Decision Trees: a Gentle Introduction

Richard D. Hector, Ph.D., M.P.H., M.A., Banner MD Anderson Cancer Center, Gilbert, Arizona

Abstract

Every car is a vehicle, but not every vehicle is a car. Similarly, classification and regression trees (CART) and decision trees look similar. Both begin with a single node followed by an increasing number of branches. However, they serve different purposes. The purpose of a family automobile is different from that of giant mining truck. This paper is a gentle introduction to decision trees using PROC DTREE. When you need to explore the relationship to factors and an outcome, CART is a useful non-parametric tool. The branching algorithm facilitates moving through the variables in the data to determine their effect on the outcome. In contrast, in decision trees the variables are chosen because the decision maker knows or can infer their effects on the outcome. Also, the decision maker knows or can estimate their distribution among relevant groups. Finally, the result (reward or pay-off) of each pathway is known or can be estimated. The goal of a decision tree is to ascertain the most desirable outcome given the combination of variables and costs (in other words, the best pathway). In addition, the amount of risk the decision maker is willing to accept can be incorporated in a decision tree analysis. This paper focuses on an example from medical care. An intermediate level of familiarity with SAS® is sufficient for this paper.

Introduction

Most situations facing individuals, organizations, communities or populations affected by disease or health condition can lead to several different outcomes or states. The outcomes have advantages or disadvantages with associated costs (or effort). Specifically, to claim an outcome is an advantage or disadvantage depends on how useful/useless it is. Greater preference is given to the state that is most useful and least costly. However, difficulty arises because actions have different chances of getting to that state. The challenges increase dramatically when intermediate states precede the most useful state. To address the challenges, one can systematically consider the chances of arriving at each intermediate state leading to the most useful state, i.e., decision analysis.

Decision analysis requires a considerable amount of time and effort. This investment is worthwhile when the most preferred state is likely to benefit individuals, organizations, communities or populations affected by disease or health condition. In contrast, classification and regression trees (CART) is a method that explores the effect of variables on the outcome. SAS® has implemented CART with both Enterprise Miner™ and Visual Analytics™. The interested reader should consult the documentation for these applications.

The approach to constructing decision trees is analogous to car manufacturers. Some automobiles (imagine a Corvette) cover the motor with fenders and the hood. Others leave the motor in full view (imagine a Hotrod). Similarly, some decision tree applications do not readily show the inner components of the software. PROC DTREE allows the user to build the decision by working directly with the inner components. The inner components are three data sets as well as the plotting procedure, PROC DTREE. The first data set is the states and successive sub-states in the domain for the decision. The second data set holds the probability of events to attain successive sub-states. The third data set is the payoffs for each successive sub-state and

the terminal state following the paths. Finally, PROC DTREE illustrates (and evaluates) the pathways.

To illustrate with a relatable example, imagine solving the problem of being hungry (perhaps for morning breakfast). The decision is what to have for breakfast (e.g., cereal with liquid, omelet, ham and eggs or bagel, lox and schmear).

The decision can lead to results that could look something like the – not exhaustive – list of breakfast possibilities:

- A bowl of cereal with a liquid (cow's milk, soy, rice, or almond "milk")
- An omelet (with cheese, vegetables, meats)
- Bacon and eggs (scrambled, over-easy/hard, poached)
- Bagel, lox and schmear

Let's pretend the person hungry for breakfast is not the person who stocked the cupboard or fridge. Each choice depends on the chances of obtaining the ingredients to have a successful result.

With a cereal, there's a chance of finding no cereal or some cereal. Only if cereal is found then does the chance of any of the liquids become relevant. If not, the desired liquid is not explored further.

If there are no eggs, any options with eggs will not be explored. However, on finding eggs, we explore the chances of finding vegetables, cheese or breakfast meats.

For this section, if bagels are not at hand, the chances of finding lox and schmear will not be explored.

In exploring the payoffs, the efforts (a.k.a costs) to obtain the ingredients and the preparation effort (a.k.a costs) for each of the options are specified. At this juncture, PROC DTREE can calculate and plot the decisions, chances and payoffs. PROC DTREE can be set to the desired outcome (e.g., minimum effort – cereal versus maximum calories – stuffed omelet).

