Matters of Discussion

More on Classification!!!!

classification and regression trees

Logistic Regression

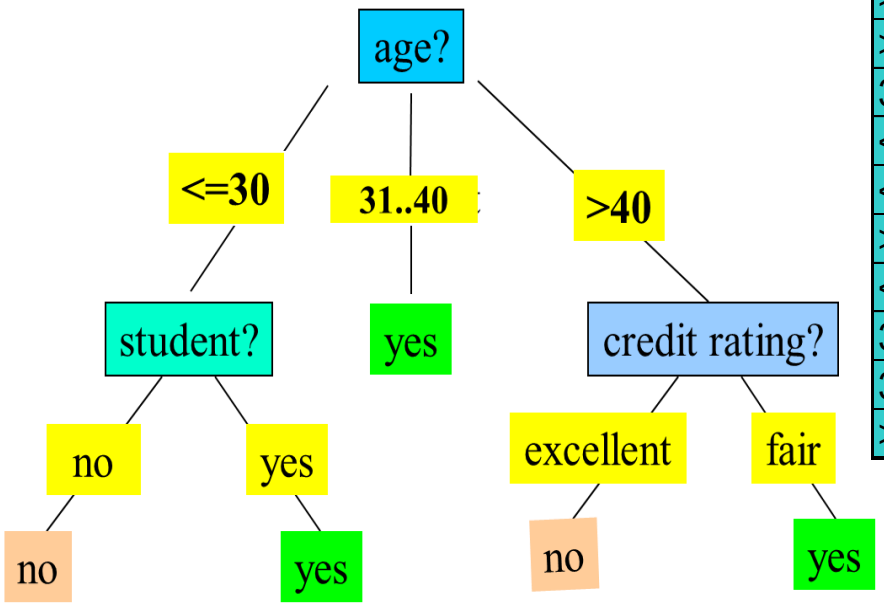
Compiled By: Dr. Nilamadhab Mishra [(PhD- CSIE) Taiwan]

Classification and Regression Tree(CART)

- CART is a term used to describe decision tree algorithms that are used for classification and regression learning tasks.
- In order to understand classification and regression trees better, decision tree plays vital role.

Decision Tree Induction: An Example

- Training data set: Buys_computer
- The data set follows an example of ID3
- Resulting tree:



age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Decision Tree Example . using ID3

- Extracting Classification Rules from the decision tree

- If **Age** (31…40) Then **Buys-Computer** (Yes)
- If **Age** (<=30) And **Student** (No) Then **Buys-Computer** (No)
- If **Age** (<=30) And **Student** (Yes) Then **Buys-Computer** (Yes)
- If **Age** (>40) And **Cr-Rating** (Excellent) Then **Buys-Computer** (No)
- If **Age** (>40) And **Cr-Rating** (Fair) Then **Buys-Computer** (Yes)

REVIEW

- Machine learning algorithms can be classified into two types- supervised and unsupervised.
- A decision tree is a supervised machine learning algorithm.
- It has a tree-like structure with its root node at the top.

The CART or Classification & Regression Trees methodology refers to these two types of decision trees.

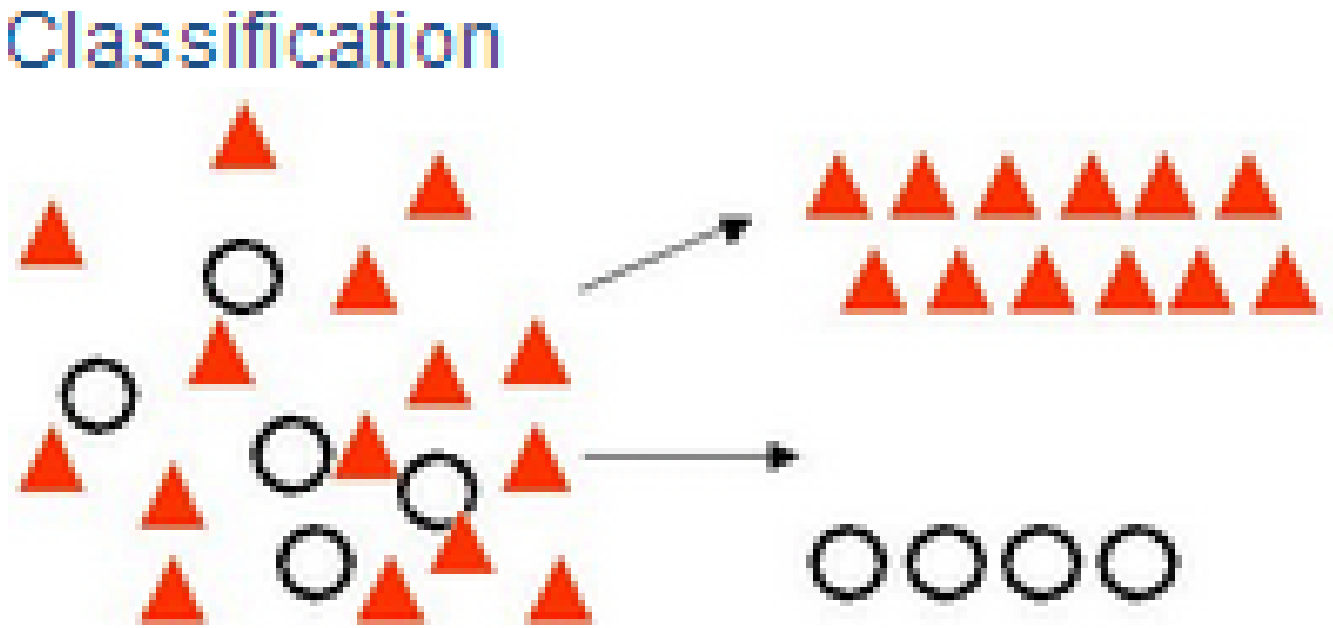
1. **Classification Trees**
2. **Regression Trees**

