Middle Tennessee State University
College of Basic and Applied Sciences
Spring 2014

CSCI 7350: Data Mining
Professor: Dr. Cen Li

Homework 1
By: Zane Colgin

January 23, 2014
modified: February 4, 2014

NOTE: We treat $0 \cdot log_2(0) = 0$. All calculations carried through to the end at double floating point precision. Numbers rounded only for presentation.

Expected encoding for information in root of the tree:

$$\sum_{k=1}^{2} P(C_k) \cdot (-\log_2 P(C_k) =$$
$$-\frac{5}{7}\log_2(\frac{5}{7}) - \frac{2}{7}\log_2(\frac{2}{7}) \approx 0.8631$$

Level 1:

A) Use Attribute A_i = Cap Shape

- bell: $P(C_1) = 1$,$P(C_2) = 0$
- flat: $P(C_1) = \frac{4}{5}$,$P(C_2) = \frac{1}{5}$
- convex: $P(C_1) = \frac{1}{4}$,$P(C_2) = \frac{3}{4}$

Expected encoding:

$$\sum_{j=1}^{3} P(A_i = V_{ij}) \cdot \left[\sum_{k=1}^{2} P(C_k|A_i = V_{ij}) \cdot (-\log_2 P(C_k|A_i = V_{ij})\right] =$$

$$P(A_i = \text{bell}) \cdot \left[\sum_{k=1}^{2} P(C_k|A_i = \text{bell}) \cdot (-\log_2 P(C_k|A_i = \text{bell})\right] +$$
$$P(A_i = \text{flat}) \cdot \left[\sum_{k=1}^{2} P(C_k|A_i = \text{flat}) \cdot (-\log_2 P(C_k|A_i = \text{flat})\right] +$$
$$P(A_i = \text{convex}) \cdot \left[\sum_{k=1}^{2} P(C_k|A_i = \text{convex}) \cdot (-\log_2 P(C_k|A_i = \text{convex})\right] =$$

$$\frac{5}{14} \cdot [-(1)\log_2(1) - (0)\log_2(0)] +$$
$$\frac{5}{14} \cdot \left[-\frac{4}{5}\log_2(\frac{4}{5}) - \frac{1}{5}\log_2(\frac{1}{5})\right] +$$
$$\frac{2}{7} \cdot \left[-\frac{1}{4}\log_2(\frac{1}{4}) - \frac{3}{4}\log_2(\frac{3}{4})\right] \approx$$

$$\frac{5}{14} \cdot [0+0] + \frac{5}{14} \cdot [0.2575 + 0.4644] + \frac{2}{7} \cdot [0.5 + 0.3113] =$$

$$0 + 0.2578 + 0.2318 = 0.4896$$

Gain $\approx 0.8631 - 0.4896 = 0.3735$

B) Use Attribute A_i = Cap Color

- brown: $P(C_1) = \frac{5}{6}$,$P(C_2) = \frac{1}{6}$
- grey: $P(C_1) = \frac{5}{8}$,$P(C_2) = \frac{3}{8}$

Expected encoding:

$$\sum_{j=1}^{2} P(A_i = V_{ij}) \cdot \left[\sum_{k=1}^{2} P(C_k|A_i = V_{ij}) \cdot (-\log_2 P(C_k|A_i = V_{ij}) \right] =$$

$$P(A_i = \text{brown}) \cdot \left[\sum_{k=1}^{2} P(C_k|A_i = \text{brown}) \cdot (-\log_2 P(C_k|A_i = \text{brown}) \right] +$$

$$P(A_i = \text{grey}) \cdot \left[\sum_{k=1}^{2} P(C_k|A_i = \text{grey}) \cdot (-\log_2 P(C_k|A_i = \text{grey}) \right] =$$

$$\frac{3}{7} \cdot \left[-\frac{5}{6} \log_2(\frac{5}{6}) - \frac{1}{6} \log_2(\frac{1}{6}) \right] +$$

$$\frac{4}{7} \cdot \left[-\frac{5}{8} \log_2(\frac{5}{8}) - \frac{3}{8} \log_2(\frac{3}{8}) \right] \approx$$

$$\frac{3}{7} \cdot [0.2192 + 0.4308] + \frac{4}{7} \cdot [0.4238 + 0.5306] =$$

$$0.2786 + 0.5454 = 0.8240$$

Gain $\approx 0.8631 - 0.8240 = 0.0391$

C) Use Attribute A_i = Odor

- almond: $P(C_1) = 1$,$P(C_2) = 0$
- spicy: $P(C_1) = \frac{1}{2}$,$P(C_2) = \frac{1}{2}$
- foul: $P(C_1) = 0$,$P(C_2) = 1$