Another relatable example, if less appealing than a delicious breakfast, is the annual threat of influenza for a community. Imagine that the department of health and human Services (DHHS) serving that community might need to consider how aggressively to promote a vaccination campaign. Although, many people make the decision to get vaccinated, a substantial number forgo vaccination for various reasons. DHHS stakeholders ask for a decision analysis to aid in the decision.

In summary, the breakfast example involves only the hungry diner and the person stocking the cupboard or fridge. In contrast, the threat of influenza involves the individuals and the community (family members, coworkers, teammates, fellow church/temple members, audiences and so on).

Addressing an Influenza Threat

The chances of contracting the flu with and without being vaccinated is obtained from government sources such as the state or federal department of health services or reports in

scientific journals. In addition, the chances of receiving helpful medication or adequate non-medication supportive care is also obtained from sources mentioned. The decision analysis will use an example from the Cost-Effectiveness Analysis of influenza text by Peter Muennig. All costs and probabilities in this paper come from this text.

The state dataset begins with the decision at the first observation and followed by the chance observations. In decision analysis terminology, these are called nodes, namely, decision, chance and terminal nodes.

Table 1 shows the decision node with the associated outcomes and ensuing events. The person can decide to obtain vaccination against the flu, seek treatment from a medical provider (with or without obtaining vaccination) or seek supportive care without consulting a medical provider. The letter 'D' indicates a decision node and 'C' indicates a chance node.

Table 1
Decision Node for Decision Analysis

State	Node Type	Outcome	Successor Sub state
Influenza threat	D	Support	Support result
		Treat flu	Treatment result
		Vaccinate	Vaccinate result

The successor sub states each has chance nodes. Table 2 shows sub states for the result of the decision not to get the vaccination. The result is the chance outcome of falling ill or remaining well. If the person remains well, we have no further considerations (similar to not having cereal or eggs). In contrast, we find two outcomes if the person fall ill, namely make a physician visit or not (presumably some folks use over-the-counter flu remedies without a medical visit). Similar methods are used for the treatment and vaccinate successor sub states.

Table 2
Chance Nodes for Decision Analysis

State	Node Type	Outcome	Successor Sub state
Support result	C	Ill	Support if ill
		Well	
Support if ill	C	Physician visit	
		No physician visit	

The chance outcomes have just two options. This simplifies the analysis and keeps chances from summing to more than 1.0. The chance nodes fit into probabilities and probabilities cannot sum to more than 1.0. So when the chance of an event has a probability (e.g., 0.6) the chance of the event not occurring will be 1 minus the probability of the event (1.0 - 0.6 = 0.4). For the relevant aspects of flu threats, the probabilities can be obtained from government sources and scientific reports. Table 3 shows the setup of the events and probability dataset.

Table 3
Events and Probabilities of the Decision

Event 1	Event 2	Probability 1	Probability 2
Support ill	Support well	Probability of falling ill	Probability of staying well

After specifying the events and their probabilities, it is time to specify the payoffs dataset. Again, the costs can be gleaned from government sources and scientific reports. The payoffs follow

directly from the first dataset. The sub states can be enumerated as they flow from the decision and chance outcomes. For example, after the decision not to obtain a flu vaccination, the first states are the outcomes (falling ill or staying well). Each state has an associated value. The person does not have any costs related to the flu if he or she remains well. In the case that the person falls ill and makes an ambulatory visit, that visit is the second state. The person would incur cost of the ambulatory visit. If the ill person did not make an ambulatory medical visit, we can consider the cost of over-the-counter drugs. In our example, obtaining treatment after falling ill was a first state. The treatment could have secondary complications. So visiting the doctor would be the second state and the secondary complications would be the third state.

In addition, the flu can lead to hospitalization and the treatment drug can have side-effects. These can be put in the terminal node as calculations or as separate nodes. The focus of the decision analysis will dictate if these be put in terminal node calculations or separate arms.

Specifying PROC DTREE

SAS procedures have the keyword 'data' followed by the equal sign after the name of the procedure (data =). This specifies the input dataset. In contrast, PROC DTREE has requirements for the three input datasets, data, probability and payoffs.

Figure 1 shows the decision tree diagram produced by PROC DTREE. The goal is to find the minimum cost approach to the threat of influenza from the point of view of a DHHS.