1. Classification Trees

- A classification tree is an algorithm where the target variable is fixed or categorical.
- The algorithm is then used to identify the "class".
- **binary classifications**: classification-type problem would be determining
  - who will or will not subscribe to a digital platform;
  - who will or will not graduate from high school.

Cont..

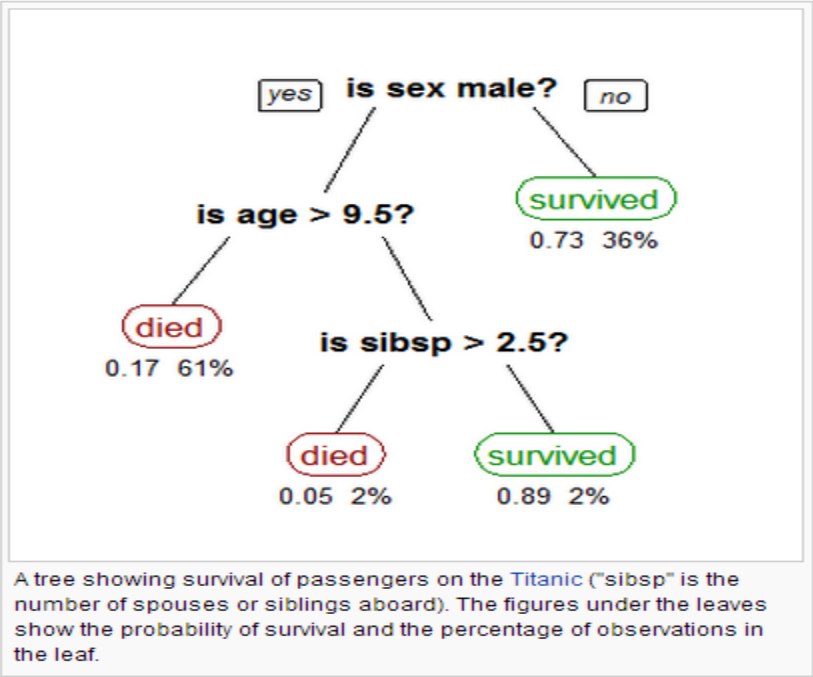
- Classification Trees: where the target variable is categorical and the tree is used to identify the "class" within which a target variable would likely fall into.



2. Regression Trees

- refers to an algorithm where for the target variable is Y and the algorithm is used to predict it's value based on the input parameter X. [Y = mX + epsilon[error]].
- As an example of a regression type problem, you may want to predict the selling prices of a residential house [Y],
- In dependent variables [**X**] like square footage as well as categorical factors like the style of home, area in which the property is located and so on.

- **Regression Trees**: where the target variable [Y] is continuous and tree is used to predict it's value.



A tree showing survival of passengers on the Titanic ("sibsp" is the number of spouses or siblings aboard). The figures under the leaves show the probability of survival and the percentage of observations in the leaf.

- The CART algorithm is structured as a sequence of questions, the answers to which determine what the next question, if any should be.
- The result of these questions is a tree like structure where the ends are terminal nodes at which point there are no more questions.