Expected encoding:

$$\sum_{j=1}^{3} P(A_i = V_{ij}) \cdot \left[\sum_{k=1}^{2} P(C_k|A_i = V_{ij}) \cdot (-\log_2 P(C_k|A_i = V_{ij})\right] =$$

$$P(A_i = \text{almond}) \cdot \left[\sum_{k=1}^{2} P(C_k|A_i = \text{almond}) \cdot (-\log_2 P(C_k|A_i = \text{almond})\right] +$$

$$P(A_i = \text{spicy}) \cdot \left[\sum_{k=1}^{2} P(C_k|A_i = \text{spicy}) \cdot (-\log_2 P(C_k|A_i = \text{spicy})\right] +$$

$$P(A_i = \text{foul}) \cdot \left[\sum_{k=1}^{2} P(C_k|A_i = \text{foul}) \cdot (-\log_2 P(C_k|A_i = \text{foul})\right] =$$

$$\frac{4}{7} \cdot [-(1)\log_2(1) - (0)\log_2(0)] +$$

$$\frac{2}{7} \cdot \left[-\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2})\right] +$$

$$\frac{1}{7} \cdot [-(0)\log_2(0) - (1)\log_2(1)] \approx$$

$$\frac{4}{7} \cdot [0 + 0] + \frac{2}{7} \cdot [0.5 + 0.5] + \frac{1}{7} \cdot [0 + 0] =$$

