Assignment 4 - Decision Tree

Keith G. Williams - 800690755

Friday, June 12, 2015

Problem Description

A company sent out some promotion to various houses and recorded a few facts about each house and also whether the people responded or not. Please create a Decision Tree (similar to one discussed in class) for the dataset below.

```
##    district    house_type income previous_customer   outcome
## 1  suburban      detached   high                no   nothing
## 2  suburban      detached   high               yes   nothing
## 3     rural      detached   high                no responded
## 4     urban semi-detached   high                no responded
## 5     urban semi-detached    low                no responded
## 6     urban semi-detached    low               yes   nothing
## 7     rural semi-detached    low               yes responded
## 8  suburban       terrace   high                no   nothing
## 9  suburban semi-detached    low                no responded
## 10    urban       terrace    low                no responded
## 11 suburban       terrace    low               yes responded
## 12    rural       terrace   high               yes responded
## 13    rural      detached    low                no responded
## 14    urban       terrace   high               yes   nothing
```

Information Content *I(C;F)* Calculations

Each node will be split based on the information content calculation:

$$I(C;F) = \sum_{i=1}^{m}\sum_{j=1}^{d} P\,(C = c_i, F = f_j) log_2 \frac{P(C = c_i, F = f_j)}{P(C = c_i)P(F = f_j)}$$

where C is the class (in this cases `outcome`), and F is the feature matrix (in this case `district`, `house_type`, `income`, `previous_customer`)

Root Node

$I(outcome, district) =$

$$P(outcome = nothing, district = suburban) log_2 \frac{P(outcome = nothing, district = suburban)}{P(outcome = nothing)P(district = suburban)}$$
$$+P(outcome = nothing, district = rural) log_2 \frac{P(outcome = nothing, district = rural)}{P(outcome = nothing)P(district = rural)}$$
$$+P(outcome = nothing, district = urban) log_2 \frac{P(outcome = nothing, district = urban)}{P(outcome = nothing)P(district = urban)}$$
$$+P(outcome = responded, district = suburban) log_2 \frac{P(outcome = responded, district = suburban)}{P(outcome = responded)P(district = suburban)}$$
$$+P(outcome = responded, district = rural) log_2 \frac{P(outcome = responded, district = rural)}{P(outcome = responded)P(district = rural)}$$
$$+P(outcome = responded, district = urban) log_2 \frac{P(outcome = responded, district = urban)}{P(outcome = responded)P(district = urban)}$$
$$= \frac{3}{14} log_2 \frac{\frac{3}{14}}{\frac{5}{14}\frac{5}{14}} + \frac{0}{14} log_2 \frac{\frac{0}{14}}{\frac{5}{14}\frac{4}{14}} + \frac{2}{14} log_2 \frac{\frac{2}{14}}{\frac{5}{14}\frac{5}{14}} + \frac{2}{14} log_2 \frac{\frac{2}{14}}{\frac{9}{14}\frac{5}{14}} + \frac{4}{14} log_2 \frac{\frac{4}{14}}{\frac{9}{14}\frac{4}{14}} + \frac{3}{14} log_2 \frac{\frac{3}{14}}{\frac{9}{14}\frac{5}{14}}$$
$$= \mathbf{0.247}$$