Sibsp – no of siblings or spouses

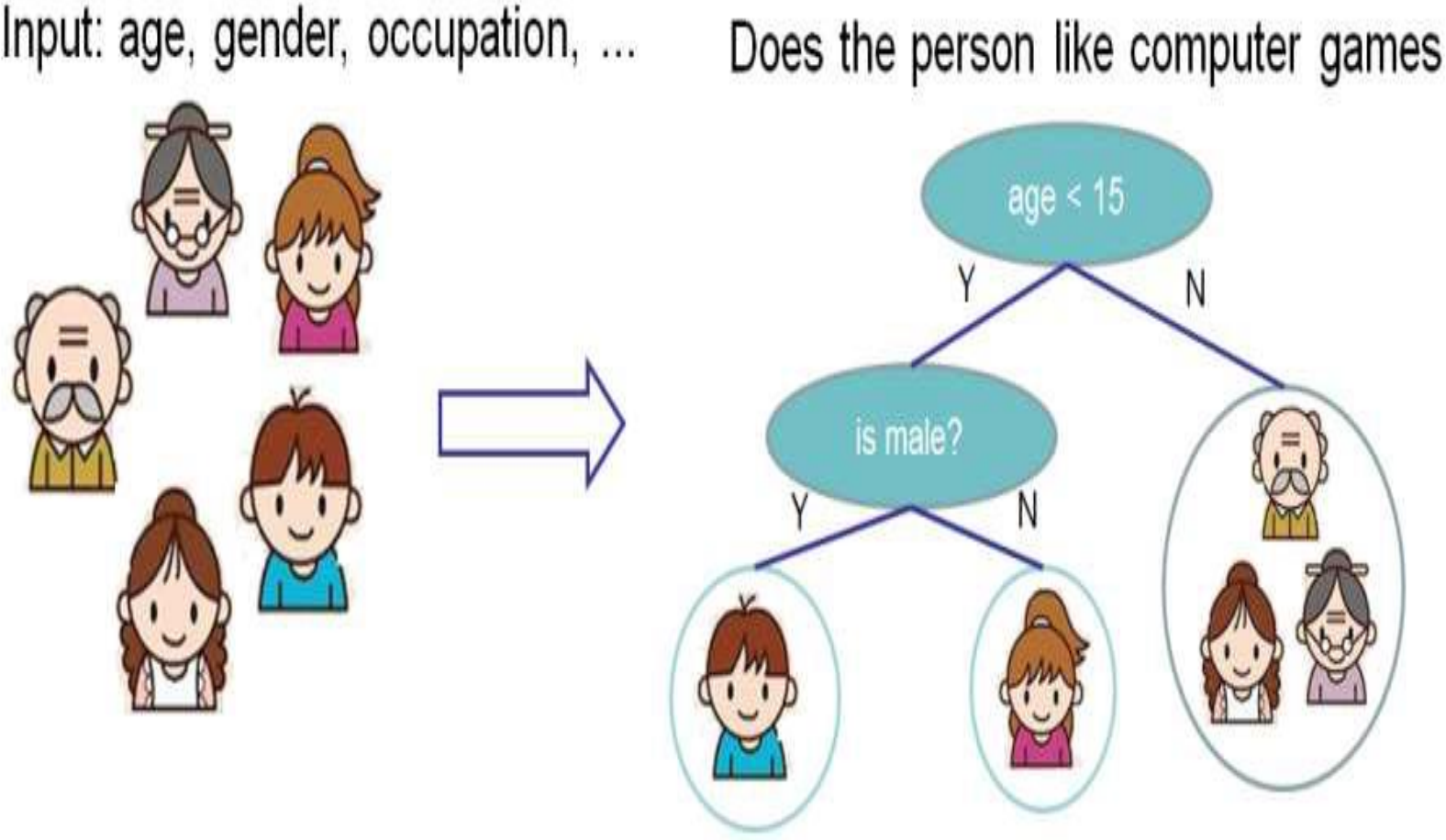
Key aim of CART and results

- create a set of if-else conditions that allow for the accurate prediction [predict the exact value] or classification of a case [class level].
- The results from classification and regression trees can be summarized in simplistic if-then conditions.

