Decision Trees: Learning Goals

1. General framework of supervised learning
2. Decision trees for binary classification
3. Extensions

- Russell 18.2-18.3

Supervised Learning

- Recall that *Clustering* seeks to answer:

  What are the sub-classes of objects in the dataset?

  - i.e., Identify classes and put class labels on data objects
- Next, we focus on *Classification:*

  How are the sub-classes different from each other?

  - i.e., Identify characteristics that discriminate objects with different labels
- The latter is called *supervised* learning, requiring prior labeling of data

Supervised Learning (2)

- Let X be the input data matrix, N objects/rows with k attributes/columns
- There is a training set of examples:

$$(X_1, y_1), \ldots, (X_N, y_N)$$

where $y_{i,}$ called the label, is either {0,1} and is generated by an unknown function y = f(x)

- The goal is to identify a function h, called the hypothesis or model, that approximates the true function f

Evaluating the Hypothesis

- There is the accuracy of h over the training data: predicted $\hat{y}_i = h(X_i)$ vs the true y_i
- Accuracy is improved if the training set includes both positive (i.e., $y = 1$) vs negative examples (i.e., $y = 0$)
  - We will talk about the imbalance problem when the number of positive examples is drastically different from the number of negative examples
- Actually, we care *more* about the accuracy of predicting a new example than the accuracy over the training examples; this is called generalization accuracy (retrospective vs prospective)

Evaluating the Hypothesis (2)

- When the generalization accuracy is lost at the expense of accuracy over the training examples, it is called overfitting
  - We will discuss various ways to guard against overfitting, e.g., regularization
  - Complex hypothesis tends to overfit, as well taking more time to compute

Everyday Examples of Prediction

- A bank loan application
  - High risk vs low risk
  - Too conservative: bank loses customers
  - Too aggressive: bank loses money for customers who default
- (Auto) insurance: discount for "good" customers
- University admissions
- Matching making

Many Classification Approaches

- Decision trees
- Linear models, e.g., regression, ridge, lasso
- Neural Nets
- Naïve Bayes
- kNN
- SVM
- Ensembles
- Random forests
- Hidden Markov Models
- Conditional Random Fields