$I(outcome, housetype) =$

$$P(outcome = nothing, housetype = detached) log_2 \frac{P(outcome = nothing, housetype = detached)}{P(outcome = nothing)P(housetype = detached)}$$
$$+P(outcome = nothing, housetype = semidetached) log_2 \frac{P(outcome = nothing, housetype = semidetached)}{P(outcome = nothing)P(housetype = semidetached)}$$
$$+P(outcome = nothing, housetype = terrace) log_2 \frac{P(outcome = nothing, housetype = terrace)}{P(outcome = nothing)P(housetype = terrace)}$$
$$+P(outcome = responded, housetype = detached) log_2 \frac{P(outcome = responded, housetype = detached)}{P(outcome = responded)P(housetype = detached)}$$
$$+P(outcome = responded, housetype = semidetached) log_2 \frac{P(outcome = responded, housetype = semidetached)}{P(outcome = responded)P(housetype = semidetached)}$$
$$+P(outcome = responded, housetype = terrace) log_2 \frac{P(outcome = responded, housetype = terrace)}{P(outcome = responded)P(housetype = terrace)}$$
$$= \frac{2}{14} log_2 \frac{\frac{2}{14}}{\frac{5}{14}\frac{4}{14}} + \frac{1}{14} log_2 \frac{\frac{1}{14}}{\frac{5}{14}\frac{5}{14}} + \frac{2}{14} log_2 \frac{\frac{2}{14}}{\frac{5}{14}\frac{5}{14}} + \frac{2}{14} log_2 \frac{\frac{2}{14}}{\frac{9}{14}\frac{4}{14}} + \frac{4}{14} log_2 \frac{\frac{4}{14}}{\frac{9}{14}\frac{5}{14}} + \frac{3}{14} log_2 \frac{\frac{3}{14}}{\frac{9}{14}\frac{5}{14}}$$
$$= \mathbf{0.050}$$

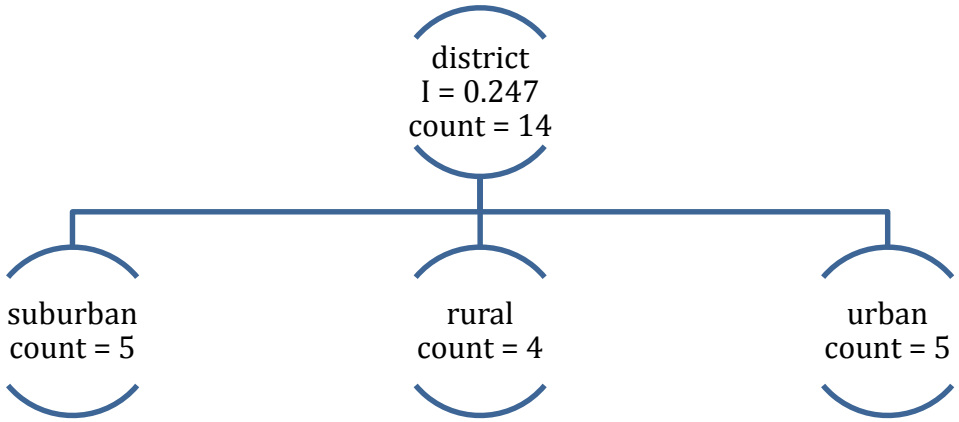
$I(outcome, income) =$

$$P(outcome = nothing, income = high) log_2 \frac{P(outcome = nothing, income = high)}{P(outcome = nothing)P(income = high)}$$
$$+P(outcome = nothing, income = low) log_2 \frac{P(outcome = nothing, income = low)}{P(outcome = nothing)P(income = low)}$$
$$+P(outcome = responded, income = high) log_2 \frac{P(outcome = responded, income = high)}{P(outcome = responded)P(income = high)}$$
$$+P(outcome = responded, income = low) log_2 \frac{P(outcome = responded, income = low)}{P(outcome = responded)P(income = low)}$$
$$= \frac{4}{14} log_2 \frac{\frac{4}{14}}{\frac{5}{14}\frac{7}{14}} + \frac{1}{14} log_2 \frac{\frac{1}{14}}{\frac{5}{14}\frac{7}{14}} + \frac{3}{14} log_2 \frac{\frac{3}{14}}{\frac{9}{14}\frac{7}{14}} + \frac{6}{14} log_2 \frac{\frac{6}{14}}{\frac{9}{14}\frac{7}{14}}$$
$$= \mathbf{0.152}$$