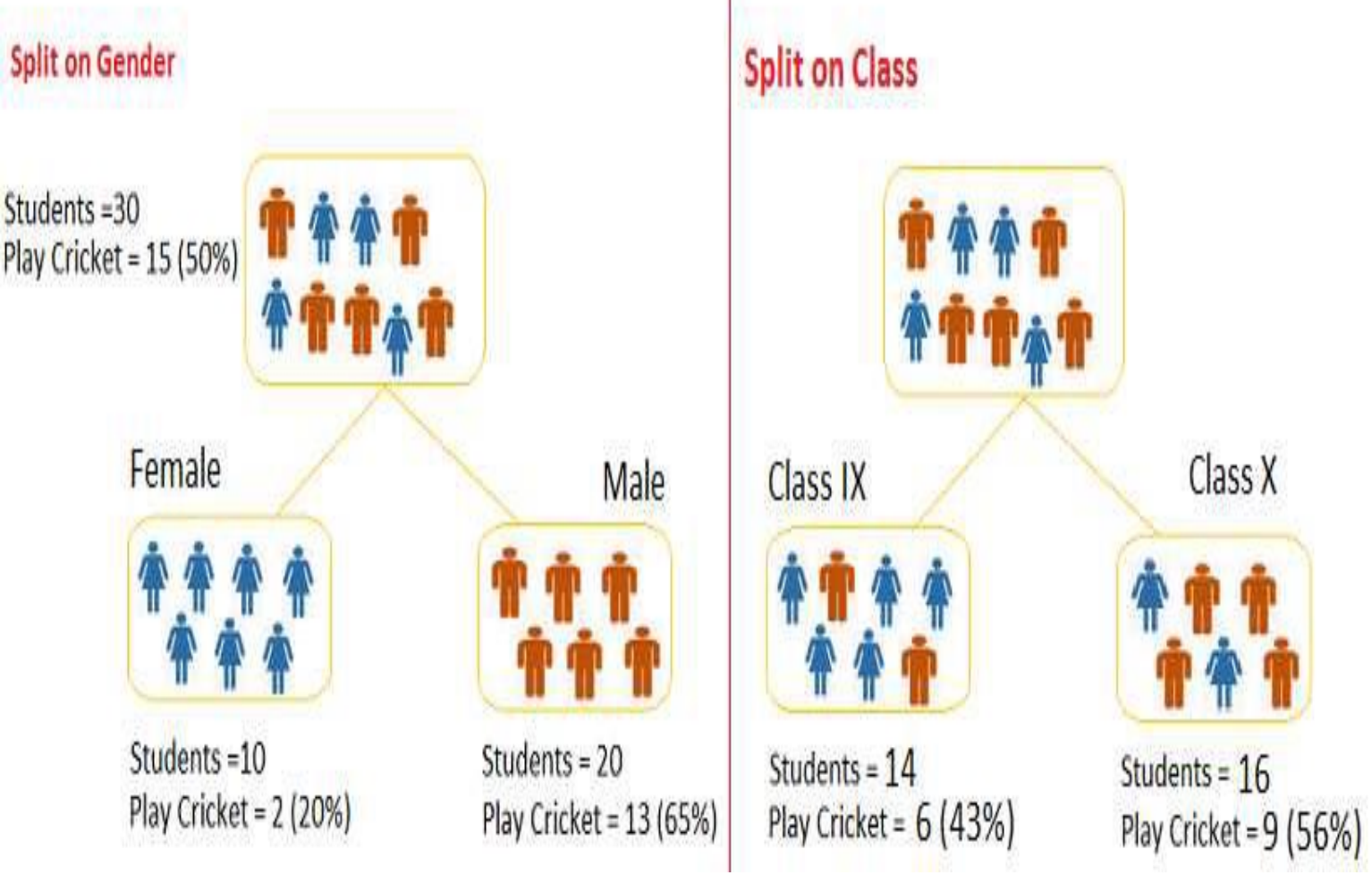
When to use Classification and Regression Trees

- Classification trees are used when the dataset needs to be split into classes which belong to the response variable. In many cases, the classes Yes or No.
- Regression trees, are used when the response variable is continuous.
  - For instance, if the response variable is something like the price of a property or the temperature of the day, a regression tree is used.

HOW CART works [Example]



If the dependent variable [Y] is categorical, CART produces a classification tree. And if the variable is continuous, it produces a regression tree.