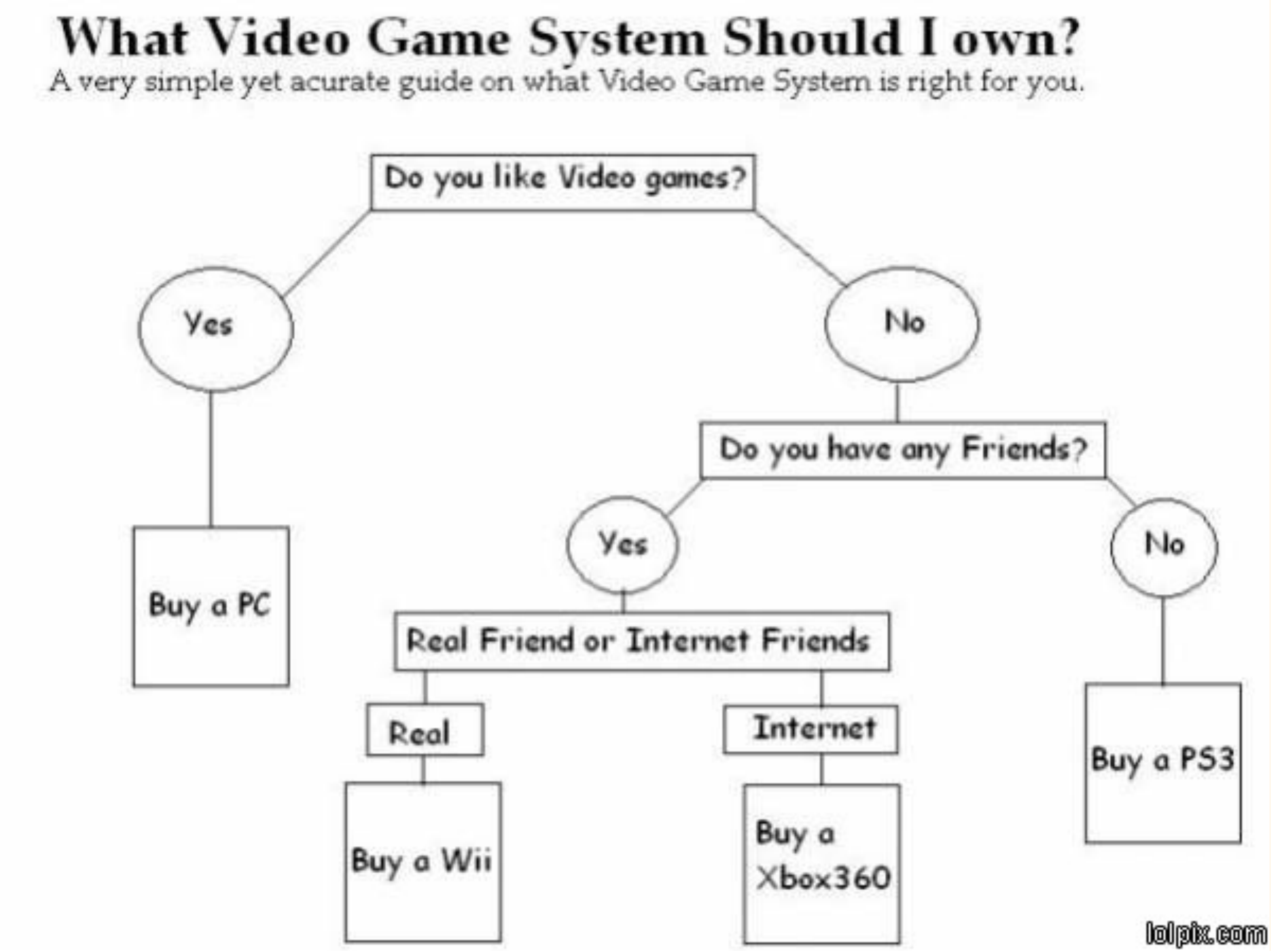
Decision Tree Structure