The path begins with the decision node which is the green rectangle on the left and ends with the terminal nodes on the right which are red triangles. Between the beginning and end are chance nodes which are blue ovals. The blue lines connect the nodes. In this tree, the recommended path has with a thick red line.

The *EV* shows, below the branch, the event value while the probability of the event is shown by *p*. We used variable names (above the branch) for the event. Using variable names has an important advantage that we discuss later in this paper.

The tree is evaluated from right to left. The event value is multiplied by the probability at each of the two end branches. The sum of the product becomes the value of the event for the branch they that give rise to them (the branch on the left). The evaluation iterates product and sum procedure until the beginning decision branches are calculated.

While this tree has a single decision note, it is not unusual to have more than one decision node in a tree. At the decision node branches, getting vaccinated was the lowest ($38,300). Persons obtaining medical care for the flu had the highest cost ($67,600) with support in between ($42,800). Therefore, the stakeholders can justify aggressively encouraging vaccination.

The code for this tree is in the Appendix. The analysis also considered the other relevant costs such as hospitalization, side effects from treatment and treatment complications. We did not include measures of health-related quality of life used to calculate quality-adjusted life years. As this is a gentle introduction, only costs are considered.

Figure 1 Decision Tree to Find MINIMUM Cost (X 1000) for Influenza Vaccination (Comparison of Vaccination, Medical Treatment and Non-medical Support)

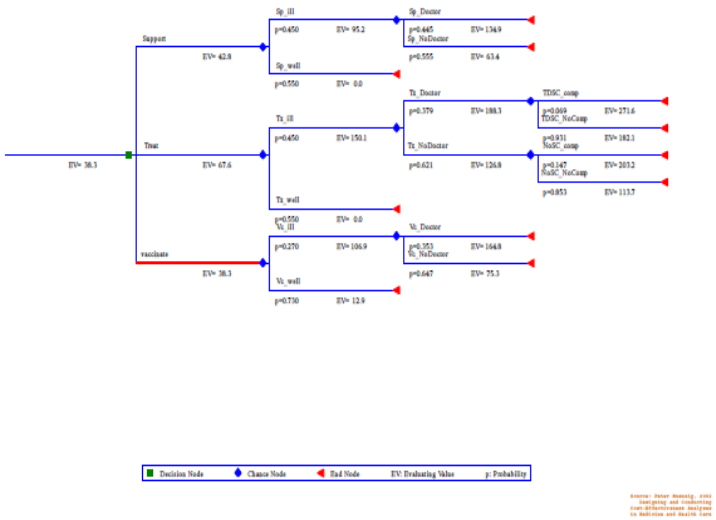

Figure 2 shows the code for decision tree. The nodes are specified with symbol1, symbol2 and symbol3 for the chance, decision and end nodes (1, 2, 3), respectively. The shape of the node marker is specified with value=. The shapes were specified with W, U and A for oval, rectangle and triangle, respectively. The title and footnotes are the usual statements used in many SAS tasks.

The decision procedure is called with PROC DTREE (4). The input datasets are stagein=, probin= and payoffs= (5). Each dataset has a convenient suffix '_ili' for 'influenza like illness'. The graphics creates plots and target= indicates the decision node for the optimal decision policy table. Warnings are suppressed with the 'nowarning' statement (6). The EVALUATE statement and its option CRITERION rolls up the decision tree to the optimal choice (minimum event cost). The optional SUMMARY statement can be used to put the results into a table (7).

PROC DTREE permits a line size for the output as well as the log (8). The TREEPLOT statement plots the tree where the scenarios are diagrammed (9). The options following the slash. The linka= and linkb= carries out the attributes of the line specified in the symbol statement. The NORC suppresses rounded corners on the plot of the tree. The decision trees can spread across many pages. So, we the COMPRESS option to keep the decision tree on a single page. Like all SAS procedures the options allows easy customization of the analysis and output.

As mentioned earlier, the use of variables is recommended instead of hardcoding. Art Carpenter provided many general methods to avoid performing hardcoding. For events, probabilities and payoffs. Using methods other than hardcoding makes the program more flexible. This is more so for decision trees.