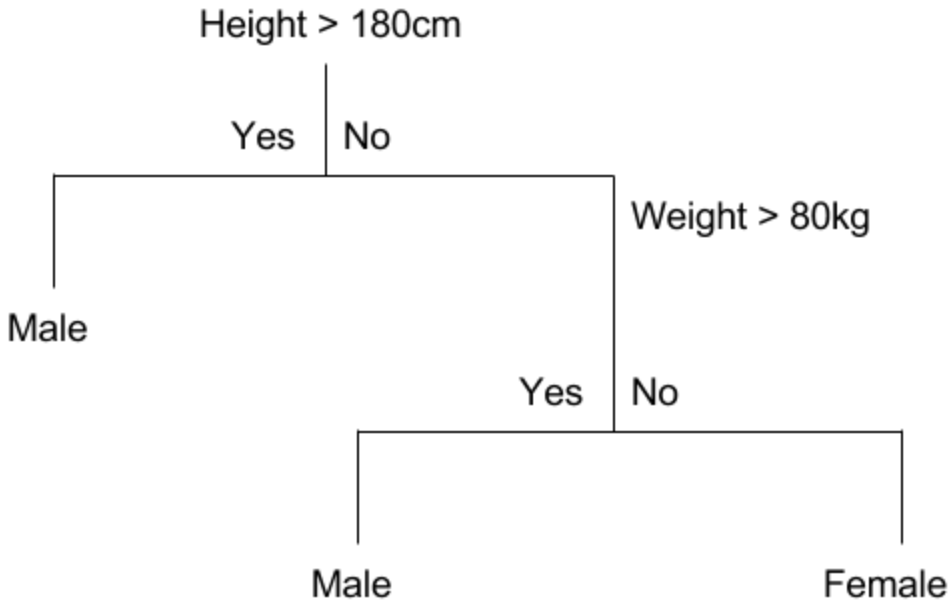
$Y = mX + C$

THE KEY IDEA

- Take all of your data.
- Consider all possible values of all variables.
- Select the variable/value (X=t1) that produces the greatest
- "separation" in the target.
- (X=t1) is called a "split".
- If (X< t1) then send the data to the "left"; otherwise, send data point to the "right".
- Now repeat same process on these two "nodes"
- You get a "tree"

Note: CART only uses binary splits.

CART model transformation example



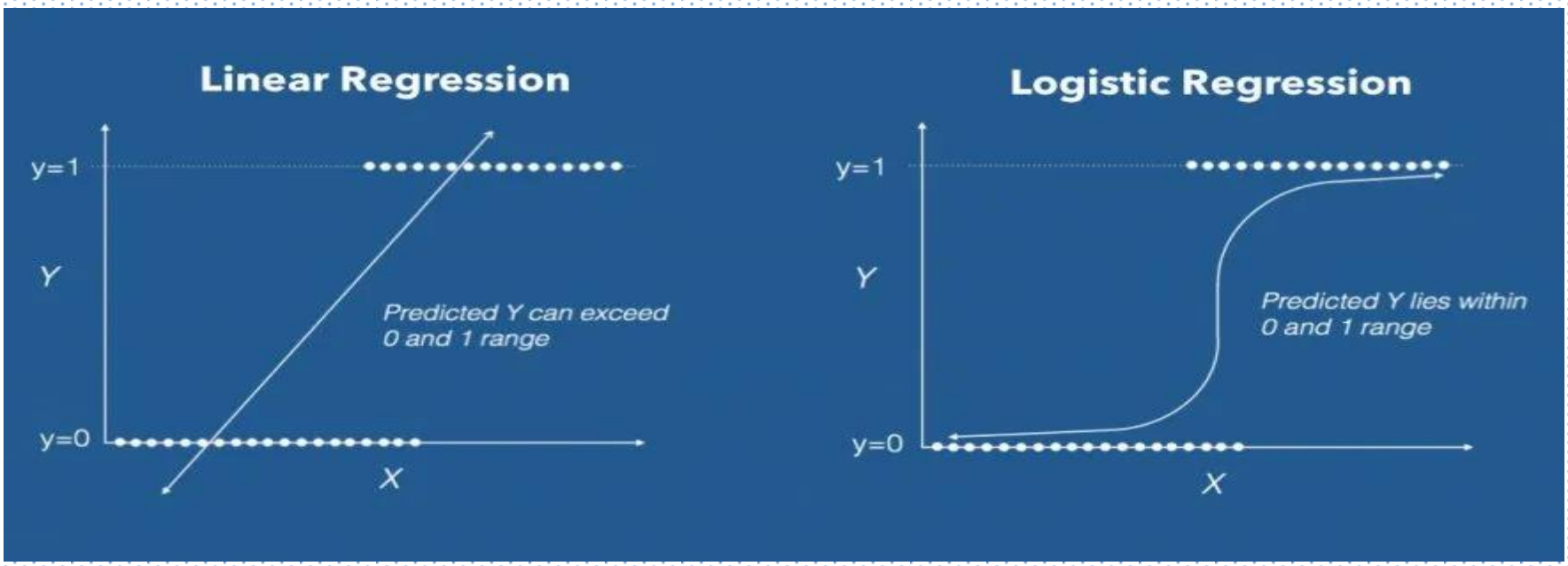
CART model

If Height > 180 cm Then Male

If Height <= 180 cm AND Weight > 80 kg Then Male

If Height <= 180 cm AND Weight <= 80 kg Then Female

Make Predictions With CART Models



When the response variable has only 2 possible values, it is desirable to have a model that predicts the value either as 0 or 1 or as a probability score that ranges between 0 and 1.