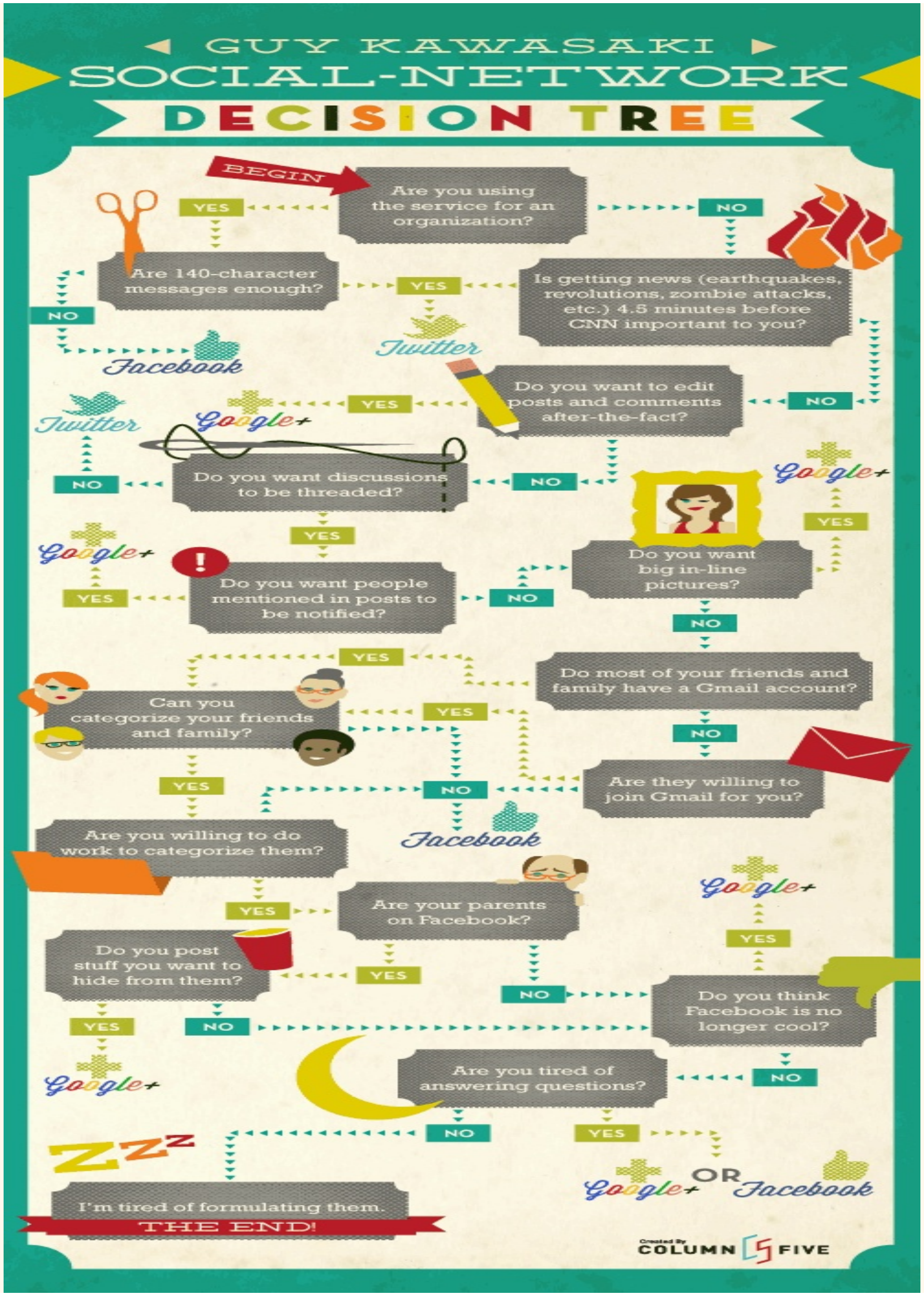
- Each node in the tree is of the form:

  attr in range

- Binary decision: if the node is internal, there are two children nodes
- An attribute can appear multiple times
- Thus, binary decision is not a restriction, as an n-ary decision can be represented as multiple binary decisions (e.g., Fig 18.2)
- If the node is external, it specifies a label value

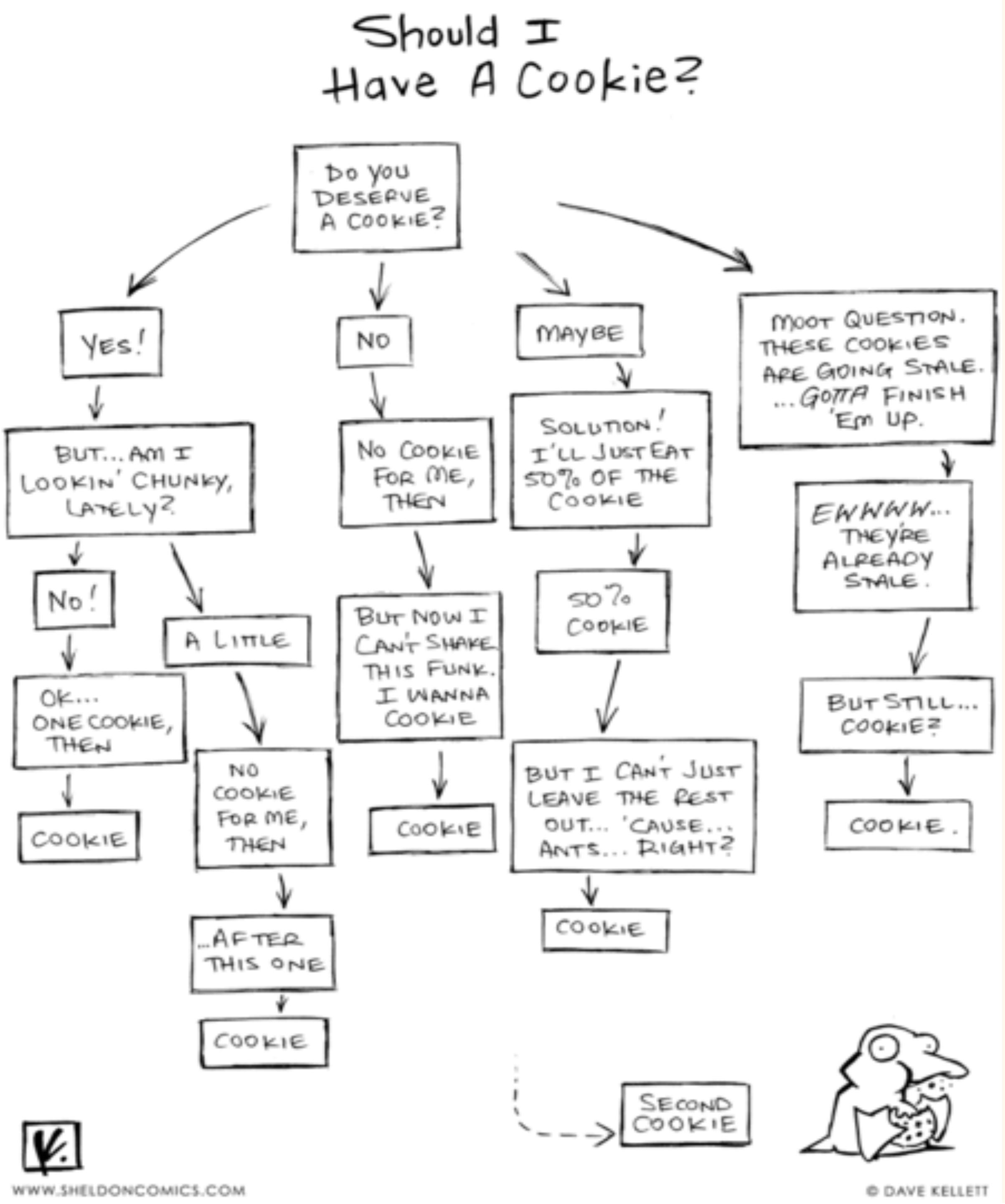
Funny Decision Tree (forums.beyond.ca)


What Video Game System Should I own?
A very simple yet acurate guide on what Video Game System is right for you.
Do you like Video games?
Yes
No
Buy a PC
Do you have any Friends?
Yes
No
Real Friend or Internet Friends
Real
Internet
Buy a Wii
Buy a Xbox360
Buy a PS3
lolpix.com


GUY KAWASAKI
SOCIAL-NETWORK
DECISION TREE
BEGIN
Are you using the service for an organization?
YES
NO
Are 140-character messages enough?
Is getting news (earthquakes, revolutions, zombie attacks, etc.) 4.5 minutes before CNN important to you?
YES
NO
Facebook
Twitter
Do you want to edit posts and comments after-the-fact?
YES
NO
Google+
Twitter
Do you want discussions to be threaded?
NO
NO
YES
Google+
Do you want big in-line pictures?
YES
Google+
Do you want people mentioned in posts to be notified?
YES
NO
NO
YES
Do most of your friends and family have a Gmail account?
Can you categorize your friends and family?
YES
NO
YES
NO
Are they willing to join Gmail for you?
Facebook
Are you willing to do work to categorize them?
YES
Are your parents on Facebook?
Google+
YES
Do you post stuff you want to hide from them?
YES
NO
YES
NO
Do you think Facebook is no longer cool?
Google+
Are you tired of answering questions?
NO
NO
YES
ZZZ
I'm tired of formulating them.
THE END!
Google+ OR Facebook
Created By COLUMN FIVE

(from pinterest.com)

