Intro. to AI

- AI
  - Goal is generally catagorized as machines acting intelligently

- Turing Test
  - Measures intelligence
  - 1 user, 1 observer, and the machine
    - observer tries to discern human and machine

- AI v. ML
  - AI: Attempt to build machines capable of simulating intelligent human behavior
  - ML: The science of getting computers to act, without being programmed

- Data Exhaust
  - The trail of data you leave behind as you journey through your day

Data + Information

- Data
  - "Something that is given"
  - Something we collect

- Data v. Info
  - Data: The raw product; Gathering of facts and figures
  - Info: The data in context; used or presented to make it meaningful

Categorizing Data

- ML
  - Composed of algorithms that req. input
  - The algorithms can be split into two types
    - Supervised Learning
    - Unsupervised learning

- Supervised Learning
  - Info's available, but not enough to learn until more data's gathered
  - Has a known label (property or result set) that can feed it data...

(Cont.) From a training set (data)
- The data being fed fall into 2 cat's
  - Independent Data
  - Dependent Data

- Independent V. Dependent Data
  - Ind. data (predictors) used to determine the dep. data (target/outcome)
  - The outcome is the dependent data

- Supervised Learning
  - Train algorithm w/ training data
    - This will improve the ability to accurately respond to future data
      - Generalization

- Overfitting
  - Model is strict w/ test data
    - Doesn't handle test data properly when presented w/ new data that is outside of the strict test data used for training

- Unsupervised Learning
  - Structure of data set is unknown
  - Data isn't classified or labeled
  - Algorithms draw inferences from (usually large) datasets, w/ unlabled responses. (usually a mix)

- Data stored in a label
  - Classification
  - Regression
- Classification
  - When independent data is defined as a class label, and has a definite discrete value.
- Regression
  - When ranges of data are stored using real #'s
- Structured v. Unstructured data
  - Structured
    - High degree of organization where each item falls into a par. type.
      - #'s, dates
- Steps to look at structured data
  - Gather example inputs and results
  - Generate a model from input & results
  - Add new inputs (test data) to the model
  - Test the results (outputs) from the model + evaluate

Why Now?

- Computational Power
  - The inc. in Computer processing power, + resources like the cloud have made it possible to process massive amounts of data quicker
  - W/ this increase in comp. power, Machine learning algorithms have led to better ways of understanding and analyzing data

- Moore's Law
  - # of transistors on microchip doubles every 2 years, and cost is halved

- Workflow

- Models
  - CRISP-DM
    - Cross-Industry Standard process for data mining

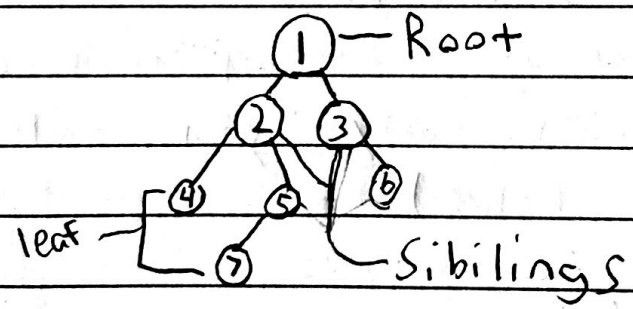
1 Business Understanding
2 Data Understanding
3 Data Preperation
4 Modeling
5 Evaluation
6 Deployment

Binary Trees

- Decision Tree Algorithms
  - Construct a model based on answers (or values) from data

- Binary Trees
  - Data structure based on the idea of nodes that have both a left + right reference to other nodes.

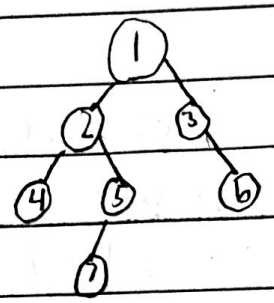
  - There are a maximum of 2 child nodes from every node



Recursion

- Recursion
  - When a method makes a call to itself

Tree Traversal



- Pre-order Traversal
  - Print given node b4 its children
  - Process to the left first
  - Ex: 1, 2, 4, 5, 7, 3, 6

- Post order Traversal
  - Print each node after its children
  - Process to the left first
  - Ex: 4, 7, 5, 2, 6, 3, 1

- In Order Traversal
  - Print the left child (including entire subtree), print the node, then visit right child.
  - Ex: 4, 2, 7, 5, 1, 3, 6

Decision Tree Algorithms

- Types:
  - Classification & regression Tree (cart)
  - Iterative Dichotomiser 3 (ID3)
  - C4.5 (succesor of ID3)
  - Chi-squared automatic interaction Det. (CHAID)
  - Decision Stump
  - Conditional Decision Trees

Information Entropy

- Data w/ a high level of uncertainty (randomness) will contain more
- If given no data, no learning takes place

- Variance
  - How far data's spread out
    - The avg of the sqrd differences from the mean

- Information Gain
  - Measures how much entropy is reduced when partioning an attribute on A
    - The higher the #, the better it classifies data