Linear regression does not have this capability. Because, If you use linear regression to model a binary response variable, the resulting model may not restrict the predicted Y values within 0 and 1.

NEXT REGRESSION TREE-CART implements in R

```
install.packages("rpart")
install.packages("rpart.plot")
install.packages("ggplot2")
library(rpart)
library(rpart.plot)
library(ggplot2)
data() # to check the availability of datasets
data(msleep)
str(msleep) # to view the structure of the dataset
df <- msleep[ , c(3,4,6,10,11)] # reduce to specific attributes
str(df) # to view the structure of new data frame
head(df) # to view the table
# sleep_total ~ brainwt, bodywt
m1 <- rpart(sleep_total~ ., data = df , method= "anova")

print(m1)

rpart.plot(m1, type=3, digits=3, fallen.leaves = TRUE)

p1 <- predict(m1, df)

print(p1)
```

REGRESSION TREE-CART implements in R

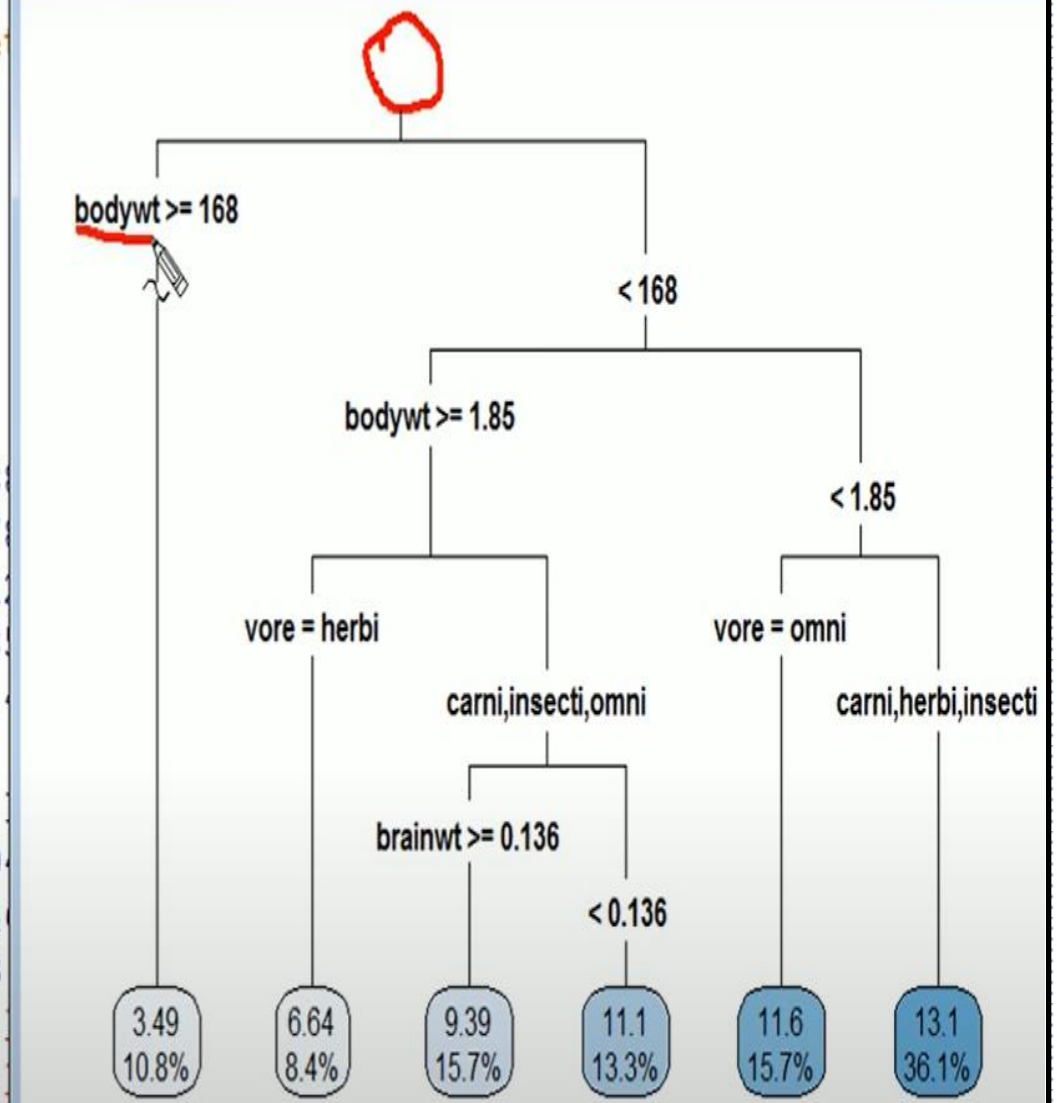
R Console

```
> m1 <- rpart(sleep_total ~ ., data=df, me
> m1
n= 83

node), split, n, deviance, yval
      * denotes terminal node

 1) root 83 1624.066000 10.433730
   2) bodywt>=167.947 9    7.868889  3.488
   3) bodywt< 167.947 74 1129.325000 11.27
     6) bodywt>=1.85 31  458.593500  9.361
      12) vore=herbi 7   88.337140  6.6428
      13) vore=carni,insecti,omni 24  303.
        26) brainwt>=0.136 13  128.669200
        27) brainwt< 0.136 11  158.307300
     7) bodywt< 1.85 43  474.662800 12.660
      14) vore=omni 13  141.370800 11.6384
      15) vore=carni,herbi,insecti 30  313
> rpart.plot(m1, type=3, digits=3, fallen.
> rpart.plot(m1, type=3, digits=3, fallen.
>
```