$I(outcome, previouscustomer) =$

$$P(outcome = nothing, previouscustomer = no) log_2 \frac{P(outcome = nothing, previouscustomer = no)}{P(outcome = nothing)P(previouscustomer = no)}$$
$$+P(outcome = nothing, previouscustomer = yes) log_2 \frac{P(outcome = nothing, previouscustomer = yes)}{P(outcome = nothing)P(previouscustomer = yes)}$$
$$+P(outcome = responded, previouscustomer = no) log_2 \frac{P(outcome = responded, previouscustomer = no)}{P(outcome = responded)P(previouscustomer = no)}$$
$$+P(outcome = responded, previouscustomer = yes) log_2 \frac{P(outcome = responded, previouscustomer = yes)}{P(outcome = responded)P(previouscustomer = yes)}$$
$$= \frac{2}{14} log_2 \frac{\frac{2}{14}}{\frac{5}{14}\frac{8}{14}} + \frac{3}{14} log_2 \frac{\frac{3}{14}}{\frac{5}{14}\frac{6}{14}} + \frac{6}{14} log_2 \frac{\frac{6}{14}}{\frac{9}{14}\frac{8}{14}} + \frac{3}{14} log_2 \frac{\frac{3}{14}}{\frac{9}{14}\frac{6}{14}}$$
$$= \mathbf{0.048}$$

The greatest information content is $I(outcome, district) = 0.247$, so the first internal node will split `outcome` three ways on `district`.



Second Node Layer

- `district == suburban`

```
##   district    house_type income previous_customer   outcome
## 1 suburban      detached   high                no   nothing
## 2 suburban      detached   high               yes   nothing
## 3 suburban       terrace   high                no   nothing
## 4 suburban semi-detached    low                no responded
## 5 suburban       terrace    low               yes responded
```

$I(outcome, housetype \mid district = suburban) =$

$$P(outcome = nothing, housetype = detached) log_2 \frac{P(outcome = nothing, housetype = detached)}{P(outcome = nothing)P(housetype = detached)}$$
$$+P(outcome = nothing, housetype = semidetached) log_2 \frac{P(outcome = nothing, housetype = semidetached)}{P(outcome = nothing)P(housetype = semidetached)}$$
$$+P(outcome = nothing, housetype = terrace) log_2 \frac{P(outcome = nothing, housetype = terrace)}{P(outcome = nothing)P(housetype = terrace)}$$
$$+P(outcome = responded, housetype = detached) log_2 \frac{P(outcome = responded, housetype = detached)}{P(outcome = responded)P(housetype = detached)}$$
$$+P(outcome = responded, housetype = semidetached) log_2 \frac{P(outcome = responded, housetype = semidetached)}{P(outcome = responded)P(housetype = semidetached)}$$
$$+P(outcome = responded, housetype = terrace) log_2 \frac{P(outcome = responded, housetype = terrace)}{P(outcome = responded)P(housetype = terrace)}$$
$$= \frac{2}{5} log_2 \frac{\frac{2}{5}}{\frac{3}{5}\frac{2}{5}} + \frac{0}{5} log_2 \frac{\frac{0}{5}}{\frac{3}{5}\frac{1}{5}} + \frac{1}{5} log_2 \frac{\frac{1}{5}}{\frac{3}{5}\frac{2}{5}} + \frac{0}{5} log_2 \frac{\frac{0}{5}}{\frac{2}{5}\frac{2}{5}} + \frac{1}{5} log_2 \frac{\frac{1}{5}}{\frac{2}{5}\frac{1}{5}} + \frac{1}{5} log_2 \frac{\frac{1}{5}}{\frac{2}{5}\frac{2}{5}}$$
$$= \mathbf{0.571}$$