The variables allow the decision makers to explore 'What Ifs'. The technical term is sensitivity analysis. The variables are located on the branches. To change the values on the branches, changing the variables avoids the excess work and risks of changing the value on each branch separately.

Figure 2 Decision Tree Explanation

```
/* define symbol characteristics for chance nodes and */
 /* links except those that represent optimal decisions */

symbol1 f=marker height=1.2 value=W color=blue weight=1 line=1; (1)

 /* define symbol characteristics for decision nodes */
 /* and links that represent optimal decisions */

symbol2 f=marker height=1.2 value=U cv=green ci=red weight=3 line=1; (2)

 /* define symbol characteristics for end nodes */

symbol3 f=marker height=1.2 value=A cv=red; (3)

 /* -- define footnotes for decision tree  -- */
footnote1 justify= right font='cumberland AMT' color= bio height= 0.8 'Source: Peter Muennig, 2002';
footnote2 justify= right font='cumberland AMT' color= bio height= 0.8 'Designing and Conducting';
footnote3 justify= right font='cumberland AMT' color= bio height= 0.8 'Cost-Effectiveness Analyses';
footnote4 justify= right font='cumberland AMT' color= bio height= 0.8 'in Medicine and Health Care';

proc dtree (4)

stagein = D_ili
probin   = Prob_ILI
payoffs = Payoff_ILI (5)

graphics nowarning target= flu_threat; (6)


evaluate / criterion= minev summary; (7)

/* plot decision tree diagram in graphics mode*/

options linesize=100; (8)

      Treeplot  /  linka=1 linkb=2
                   symbold=2
                   symbolc=1
                   symbole=3
                   norc  compress; (9)
run; quit;
```

For example, oseltamivir is used to treat the flu. It reduces the severity and duration of the flu. In the example that was published in 2002, the price of oseltamivir (e.g., Tamiflu) was approximately $50. At this price, treatment decision cost approximately $46,200. With 2017 prices approximately $100, the treatment decision cost $61,600. If the individual caught the flu and used OTC drugs for relief (e.g., Theraflu or Nyquil), the price could range from approximately $10 to $15 and cost from $37,500 to $39.500 (remember it was $36,900 at $8.48). [**https://www.goodrx.com/tamiflu – suggested retail price: $149.09 – accessed 7/28/2017**]

The DHHS stakeholders could explore different scenarios and costs. An important feature of decision analysis requires the stakeholders to consider the evidence and ramifications each decision. The research that go into the different aspects of the decision lead to a more informed choices.

Conclusion

In the SAS products, two distinct types of dendrograms are called decision trees. This paper is a gentle introduction to decision trees where one desires making the optimal choice among several options. It provides a graphical representation of the states and events leading to the different options. This allows others to see the issues considered on arriving at the decision. Even if the recommended path is not followed, stakeholders would be aware of the risks in the not recommended paths and can preparations can be made to lessen any undesirable effects.

References

Carpenter, A. (2012). Carpenter's Guide to Innovative SAS® Techniques. Cary, NC: SAS Institute Inc.

Muennig, P., & Kamran, K. (2002). Designing and Conducting Cost-effectiveness Analyses in Medicine and Health Care. *San Francisco: Jossey-Bass*.