R Graphics: Device 2 (ACTIVE)



Logistic Regression

- Logistic regression is forcefully not a classification algorithm on its own.
- It is only a classification algorithm in combination with a decision rule that makes the predicted probabilities of the outcome.

Application analysis for Logistic Regression

- As an example, consider the task of predicting someone's gender (Male/Female) based on their Weight and Height.
- For this, we will train a machine learning model from a data set of 10,000 samples of people's weight and height.

Logistic Regression for R Implements

- The Logistic Regression is a regression model in which the response variable (dependent variable) has categorical values such as True/False or 0/1.
- It actually measures the probability of a binary response as the value of response variable based on the mathematical equation relating it with the predictor variables.

Cont.. e- scientific notation

The general mathematical equation for logistic regression is –Sigmoid Function

y = 1/(1+e^-(a+b1x1+b2x2+b3x3+...))

- ✓ y is the response variable.
- ✓ x is the predictor variable.
- ✓ a and b are the coefficients which are numeric constants.

The function used to create the regression model ---- glm() function.

Cont..

glm(formula,data,family)

- ✓ formula is the symbol presenting the relationship between the variables.
- ✓ data is the data set giving the values of these variables.
- ✓ family is R object to specify the details of the model. It's value is binomial for logistic regression.

Logistic Regression

- Logistic regression is a binary classification algorithm.
- You can implement using the glm() function by setting the family argument to "binomial".

Step 1: Build Logit Model on Training Dataset

```
logitMod <- glm(Y ~ X1 + X2, family="binomial", data = trainingData)
```

Step 2: Predict Y on Test Dataset

```
predictedY <- predict(logitMod, testData, type="response")
```

Cont..

```
# Select some columns form mtcars.
input <- mtcars[,c("am","cyl","hp","wt")]
print(head(input))
```

	am	cyl	hp	wt
Mazda RX4	1	6	110	2.620
Mazda RX4 Wag	1	6	110	2.875
Datsun 710	1	4	93	2.320
Hornet 4 Drive	0	6	110	3.215
Hornet Sportabout	0	8	175	3.440
Valiant	0	6	105	3.460

Cont..