$$0 + 0.2857 + 0 = 0.2857$$

Gain $\approx 0.8631 - 0.2857 = 0.5774$

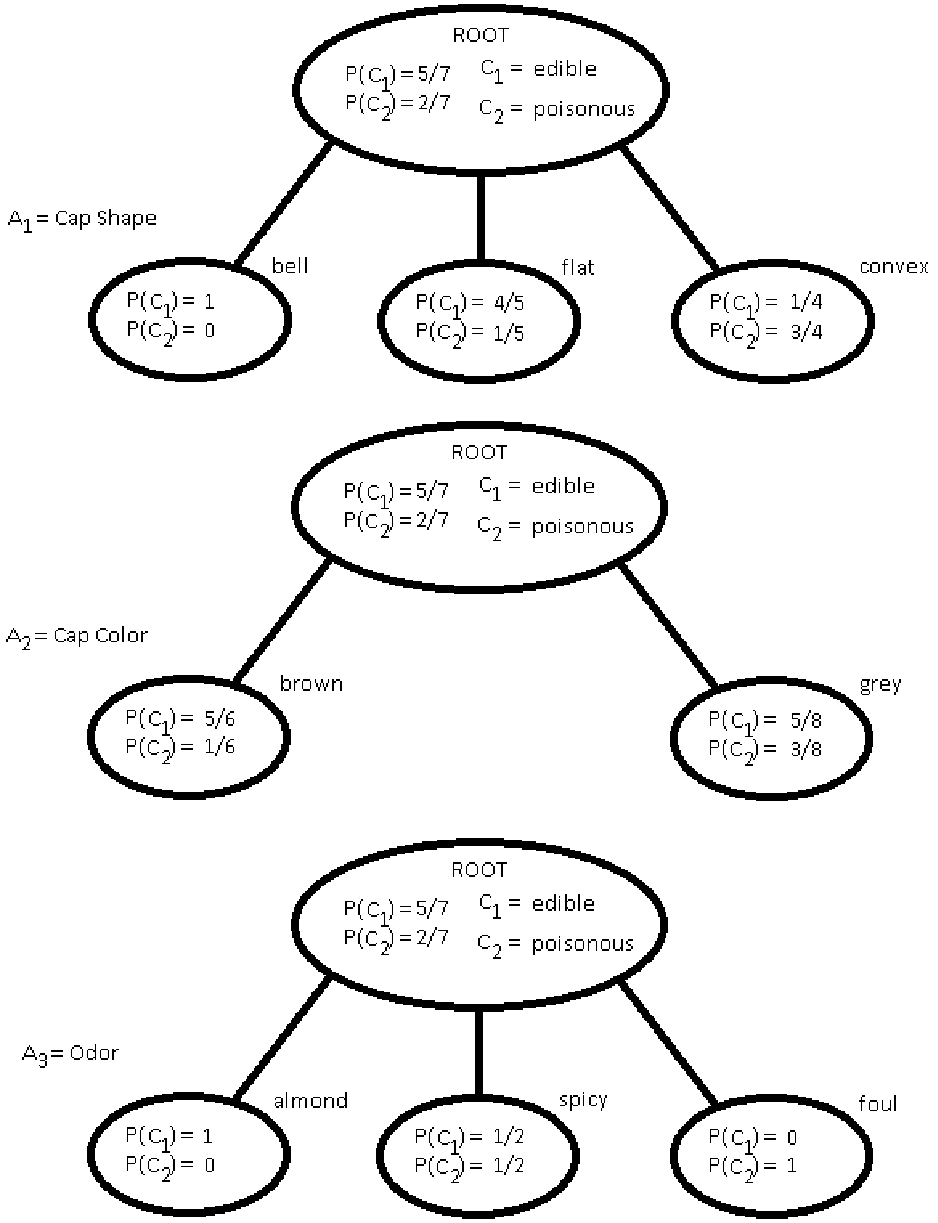

Figure 1: partial decision tree

```
% hw1.m
% CSCI 7350 - Homework 1
% Professor: Cen Li
%
% AUTHOR:
%       Zane Colgin
%       Middle Tennessee State University
%       January 2014
%
% NOTE:
%       "!!!" in comments indicates programming note
%       i.e. look here for debugging notes, optimizations

clear all; close all; clc; %format rational
% ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
% ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
%% ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ INITIALIZE / DATA
div='~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ ';
halfdiv='~~~~~~~~~~~~~~~~~~~~~~ ';
% Cap Shape: bell, flat, or convex
% Cap Color: brown, grey
% Odor: almond, spicy, foul

C = {'edible','poisonous'};
A = { 'CapShape'; 'CapColor'; 'Odor' };

CapShapes = { 'bell'; 'flat'; 'convex' };
CapColors = { 'brown'; 'grey' };
Odors = { 'almond'; 'spicy'; 'foul' };
V = { CapShapes, CapColors, Odors };
clear CapShapes CapColors Odors
className = 'Class';

% Data:
% Object    Cap Shape   Cap color   Odor        class

% X1        bell        brown       almond      edible
% X2        flat        grey        almond      edible
% X3        convex      grey        spicy       poisonous
% X4        bell        brown       almond      edible
% X5        flat        grey        almond      edible
% X6        flat        grey        spicy       edible
% X7        convex      grey        almond      edible
% X8        bell        brown       almond      edible
% X9        convex      brown       foul        poisonous
% X10       bell        brown       spicy       edible
```

```
% X11       bell        grey        almond      edible
% X12       convex      grey        spicy       poisonous
% X13       flat        brown       almond      edible
% X14       flat        grey        foul        poisonous

% allocate X by initializing last date object first
X(14)=struct('CapShape','flat','CapColor','grey',...
    'Odor','foul','Class','poisonous');
X(1)=struct('CapShape','bell','CapColor','brown',...
    'Odor','almond','Class','edible');
X(2)=struct('CapShape','flat','CapColor','grey',...
    'Odor','almond','Class','edible');
X(3)=struct('CapShape','convex','CapColor','grey',...
    'Odor','spicy','Class','poisonous');
X(4)=struct('CapShape','bell','CapColor','brown',...
    'Odor','almond','Class','edible');
X(5)=struct('CapShape','flat','CapColor','grey',...
    'Odor','almond','Class','edible');
X(6)=struct('CapShape','flat','CapColor','grey',...
    'Odor','spicy','Class','edible');
X(7)=struct('CapShape','convex','CapColor','grey',...
    'Odor','almond','Class','edible');
X(8)=struct('CapShape','bell','CapColor','brown',...
    'Odor','almond','Class','edible');
X(9)=struct('CapShape','convex','CapColor','brown',...
    'Odor','foul','Class','poisonous');
X(10)=struct('CapShape','bell','CapColor','brown',...
    'Odor','spicy','Class','edible');
X(11)=struct('CapShape','bell','CapColor','grey',...
    'Odor','almond','Class','edible');
X(12)=struct('CapShape','convex','CapColor','grey',...
    'Odor','spicy','Class','poisonous');
X(13)=struct('CapShape','flat','CapColor','brown',...
    'Odor','almond','Class','edible');


% ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
% ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
%% ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ ROOT
% Expected encoding for information in root of the tree:
P0 = P(X,C,className);
H0 = H(P0);

disp('Root:');
for i=1:length(C)
    fprintf('%s\t%s\n',strtrim(rats(P0(i))), C{i});
end
fprintf('Gain:   %f\n',H0);
```

```
% ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
% ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
%% ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ LEVEL 1
Child_Node(1) = X(1); % define array type

for k=1:length(A)       % for each Attribute, create child node k
    fprintf(’%s\n’,[div,A{k}]);
    T = P(X,V{k},A{k})
    Hsum = 0;
    for j=1:length(V{k})% for each Attribute Type for current Attribute k

        size = 0;       % initialize size of current node for Attribute k

        % for each element in the root, find if belongs in node k
        for i=1:length(X)
            if (strcmp(X(i).(A{k}),V{k}{j}))
                size = size+1;
                Child_Node(size) = X(i);
            end
        end

        P1 = P(Child_Node(1:size),C,className);
        Hnode = H(P1)

        disp(V{k}{j});
        for i=1:length(C)
            fprintf(’%s\t%s\n’,strtrim(rats(P1(i))), C{i});
        end
        disp(halfdiv);
        Hsum = Hsum + T(j)*Hnode;

    end

    fprintf(’Gain:   %f\n’,H0 - Hsum);
end
```

```
function out = H(P)
n = length(P);      % number of items
    out = 0;
    for i=1:n % number of items
        partial = -P(i)*log2(P(i))
        if(P(i)~=0)
            out = out + partial;
        end
    end
end


function out = P(S,C,structFieldName)
n = length(S);      % number of items
m = length(C);      % number of classes
out = zeros(m,1);

    for i=1:n % number of items
        for j=1:m % number of classes
            if (strcmp(S(i).(structFieldName),C(j)))
                out(j) = out(j) + 1;
                break;
            end
        end
    end

out = out./n;
end
```

```
partial =
    0.3467
partial =
    0.5164
Root:
5/7 edible
2/7 poisonous
Gain:   0.863121
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ CapShape
T =
    0.3571
    0.3571
    0.2857
partial =
     0
Hnode =
     0
bell
1 edible
0 poisonous
~~~~~~~~~~~~~~~~~~~~~~
partial =
    0.2575
partial =
    0.4644
Hnode =
    0.7219
flat
4/5 edible
1/5 poisonous
~~~~~~~~~~~~~~~~~~~~~~
partial =
    0.5000
partial =
    0.3113
Hnode =
    0.8113
convex
1/4 edible
3/4 poisonous
~~~~~~~~~~~~~~~~~~~~~~
Gain:   0.373495
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ CapColor
T =
    0.4286
    0.5714
```

```
partial =
    0.2192
partial =
    0.4308
Hnode =
    0.6500
brown
5/6 edible
1/6 poisonous
~~~~~~~~~~~~~~~~~~~~~~
partial =
    0.4238
partial =
    0.5306
Hnode =
    0.9544
grey
5/8 edible
3/8 poisonous
~~~~~~~~~~~~~~~~~~~~~~
Gain:   0.039149
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ Odor
T =
    0.5714
    0.2857
    0.1429
partial =
     0
Hnode =
     0
almond
1 edible
0 poisonous
~~~~~~~~~~~~~~~~~~~~~~
partial =
    0.5000
partial =
    0.5000
Hnode =
     1
spicy
1/2 edible
1/2 poisonous
~~~~~~~~~~~~~~~~~~~~~~
partial =
     0
Hnode =
     0
foul
```

```
0 edible
1 poisonous
~~~~~~~~~~~~~~~~~~~~~~
Gain:   0.577406
```