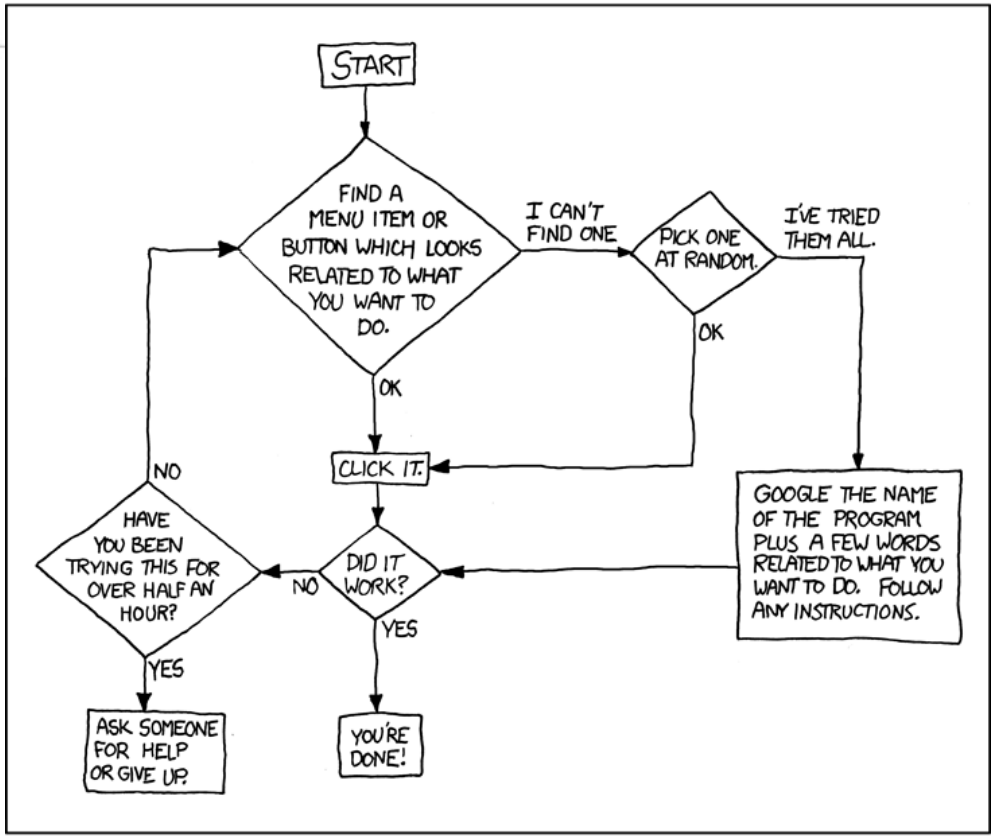
Decision Trees

CS 450 – Machine Learning and Data Mining

The Decision Tree

DEAR VARIOUS PARENTS, GRANDPARENTS, CO-WORKERS, AND OTHER "NOT COMPUTER PEOPLE."

WE DON'T MAGICALLY KNOW HOW TO DO EVERYTHING IN EVERY PROGRAM. WHEN WE HELP YOU, WE'RE USUALLY JUST DOING THIS:



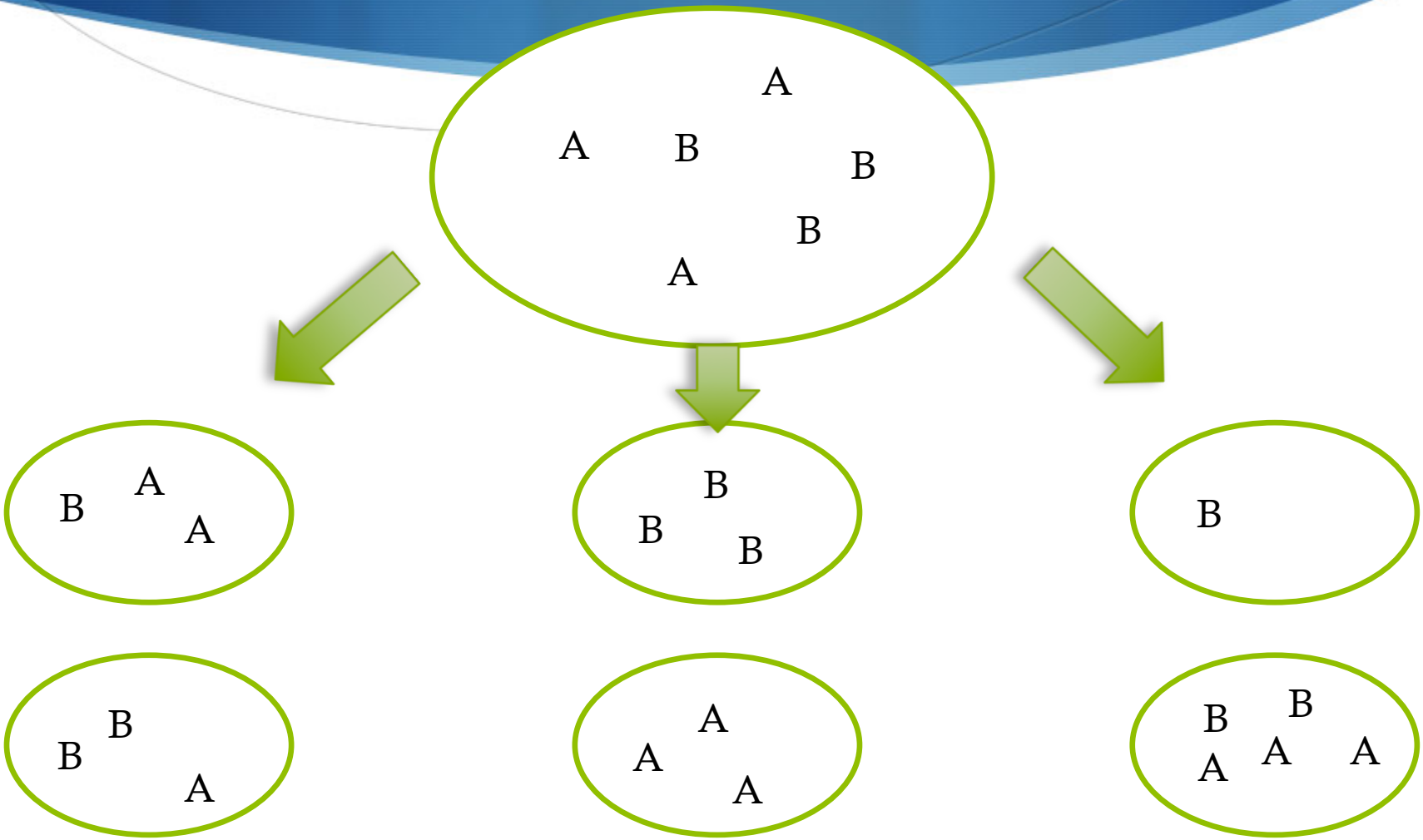
PLEASE PRINT THIS FLOWCHART OUT AND TAPE IT NEAR YOUR SCREEN. CONGRATULATIONS; YOU'RE NOW THE LOCAL COMPUTER EXPERT!

https://xkcd.com/627/

The Decision Tree

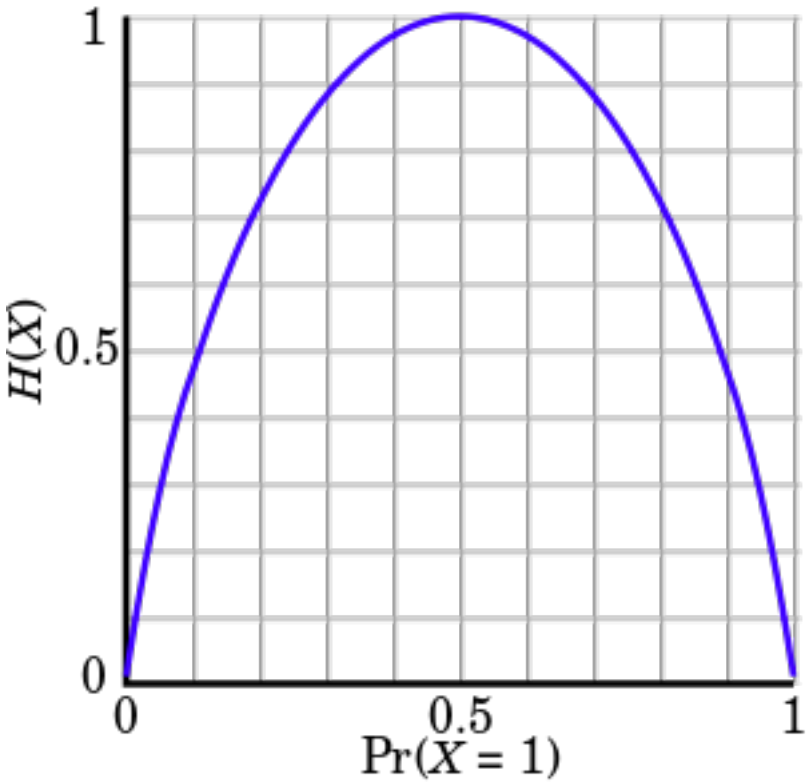
How to build it?