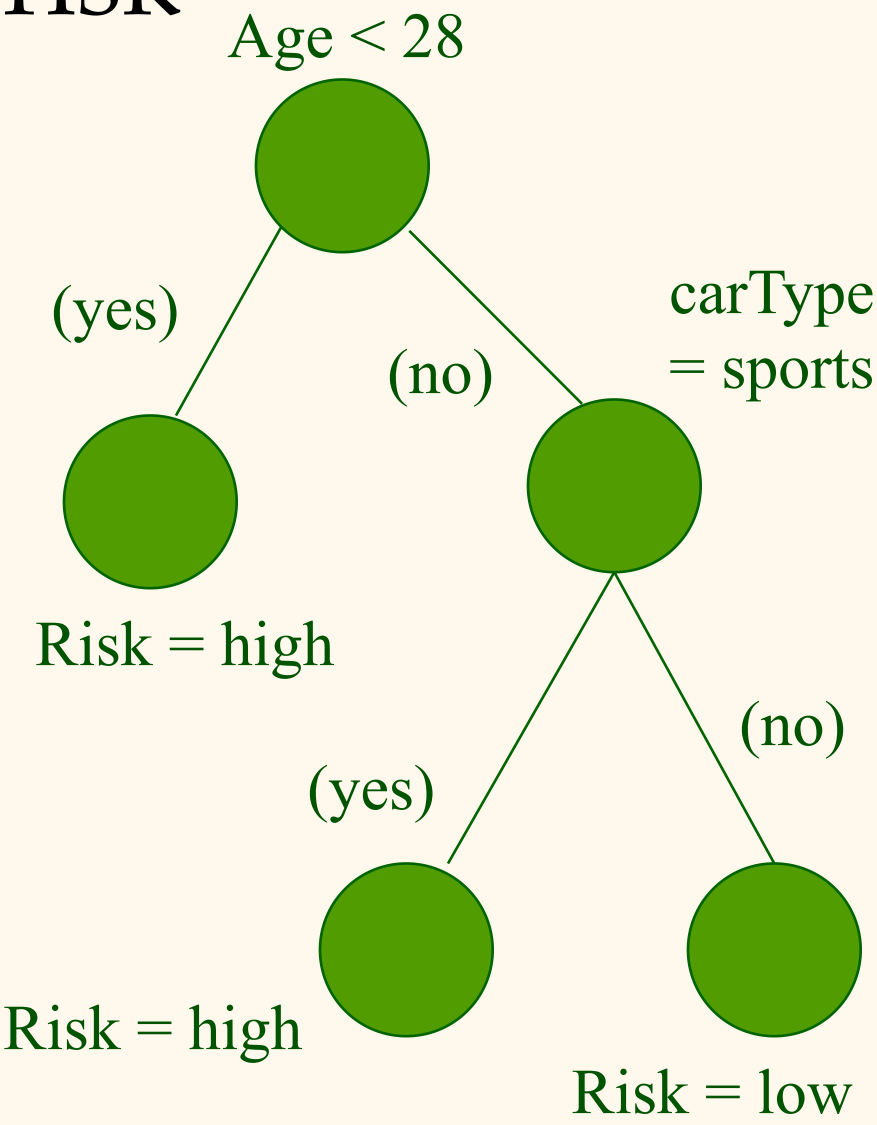
Is this a decision tree? More a deception tree! (from Sheldon Comics)


Should I Have A Cookie?
Do you deserve a cookie?
Yes!
But... am I lookin' chunky, lately?
No!
Ok... one cookie, then
Cookie
A little
No cookie for me, then
...After this one
Cookie
No
No cookie for me, then
But now I can't shake this funk. I wanna cookie
Cookie
Maybe
Solution! I'll just eat 50% of the cookie
50% cookie
But I can't just leave the rest out... 'cause... ants... right?
Cookie
Moot question. These cookies are going stale. ...Gotta finish 'em up.
Ewwww... they're already stale.
But still... cookie?
Cookie.
Second cookie
WWW.SHELDONCOMICS.COM
© DAVE KELLETT

Simple Example Predicting Auto insurance risk

Age	CarType	Risk
23	Family	High
17	Sports	High
43	Sports	High
68	Family	Low
32	Truck	Low
20	Family	High

Notice the mixing of numeric and categorical attrs



Age < 28

(yes)

(no)

Risk = high

carType = sports

(yes)

(no)

Risk = high

Risk = low

Clicker Question

- What is the accuracy of the tree on the training data?