Appendix

Implement Decision Tree for Least Costly Path to Influenza Threat

```
%let cAbx           = 46.42;    /* cost of antibiotics                          */
%let cCare          = 53.56;    /* cost of caregiving                           */
%let cHosp          = 5728;     /* cost of hospitalization                      */
%let cMD            = 68.41;    /* cost of medical visit                        */
%let cRx            = 49.82;    /* cost of oseltamivir                          */
%let cOTC           =  8.48;    /* cost of over the counter medications         */
%let cSE            = 45.11;    /* cost of side effects                         */
%let cVaccine       = 12.47;    /* cost of vaccine                              */
%let pAbx_SC        =  0.065;   /* probability of antibiotics, support          */
%let pAbx_Rx        =  0.45;    /* probability of antibiotics, treat            */
%let pAbx_Vac       =  0.45;    /* probability of antibiotics, vaccinate        */
%let pHosp          =  0.00024; /* probability of hospitalization               */
%let Vacc_eff_ill   =  0.40;    /* vaccine efficacy influenza type illness      */
%let Vacc_eff_infl  =  0.803;   /* vaccine efficacy influenza                   */
%let Osel_eff_ill   =  0.2266;  /* efficacy of oseltamivir reducing flu severity */
%let pMD_SC         =  0.445;   /* probability of medical visit, support        */
%let pMD_Rx         =  0.379;   /* probability of medical visit, treat          */
%let pMD_Vac        =  0.353;   /* probability of medical visit, vaccinate      */
%let pSec_Rx        =  0.069;   /* probability of secondary complications, treat */
%let pSec_NoRx      =  0.147;   /* probability of secondary complications, support */
%let pSE            =  0.01;    /* probability of side effects                  */
%let IR_flu         =  0.45;
%let pSp_ill        =  &IR_flu;
%let pTx_ill        =  &IR_flu;
%let pVc_ill        =  &pSp_ill * (1 - &Vacc_eff_ill);
%let pHosp_Vac      = (&pHosp * (1 - &Vacc_eff_infl))/&pVc_ill;

/********************** equation 9.6 ********************************************/

* -- create the STAGEIN= data set shell -- -- match the 'I;
data D_ILI;
format _STNAME_ $15. _STTYPE_ $2. _OUTCOM_ $15. _SUCCES_ $15.;

input _STNAME_ $ _STTYPE_ $ _OUTCOM_ $ _SUCCES_ $ ;
/*************************************************
 * note that cVc_ill does
 *************************************************/
datalines;
flu_threat      D    Support        Sup_result
.               .    Treat          Tx_result
.               .    vaccinate      Vc_result

Sup_result      C    Sp_ill         cSp_ill
.               .    Sp_well        .
cSp_ill         C    Sp_Doctor      .
.               .    Sp_NoDoctor    .

Tx_result       C    Tx_ill         cTx_ill
.               .    Tx_well        .
cTx_ill         C    Tx_Doctor      TD_SC
.               .    Tx_NoDoctor    TD_NoSC
TD_SC           C    TDSC_comp      .
.               .    TDSC_NoComp    .
TD_NoSC         C    NoSC_comp      .
.               .    NoSC_NoComp    .

Vc_result       C    Vc_ill         cVc_ill
.               .    Vc_well        .
cVc_ill         C    Vc_Doctor      .
.               .    Vc_NoDoctor    .
;
run;

data Prob_ILI; /* probability of Influenza-Like-Illness */
```

```
        format _event1 _event2 $12. _prob1 _prob2 5.3;

        *set ieq.Prob_ILI;

        _event1 = 'Sp_ill';           _prob1 = &pSp_ill;
        _event2 = 'Sp_well';          _prob2 = 1 - _prob1;
        output;
        _event1 = 'Tx_ill';           _prob1 = &pTx_ill;
        _event2 = 'Tx_well';          _prob2 = 1 - _prob1;
        output;
        _event1 = 'Vc_ill';           _prob1 = &pVc_ill;
        _event2 = 'Vc_well';          _prob2 = 1 - _prob1;
