

FIT3152 Data analytics. Tutorial 07:

Decision Trees

Objectives:

- Understand entropy and information gain in relation to decision tree algorithms
- Create decision trees using R
- Using decision trees for profiling
- Using decision trees for classifying unseen instances
- Evaluation of decision tree models – confusion matrices

References: Introduction to Data Mining, Tan, Steinbach and Kumar (available from Hargrave-Andrew library); An Introduction to Statistical Learning with applications in R (Springer Texts in Statistics), James, Witten, Hastie and Tibshirani, Chapter 8 (available on-line from Monash Library)

- 1 Work through the examples in the lecture slides. For the examples using R you will need to download and install the ‘tree’ package using the following code.

```
install.packages("tree")  
library(tree)
```

- 2 Calculation of entropy and information gain

Table 1 below includes data for 10 different types of aliens. The data is to be used to determine which aliens are friendly and which are not.

- a. What is the entropy of the IsFriendly class?
- b. Which decision attribute should you choose as the root of the decision tree – you can work this out without doing any calculations. Explain why you chose that attribute.
- c. What is the information gain of the attribute you chose in b.?
- d. Using the attribute you chose in b. as the root node, and using examples 1 to 10, build a decision tree for classifying aliens as friendly or not.
- e. Using your decision tree, what are the predictions for aliens 11, 12, 13 in table 2?

Table 1: Training Set

ID	colour	size	teeth	IsFriendly
1	red	medium	yes	no
2	blue	big	no	yes
3	green	medium	no	no
4	green	small	yes	no
5	blue	big	yes	yes
6	blue	small	yes	yes
7	red	small	no	yes
8	red	medium	no	yes
9	blue	medium	yes	yes
10	green	small	no	no

Table 2: Test Set

ID	colour	size	teeth	IsFriendly
11	red	big	yes	?
12	green	big	yes	?
13	blue	small	no	?

- 3 The built-in data set `mtcars` describes the fuel consumption and 10 other variables for 32 cars produced during 1973 – 1974. Fuel consumption is determined as miles per gallon (mpg). Create a decision tree to classify cars as either high consumption (greater than the median), or low consumption. To do this, follow the steps below.

- a. Convert the mpg variables in to a class using the script below and create new data set: `carsclass`.

```
data(mtcars)
attach(mtcars)
attach(mtcars)
summary(mpg)
cons = ifelse(mpg >= 19.20, "yes", "no")
carsclass = cbind(cons, mtcars)
head(carsclass)
```

- b. Partition your new data set into 70% training and 30% test.
- c. Fit the ‘tree’ model to your data. Make sure you don’t include mpg as an attribute. You may need to first create a synthetically larger training data set using resampling with replacement as was done for the playtennis example.
- d. Examine your decision tree using summary and plot functions. What are the important attributes for determining fuel economy?
- e. Using the test data set, calculate the accuracy of your model. How well does it predict fuel economy?

- 4 The Zoo data set (`zoo.data.csv`) contains data relating to seven classes of animals.

Using all the data, construct a decision tree to predict class “type” based on the other attributes. Note that you will need to specify that “type” is a factor. Which attributes are most influential in determining the class of an animal. What classes have less than 100% accuracy?

Now split your data into a training set 70% and test set 30% and refit the model. What is the overall accuracy of the model when measured on the test set?

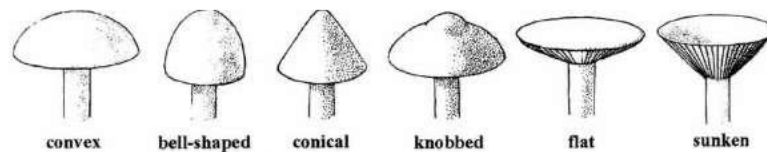
- 5 The Mushroom data set (`mushroom.data.csv`) contains 22 pieces of information about 8000+ species of mushrooms. This data set was obtained from the UCI Machine Learning Repository. Further information about the data can be obtained from: <https://archive.ics.uci.edu/ml/datasets/Mushroom>

Using the data, construct a decision tree to predict whether an unclassified mushroom is of “class” poisonous or edible based on the other attributes:

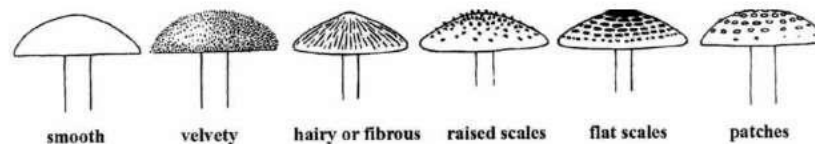
1. cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
2. cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s
3. cap-color: brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p, purple=u, red=e, white=w, yellow=y
4. bruises?: bruises=t,no=f
5. odor: almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s
- ...

You should create a training and test data set and report the accuracy of your model. Which attributes are most important in distinguishing between poisonous and edible mushrooms?

Mushroom cap shapes



Mushroom cap surfaces



- 6 Using data from the Kaggle competition: Titanic: Machine Learning from Disaster (ref <https://www.kaggle.com/c/titanic/data>) develop a decision tree model to predict whether a passenger would have or have not survived. The data has been portioned into a training and a test set: (Kaggle.Titanic.train.csv, Kaggle.Titanic.test.csv) The main details of the attributes, from the Kaggle site, are:

VARIABLE DESCRIPTIONS:

survival	Survival (0 = No; 1 = Yes)
pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
name	Name
sex	Sex
age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare
cabin	Cabin
embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

How accurate is your model? Based on your model, what are the most important predictors for survival?