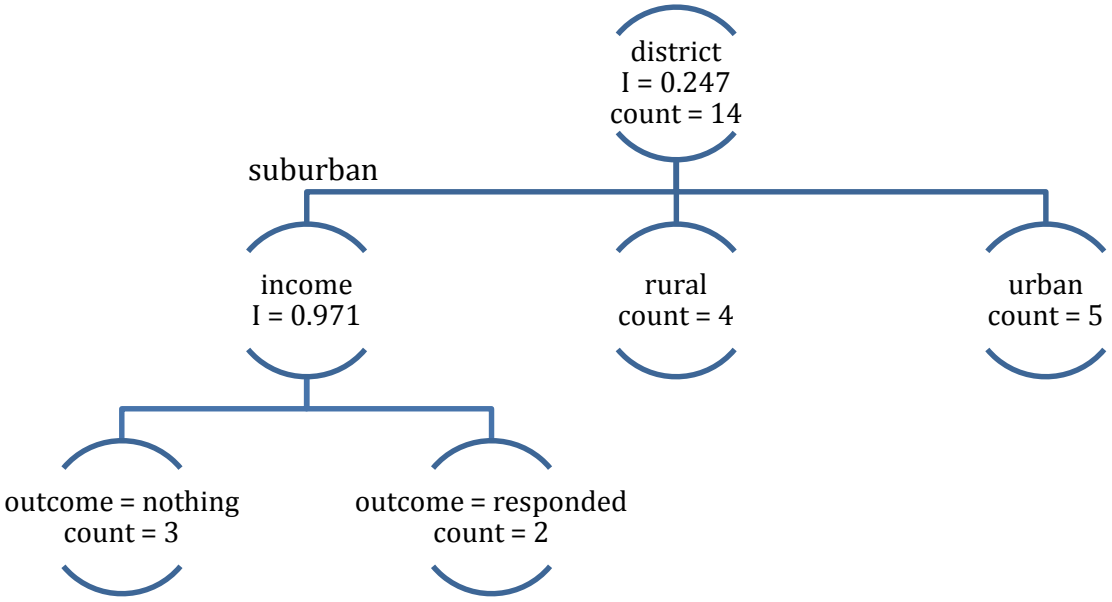
$I(outcome, income \mid district = suburban) =$

$$P(outcome = nothing, income = high) log_2 \frac{P(outcome = nothing, income = high)}{P(outcome = nothing)P(income = high)}$$
$$+P(outcome = nothing, income = low) log_2 \frac{P(outcome = nothing, income = low)}{P(outcome = nothing)P(income = low)}$$
$$+P(outcome = responded, income = high) log_2 \frac{P(outcome = responded, income = high)}{P(outcome = responded)P(income = high)}$$
$$+P(outcome = responded, income = low) log_2 \frac{P(outcome = responded, income = low)}{P(outcome = responded)P(income = low)}$$
$$= \frac{3}{5} log_2 \frac{\frac{3}{5}}{\frac{3}{5}\frac{3}{5}} + \frac{0}{5} log_2 \frac{\frac{0}{5}}{\frac{3}{5}\frac{2}{5}} + \frac{0}{5} log_2 \frac{\frac{0}{5}}{\frac{2}{5}\frac{3}{5}} + \frac{2}{5} log_2 \frac{\frac{2}{5}}{\frac{2}{5}\frac{2}{5}}$$
$$= \mathbf{0.971}$$

$I(outcome, previouscustomer \mid district = suburban) =$

$$P(outcome = nothing, previouscustomer = no) log_2 \frac{P(outcome = nothing, previouscustomer = no)}{P(outcome = nothing)P(previouscustomer = no)}$$
$$+P(outcome = nothing, previouscustomer = yes) log_2 \frac{P(outcome = nothing, previouscustomer = yes)}{P(outcome = nothing)P(previouscustomer = yes)}$$
$$+P(outcome = responded, previouscustomer = no) log_2 \frac{P(outcome = responded, previouscustomer = no)}{P(outcome = responded)P(previouscustomer = no)}$$
$$+P(outcome = responded, previouscustomer = yes) log_2 \frac{P(outcome = responded, previouscustomer = yes)}{P(outcome = responded)P(previouscustomer = yes)}$$
$$= \frac{2}{5} log_2 \frac{\frac{2}{5}}{\frac{3}{5}\frac{3}{5}} + \frac{1}{5} log_2 \frac{\frac{1}{5}}{\frac{3}{5}\frac{2}{5}} + \frac{1}{5} log_2 \frac{\frac{1}{5}}{\frac{2}{5}\frac{3}{5}} + \frac{1}{5} log_2 \frac{\frac{1}{5}}{\frac{2}{5}\frac{2}{5}}$$
$$= \mathbf{0.020}$$