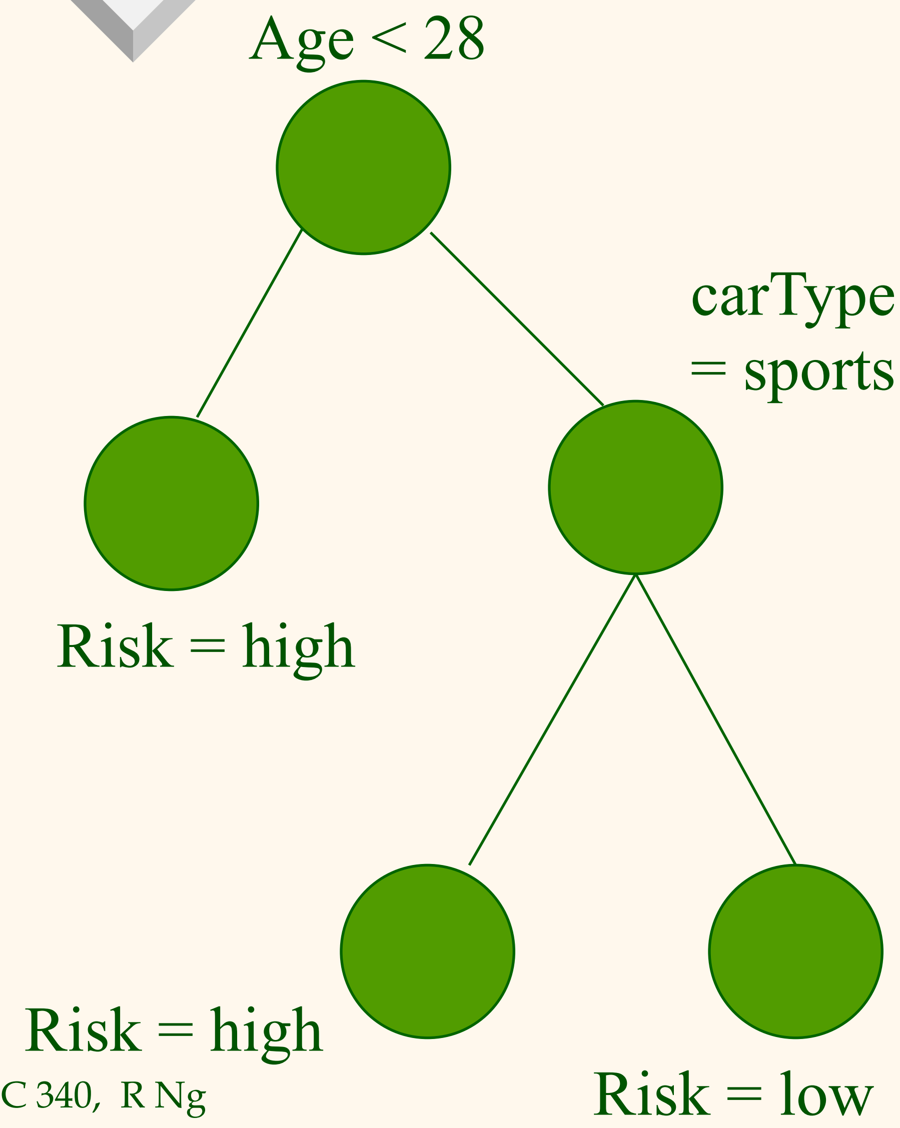
a) Perfect for high risk but not perfect for low risk.

b) Perfect for low risk but not perfect for high risk.

c) Imperfect for both low and high risks.

d) Perfect for both low and high risks.

