 | Yes |
| No | Single | 75,000 | Yes |

Consider Martial Status (3 possible splits):

Entropy(parent) =

$-(4/7 \log_2(4/7) + 3/7 \log_2(3/7) \approx 0.99$

1 of 3 possible splits:

- (single) to the left
- (married/divorced) right

$$\text{Left} = {}^4/_7 * -({}^2/_4 \log_2 {}^2/_4 + {}^2/_4 \log_2 {}^2/_4)$$

# Information Gain: An Example for a Split Candidate

| Home Owner | Martial Status | Annual Income | Defaulted Borrower |
|---|---|---|---|
| Yes | Single | 120,000 | No |
| No | Married | 100,000 | No |
| Yes | Single | 70,000 | No |
| No | Single | 150,000 | Yes |
| Yes | Divorced | 85,000 | No |
| No | Married | 80,000 | Yes |
| No | Single | 75,000 | Yes |

Consider Martial Status (3 possible splits):

Entropy(parent) =

-(4/7 log2(4/7) + 3/7 log2(3/7) ≈ 0.99

1 of 3 possible splits:

- (single) to the left
- (married/divorced) right

$$\text{Left} = {}^{4}/_{7} * -1 \; * \; ({}^{2}/_{4} \log_2 {}^{2}/_{4} + {}^{2}/_{4} \log_2 {}^{2}/_{4})$$

$$Right = {}^{3}/_{7} * -1 \; * ({}^{2}/_{3}\, log_2 {}^{2}/_{3} + {}^{1}/_{3}\, log_2 {}^{1}/_{3})$$

# Information Gain: An Example for a Split Candidate

| Home Owner | Martial Status | Annual Income | Defaulted Borrower |
|---|---|---|---|
| Yes | Single | 120,000 | No |
| No | Married | 100,000 | No |
| Yes | Single | 70,000 | No |
| No | Single | 150,000 | Yes |
| Yes | Divorced | 85,000 | No |
| No | Married | 80,000 | Yes |
| No | Single | 75,000 | Yes |

Consider Martial Status (3 possible splits):

Entropy(parent) =

-(4/7 log2(4/7) + 3/7 log2(3/7) ≈ 0.99

1 of 3 possible splits:

- (single) to the left
- (married/divorced) right

$$\mathrm{Left} = {}^{4}/_{7} * -1 \ * \ ({}^{2}/_{4} \log_2 {}^{2}/_{4} + {}^{2}/_{4} \log_2 {}^{2}/_{4})$$

$$Right \ = \ {}^{3}/_{7} * -1 \ * ({}^{2}/_{3}\, log_2 \, {}^{2}/_{3} + {}^{1}/_{3}\, log_2 \, {}^{1}/_{3})$$

Info Gain = $Entropy(Parent) - \sum_{v \in (Left,right)} \frac{N(v)}{N} Entropy(v)$

**Info Gain = 0.99 $-(.57+.39) = 0.03$**

# Information Gain: Continuous Attributes

| Home Owner | Martial Status | Annual Income | Defaulted Borrower |
|---|---|---|---|
| Yes | Single | 120,000 | No |
| No | Married | 100,000 | No |
| Yes | Single | 70,000 | No |
| No | Single | 150,000 | Yes |
| Yes | Divorced | 85,000 | No |
| No | Married | 80,000 | Yes |
| No | Single | 75,000 | Yes |

For annual income, where to split?

- Sort the feature and make the midpoint between adjacent values the candidate split point.
- Compute the info gain for each of these splits.

# Bounds on Split Points for a Single Feature

**Discussion**

# For Next time

**Homework**:

- Work on PA 0
- Complete Lab/PDF and submit to Canvas by Wed at 9 PM.

**Reading**: IDD Sections 2.1 and 3.3

**Canvas Quiz** Short Reading Quiz (due at 11:59 pm on Wednesday)

**Lab** for next class will use **Jupyter Notebooks.** Make sure you can download the lab from the class website and start the notebook on your computer (the Resources area on the website has instructions on starting the notebook).