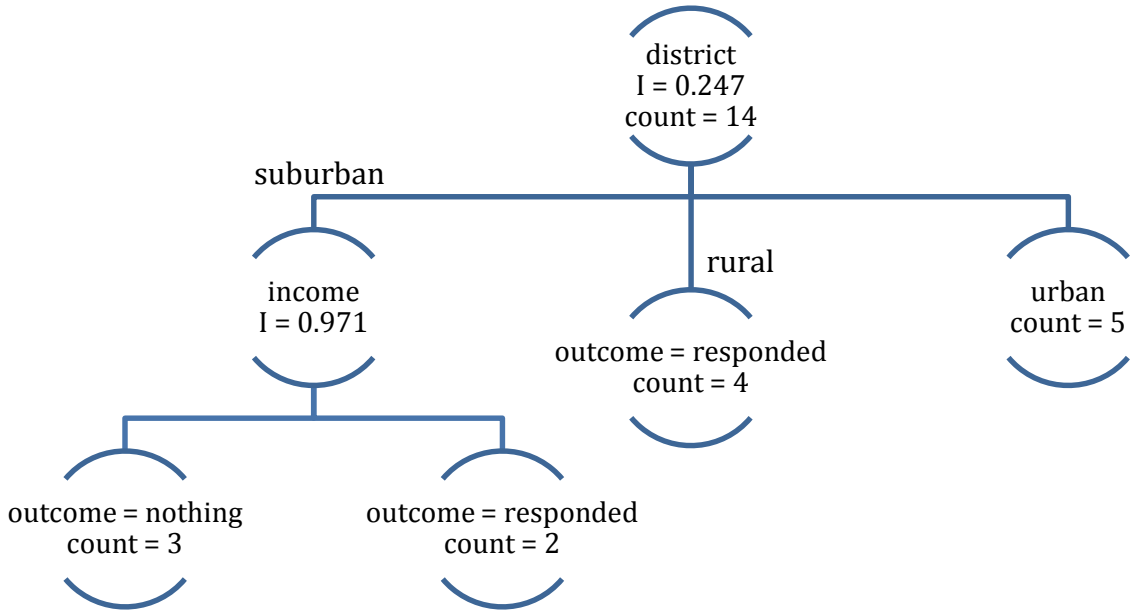
The greatest information content is $I(outcome, income \mid district = suburban) = 0.971$, so the suburban internal node will split `outcome` on `income`. In fact, this split gives perfect separation, so this split will result in two leaf nodes.



- `district == rural`

```
##   district    house_type income previous_customer   outcome
## 1    rural      detached   high                no responded
## 2    rural semi-detached    low               yes responded
## 3    rural       terrace   high               yes responded
## 4    rural      detached    low                no responded
```

Since $P(outcome = responded \mid district = rural) = 1$, no information can be gained by splitting this node further, so this path will terminate in a leaf node.