Simple Test on Generalization



Age < 28

carType = sports

Risk = high

Risk = high

Risk = low

Age	CarType	Risk
29	Family	High
45	Sports	Low
22	Truck	High

Clicker Question

- What is the accuracy on the test data?

a) 0/3

b) 1/3

c) 2/3

d) 3/3

Building a Decision Tree

- Start with a single node with all the training examples
- If the node is *pure (i.e., unique label value),* done; annotate the node with the label value
- Else perform attribute selection based on an impurity function
- Select the attribute that is the best to split, and recurse

Impurity functions

- Entropy:
  - Compute the entropy, $- p \log_2 p$, if the attribute is selected
  - Compute the entropy if not selected
  - *Gain* = reduction in entropy
- Other variations exist
  - Gini index
    - Gini(S) = 1 - $\sum (p_j * p_j)$, where p_j is the probability of class j in S
    - Gini-split(S) proportional to gini(S_1) and gini (S_2)
  - Chi-squared statistic

Choosing the Best Split Point

- If age ≤ 17, two subsets $D_1 = \{r_2\}$, D_2 = rest
  - $Ent(D_1) = -[1 \log 1 + 0 \log 0] = 0$ *[* 0 log 0 defined as 0 *]*
  - $Ent(D_2) = -[\ 0.6 \log 0.6 + 0.4 \log 0.4] = 0.97$
  - $Ent(D) = -\ [0.67 \log 0.67 + 0.33 \log 0.33] = 0.92$
  - Entropy reduction = $Ent(D) - [1/6 *Ent(D_1) + 5/6*Ent(D_2)]= 0.11$

Age	CarType	Risk
23	Family	High
17	Sports	**High**
43	Sports	High
68	Family	Low
32	Truck	Low
20	Family	High

Age	CarType	Risk
23	Family	**High**
17	Sports	High
43	Sports	**High**
68	Family	**Low**
32	Truck	**Low**
20	Family	**High**

Age	CarType	Risk
23	Family	High
17	Sports	High
43	Sports	High
68	Family	Low
32	Truck	Low
20	Family	High

Choosing the Best Split Point

- If age ≤ 20, two subsets D_1 = {r_2,r_6}, D_2 = rest
  - Ent(D_1) = -[1 log 1 + 0 log 0] = 0
  - Ent(D_2) = -[0.5 log 0.5 + 0.5 log 0.5] = 1
  - Ent(D) = 0.92
  - Entropy reduction = Ent(D) – [2/6* Ent(D_1)+4/6*Ent(D_2)]= 0.25

Age	CarType	Risk
23	Family	High
17	Sports	**High**
43	Sports	High
68	Family	Low
32	Truck	Low
20	Family	**High**

Age	CarType	Risk
23	Family	**High**
17	Sports	High
43	Sports	**High**
68	Family	**Low**
32	Truck	**Low**
20	Family	High

Choosing the Best Split Point (2)

- If age ≤ 23, two subsets $D_1 = \{r_1, r_2, r_6\}$, D_2 = rest
  - $\text{Ent}(D_1) = 0$
  - $\text{Ent}(D_2) = -[\,0.67 \log 0.67 + 0.33 \log 0.33] = 0.92$
  - $\text{Ent}(D) = 0.92$
  - Entropy reduction = $\text{Ent}(D) - [3/6 * \text{Ent}(D_1) + 3/6 * \text{Ent}(D_2)] = 0.46$
- … for all possible splits based on Age
- (corrected!) If carType = family, two subsets $D_1 = \{r_1, r_4, r_6\}$, D_2 = rest
  - $\text{Ent}(D_1) = -[0.67 \log 0.67 + 0.33 \log 0.33] = 0.92$
  - $\text{Ent}(D_2) = -[0.67 \log 0.67 + 0.33 \log 0.33] = 0.92$
  - $\text{Ent}(D) = 0.92$
  - Entropy reduction = $\text{Ent}(D) - [3/6 * \text{Ent}(D_1) + 3/6 * \text{Ent}(D_2)] = 0$
- … for all possible values on carType

Continuing Example

- Age ≤ 23 turns out to reduce entropy the most, thus chosen as the split point
  - Using 23 as the threshold may overfit
  - Take the mean of 23 and 32 (next value), which is 28
- Left branch of the tree becomes pure, done
- Right branch consists of r_3, r_4, r_5
- Not pure: so recurse on those training examples

Age	CarType	Risk
23	Family	**High**
17	Sports	**High**
43	Sports	High
68	Family	**Low**
32	Truck	**Low**
20	Family	**High**