        output;
        _event1 = 'Sp_Doctor';        _prob1 = &pMD_SC;
        _event2 = 'Sp_NoDoctor';      _prob2 = 1 - _prob1;
        output;
        _event1 = 'Tx_Doctor';        _prob1 = &pMD_Rx;
        _event2 = 'Tx_NoDoctor';      _prob2 = 1 - _prob1;
        output;
        _event1 = 'Vc_Doctor';        _prob1 = &pMD_Vac;
        _event2 = 'Vc_NoDoctor';      _prob2 = 1 - _prob1;
        output;
        _event1 = 'TDSC_comp';        _prob1 = &pSec_Rx;
        _event2 = 'TDSC_NoComp';      _prob2 = 1 - _prob1;
        output;
        _event1 = 'NoSC_comp';        _prob1 = &pSec_NoRx;
        _event2 = 'NoSC_NoComp';      _prob2 = 1 - _prob1;
        output;
run;

* -- create PAYOFFS= data set -- ;

data Payoff_ILI;

format _state1-_state3 $13. _value_ 5.2;

     _state1 = 'Sp_well'; _value_ = 0; output;
     _state1 = 'Tx_well'; _value_ = 0; output;
     _state1 = 'Vc_well'; _value_ = &HRQL_well * (&cVaccine + (&pSE * &cSE)); output;

    /* Support */
    _state1 = 'Sp_ill'; _state2 = 'Sp_Doctor';   _value_ = &HRQL_ill *  (&cMD + (&cHosp * &pHosp)
+ &cOTC + &cCare + (&pAbx_SC * (&cAbx + (&pSE * &cSE)))); output;
    _state1 = 'Sp_ill'; _state2 = 'Sp_NoDoctor'; _value_ = &HRQL_ill *  ((&cHosp * &pHosp) +
&cOTC + &cCare); output;

   /* Treatment */
    _state1 = 'Tx_ill'; _state2 = 'Tx_Doctor'; _state3 = 'TDSC_comp';   _value_ = (&cMD * 2) +
&cRx + (&pSE * &cSE) + &cOTC + &cCare + (&pAbx_Rx * (&cAbx + (&pSE * &cSE))) + (&pHosp * &cHosp);
output;
    _state1 = 'Tx_ill'; _state2 = 'Tx_Doctor'; _state3 = 'TDSC_NoComp'; _value_ = &cMD + &cRx +
(&pSE * &cSE) + &cOTC + &cCare + (&pHosp * &cHosp); output;

    _state1 = 'Tx_ill'; _state2 = 'Tx_NoDoctor'; _state3 = 'NoSC_comp';   _value_ = &cMD + &cRx +
(&pSE * &cSE) + &cOTC + &cCare + (&pAbx_Rx * (&cAbx + (&pSE * &cSE))) + (&cHosp * &pHosp);
output;
    _state1 = 'Tx_ill'; _state2 = 'Tx_NoDoctor'; _state3 = 'NoSC_NoComp'; _value_ = &cRx + (&pSE
* &cSE) + &cOTC + &cCare + (&cHosp * &pHosp)); output;

   /* Vaccine */
    _state1 = 'Vc_ill'; _state2 = 'Vc_Doctor';   _state3 = ''; _value_ = &cVaccine + (&pSE *
&cSE) + &cMD + (&cHosp * &pHosp_Vac) + &cOTC + &cCare + (&pAbx_Vac * (&cAbx + (&pSE * &cSE)));
output;
    _state1 = 'Vc_ill'; _state2 = 'Vc_NoDoctor'; _state3 = ''; _value_ = &cVaccine + (&pSE *
&cSE) + (&cHosp * &pHosp_Vac) + &cOTC + &cCare; output;
run;

/* define symbol characteristics for chance nodes and */
/* links except those that represent optimal decisions */
symbol1 f=marker h=1.2 v=W c=blue w=1 l=1;
```

```
/* define symbol characteristics for decision nodes */
/* and links that represent optimal decisions */
symbol2 f=marker h=1.2 v=U cv=green ci=red w=3 l=1;

/* define symbol characteristics for end nodes */
symbol3 f=marker h=1.2 v=A cv=red;

/* define graphics options */
goptions htext=1.2;

/* -- define title -- */
title1 font= 'cumberland AMT' 'Find MINIMUM cost (X 1000) for Flu Season';
title2 font= 'cumberland AMT' 'No shot (then SUPPORT or TREAT) versus VACCINATE';
footnote1 j= r font='cumberland AMT' c= bio h= 0.8 'Source: Peter Muennig, 2002';
footnote2 j= r font='cumberland AMT' c= bio h= 0.8 'Designing and Conducting';
footnote3 j= r font='cumberland AMT' c= bio h= 0.8 'Cost-Effectiveness Analyses';
footnote4 j= r font='cumberland AMT' c= bio h= 0.8 'in Medicine and Health Care';

ods output policy= work.policy (keep= out1 value optmark);
proc dtree
 stagein = D_ili
 probin  = Prob_ILI
 payoffs = Payoff_ILI
 graphics nowarning target= flu_threat;

 evaluate / criterion= minev;

 /* plot decision tree diagram in graphics mode*/
 OPTIONS LINESIZE=100;
 treeplot/     linka=1 linkb=2
                       symbold=2 symbolc=1 symbole=3
                       norc   compress;
run; quit;
```