Finding the Best Attribute



A Measure of Purity (or rather, impurity)

$$H(S) = -\sum_{x \in X} p(x) \log_2 p(x)$$

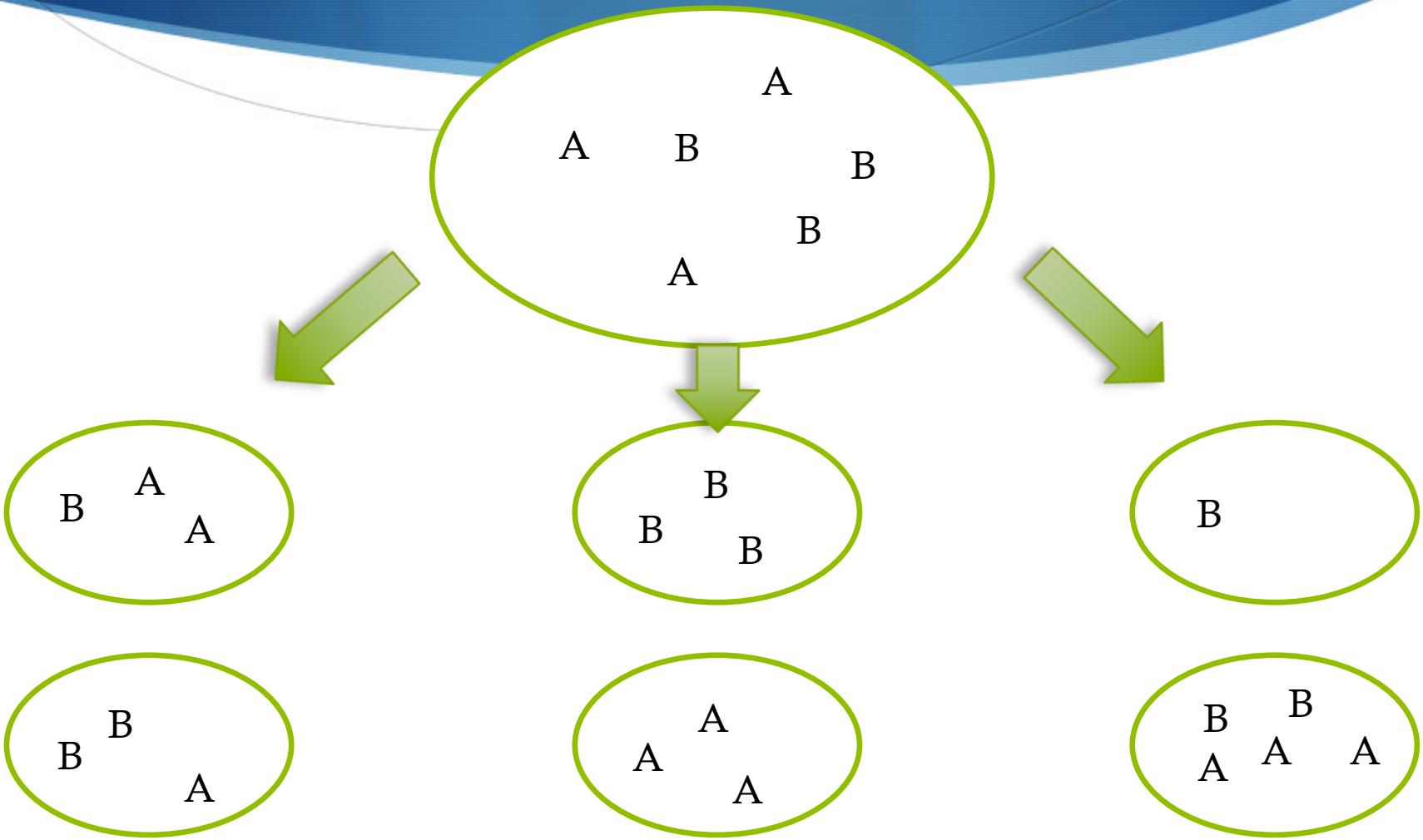


A Measure of Purity (or rather, impurity)

$$H(S) = -\sum_{x \in X} p(x) \log_2 p(x)$$

x1	x2	x3	class
1	2	3	A
1	4	5	A
2	2	5	B
5	2	1	B
3	5	1	A

Information Gain



The ID3 Algorithm

If all examples have the same label
    return a leaf with that label
Else if there are no features left to test
    return a leaf with the most common label
Else
    Consider each available feature
    Choose the one that maximizes information gain
    Create a new node for that feature