Age	CarType	Risk
23	Family	High
17	Sports	High
43	Sports	**High**
68	Family	**Low**
32	Truck	**Low**
20	Family	High

Re-doing Slide 20 using Gini Index

- If age ≤ 23, two subsets $D_1 = \{r_1, r_2, r_6\}$, D_2 = rest
  - Gini(D_1) = 1 – (1 * 1) – (0 * 0) = 0
  - Gini(D_2) = 1 – (0.67 * 0.67) – (0.33 * 0.33) = 0.44
  - Gini(D) = 1 – (0.67 * 0.67) – (0.33 * 0.33) = 0.44
  - **Gini** reduction = Gini(D) – [3/6 * Gini(D_1) + 3/6* Gini(D_2)]= 0.22
  - (actually, Gini(D) is fixed per node)

Age	CarType	Risk
23	Family	**High**
17	Sports	**High**
43	Sports	High
68	Family	Low
32	Truck	Low
20	Family	**High**

Age	CarType	Risk
23	Family	High
17	Sports	High
43	Sports	**High**
68	Family	**Low**
32	Truck	**Low**
20	Family	High

Stopping Criteria

- One criterion is clearly when the node is pure
- It is always possible to build a decision tree with all pure nodes
- But experience indicates that the accuracy for test data drops, i.e., overfitting
- One way is to apply early stopping
  - stop (i) after a maximum depth, or (ii) when the impurity gain is too small, or (iii) the number of training samples for a node is too small
  - If a leaf node is impure, then use majority voting to determine the label value

Tree Pruning

- Early stopping may stop too early when a combination of attributes leads to significant gain
- A better way, which is more expensive, is to prune a generated tree
- Start with parents of leaf nodes
- Identify irrelevant nodes to be pruned, i.e., replaced by the lead descendants

Tree Pruning (2)

- Use statistical testing to identify irrelevant nodes
- Test if the proportion of positive and negative examples in the subtree is different from the proportion in the training set
  - Chi-squared test,
  - Student T-test
- Experience shows that pruned trees perform better than unpruned trees when the training set contains a lot of noisy examples
- Pruned trees are smaller and easier to understand

Extension: Multi-class Classification

- So far we've discussed binary classification
- Can easily extend decision trees to m ≥ 3 sub-classes
  - Both entropy and gini index can deal with 3+ sub-classes
  - No change to choosing the best split point

Extension: Continuous-valued Label

- So far we’ve assumed that the label attribute is categorical, i.e., discrete values
- What if the label attribute is numeric, e.g., the salary, the price?
  - Put the numeric values into discrete bins and treat it as a binary or multi-class classification problem
  - Use regression tree when the leaf node is not a single label value, but a linear regression line (later)

Extension: Mixing Continuous and Categorical Attributes

- Actually, both the entropy and gini index are not affected – we are already doing it!
  - Age is numeric whereas carType categorical
  - Ubiquitous in everyday applications
- Contrast that with our previous discussion on clustering when in the context of distance functions, we've assumed numeric attributes
  - There are more sophisticated clustering algorithms that deal with mixed attributes

Extension: Missing Values

- For many everyday applications, data may have missing values
- N/A may mean: (i) value does not exist; (ii) value exists but unknown; (iii) not sure whether it is (i) or (ii)
- Solution 1: remove the examples with missing values
  - For some applications, this may remove a significant percentage of the training data
  - If the missing values predominantly occur to one sub-class but not the other, then this phenomenon is discriminating and valuable

Extension: Missing Values (2)

- Solution 2: remove the attributes with missing values
  - Very similar to the previous solution of removing examples containing missing values
- Solution 3: if the attribute is categorical, treat N/A as a special categorical value and proceed as usual; if the attribute is numeric, ignore the missing value/example and proceed as usual
- Solution 4: apply missing value imputation (later) – which can be treated as a prediction problem!

Concluding Remarks: Big Data

- (Binary) classification is arguably the most popular tool for data mining
- Choosing the best split points dominates decision tree computation
- That step is inherently very parallelizable
- Thus, fast parallel schemes for decision trees have been well developed