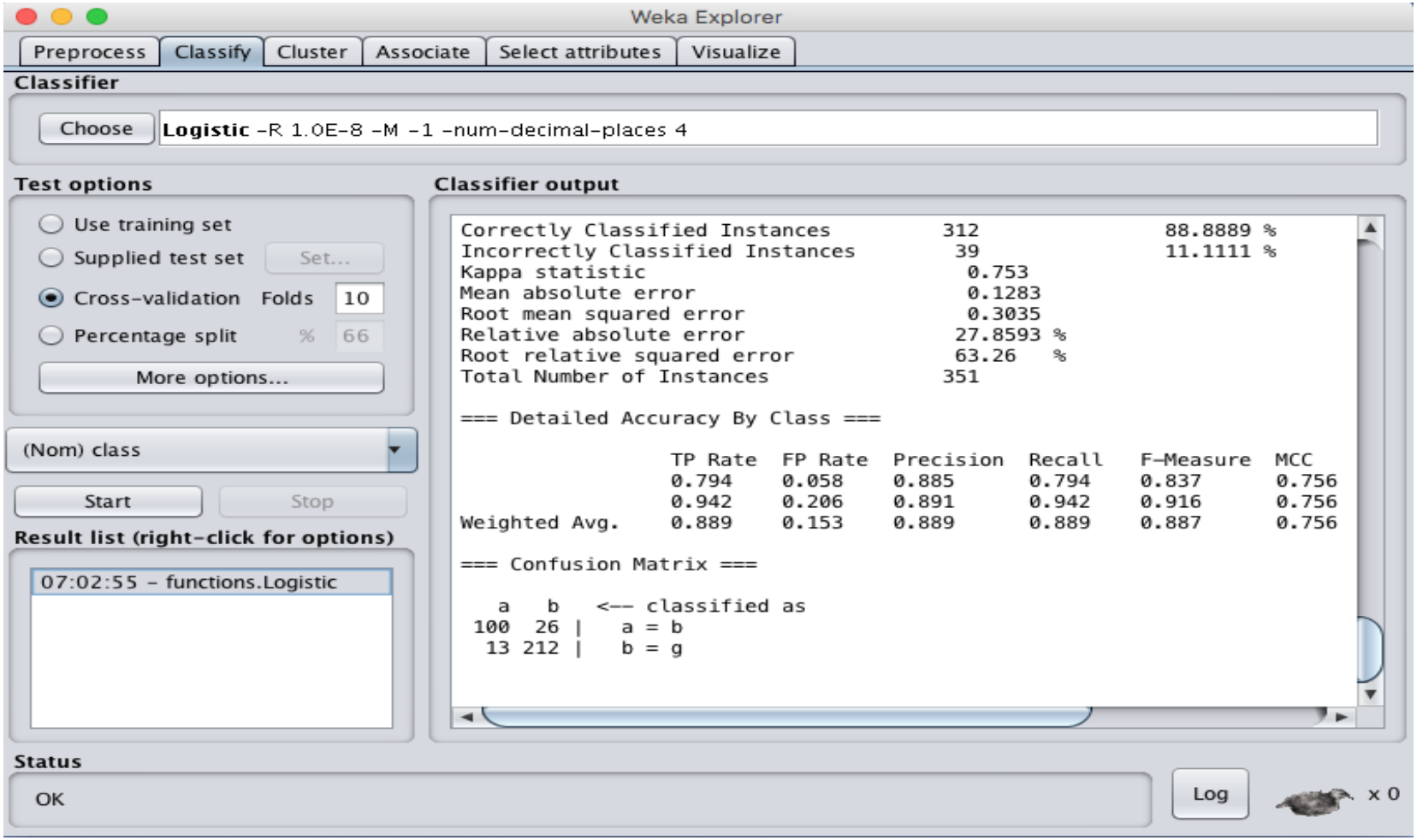
```
input <- mtcars[,c("am","cyl","hp","wt")] am.data = glm(formula = am ~ cyl + hp + wt, data = input, family = binomial) print(summary(am.data))
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 19.70288    8.11637   2.428   0.0152 *
cyl          0.48760    1.07162   0.455   0.6491
hp           0.03259    0.01886   1.728   0.0840 .
wt          -9.14947    4.15332  -2.203   0.0276 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Result analysis

- In the summary as the p-value in the last column is more than 0.05 for the variables "cyl" and "hp",
- we consider them to be insignificant in contributing to the value of the variable "am".
- Only weight (wt) impacts the "am" value in this regression model.

Logistic Regression as classification algorithm



Analysis

- This method seeks to simplify the model during training by minimizing the coefficients learned by the model.
- The ridge parameter defines how much pressure to put on the algorithm to reduce the size of the coefficients.
- You can see that with the default configuration that logistic regression achieves an accuracy of 88%.

ACTIVITY-12(LAB—06)

Stepwise investigate the implementations of logistic regression algorithm by considering any application , and analyze the results in detail.

Cheers For the Great Patience!

Query Please?

Compiled By: Dr. Nilamadhab Mishra [(PhD- CSIE) Taiwan]