    For each possible value of the feature
        Create a branch for this value
        Create a subset of the examples for each branch
        Recursively call the function to create a new node at that branch

ID3 Demo

Row	Type	Plot	Star Actors	Profit
1	Comedy	Deep	Yes	Low
2	Comedy	Shallow	Yes	High
3	Drama	Deep	Yes	High
4	Drama	Shallow	No	Low
5	Comedy	Deep	No	High
6	Comedy	Shallow	No	High
7	Drama	Deep	No	Low

Let's Build a Tree

Let's Build a Tree

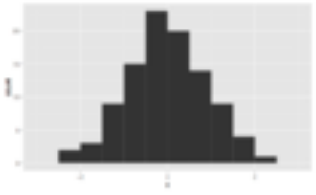
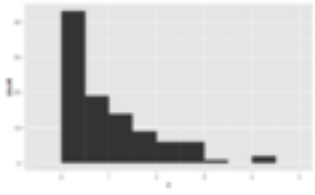
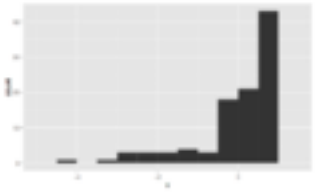
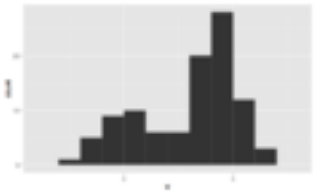
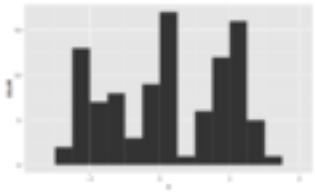
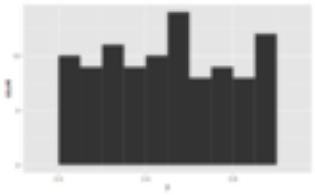
Row #	Credit Score	Income	Collateral	Job History	Should Loan
1	Good	High	Good	Long	Yes
2	Good	High	Poor	Short	No
3	Good	Low	Good	Long	Yes
4	Good	Low	Poor	Long	No
5	Average	High	Good	Long	Yes
6	Average	Low	Poor	Long	No
7	Average	Low	Poor	Short	No
8	Average	High	Poor	Long	Yes
9	Average	Low	Good	Long	No
10	Low	High	Good	Long	Yes
11	Low	High	Poor	Long	No
12	Low	High	Good	Short	No
13	Low	Low	Poor	Long	No

Numeric Data

Temperature (°F)	Wind Speed (mph)	Precipitation	Activity
45	0	None	Road Bike
50	10	Rain	Road Bike
65	18	None	Road Bike
74	35	None	Mountain Bike
34	14	Snow	None
52	23	Rain	None
60	28	None	Mountain Bike

Discretizing

Different Distributions

Overfitting

An Overfit Tree

Pruning

Pre-pruning Ideas

Post-pruning (basic idea)

1. Build tree completely
2. Iterate through each node
    1. Replace the sub-tree with a leaf labeled with the most common classification
    2. Compare leaf node vs. sub-tree on a validation set
    3. Keep the one with better performance

Assignment: ID3 Assignment

Minimum Standard Requirements

- Implement the basic ID3 decision tree algorithm
- Handle nominal and numeric data
- Handle missing data
- Produce a textual view of your resulting tree
- Basic experimentation
  - Use the supplied datasets
  - 10-fold cross validation
  - Compare to existing implementations

Ideas for Above and Beyond

- Exploration of additional approaches to handle numeric and/or missing data
- Pruning
- Experimentation on many more datasets
- Implement a technique to handle splitting on multiple attributes in the same node
- Regression
- Any other ideas you have

Team Project Requirements

- Demonstrates understanding of the business needs / context of the data
- Demonstrates ability to handle non-trivial dataset
  - Size, complex features, missing data, redundant attributes, obtaining the data
- Demonstrates proper algorithm selection and application
- Discovered something interesting / of value
  - About the data?
  - About the algorithms and their limitations in this context?
  - Something actionable for a stakeholder?
- Demonstrates understanding of limitations of the solution and potential ethical issues
- Conversant in appropriate terminology and technologies