- district == urban

```
##   district    house_type income previous_customer   outcome
## 1    urban semi-detached   high                no responded
## 2    urban semi-detached    low                no responded
## 3    urban semi-detached    low               yes   nothing
## 4    urban       terrace    low                no responded
## 5    urban       terrace   high               yes   nothing
```

$I(outcome, housetype \mid district = urban) =$

$$P(outcome = nothing, housetype = semidetached) log_2 \frac{P(outcome = nothing, housetype = semidetached)}{P(outcome = nothing)P(housetype = semidetached)}$$
$$+P(outcome = nothing, housetype = terrace) log_2 \frac{P(outcome = nothing, housetype = terrace)}{P(outcome = nothing)P(housetype = terrace)}$$
$$+P(outcome = responded, housetype = semidetached) log_2 \frac{P(outcome = responded, housetype = semidetached)}{P(outcome = responded)P(housetype = semidetached)}$$
$$+P(outcome = responded, housetype = terrace) log_2 \frac{P(outcome = responded, housetype = terrace)}{P(outcome = responded)P(housetype = terrace)}$$
$$= \frac{1}{5} log_2 \frac{\frac{1}{5}}{\frac{2}{5}\frac{3}{5}} + \frac{1}{5} log_2 \frac{\frac{1}{5}}{\frac{2}{5}\frac{2}{5}} + \frac{2}{5} log_2 \frac{\frac{2}{5}}{\frac{3}{5}\frac{3}{5}} + \frac{1}{5} log_2 \frac{\frac{1}{5}}{\frac{3}{5}\frac{2}{5}}$$
$$= \mathbf{0.020}$$

$I(outcome, income \mid district = urban) =$

$$P(outcome = nothing, income = high) log_2 \frac{P(outcome = nothing, income = high)}{P(outcome = nothing)P(income = high)}$$
$$+P(outcome = nothing, income = low) log_2 \frac{P(outcome = nothing, income = low)}{P(outcome = nothing)P(income = low)}$$
$$+P(outcome = responded, income = high) log_2 \frac{P(outcome = responded, income = high)}{P(outcome = responded)P(income = high)}$$
$$+P(outcome = responded, income = low) log_2 \frac{P(outcome = responded, income = low)}{P(outcome = responded)P(income = low)}$$
$$= \frac{1}{5} log_2 \frac{\frac{1}{5}}{\frac{2}{5}\frac{2}{5}} + \frac{1}{5} log_2 \frac{\frac{1}{5}}{\frac{2}{5}\frac{3}{5}} + \frac{1}{5} log_2 \frac{\frac{1}{5}}{\frac{3}{5}\frac{2}{5}} + \frac{2}{5} log_2 \frac{\frac{2}{5}}{\frac{3}{5}\frac{3}{5}}$$
$$= \mathbf{0.020}$$

$I(outcome, previouscustomer \mid district = urban) =$

$$P(outcome = nothing, previouscustomer = no) log_2 \frac{P(outcome = nothing, previouscustomer = no)}{P(outcome = nothing)P(previouscustomer = no)}$$
$$+P(outcome = nothing, previouscustomer = yes) log_2 \frac{P(outcome = nothing, previouscustomer = yes)}{P(outcome = nothing)P(previouscustomer = yes)}$$
$$+P(outcome = responded, previouscustomer = no) log_2 \frac{P(outcome = responded, previouscustomer = no)}{P(outcome = responded)P(previouscustomer = no)}$$
$$+P(outcome = responded, previouscustomer = yes) log_2 \frac{P(outcome = responded, previouscustomer = yes)}{P(outcome = responded)P(previouscustomer = yes)}$$
$$= \frac{0}{5} log_2 \frac{\frac{0}{5}}{\frac{2}{5}\frac{3}{5}} + \frac{2}{5} log_2 \frac{\frac{2}{5}}{\frac{2}{5}\frac{2}{5}} + \frac{3}{5} log_2 \frac{\frac{3}{5}}{\frac{3}{5}\frac{3}{5}} + \frac{0}{5} log_2 \frac{\frac{0}{5}}{\frac{3}{5}\frac{2}{5}}$$
$$= \mathbf{0.971}$$

The greatest information content is $I(outcome, previous \mid district = urban) = 0.971$, so the urban internal node will split `outcome` on `previous customer`. In fact, this split gives perfect separation, so this split will result in the final two leaf nodes.

Final Decision Tree