Introduction to Data Science

DSA1101

Semester 1, 2018/2019
Week 6

Classification methods: Decision Trees

Classification methods: Decision Trees

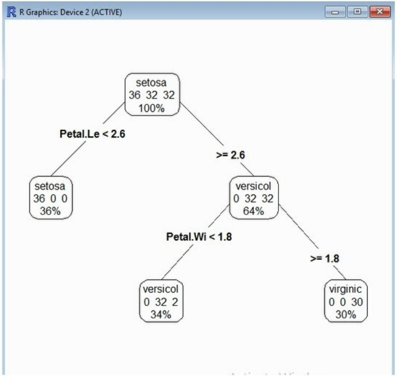
- Classification is widely used for prediction purposes.
- For example, by building a classifier on the transcripts of United States Congressional floor debates, it can be determined whether the speeches represent support or opposition to proposed legislation.
- Classification can help health care professionals diagnose heart disease patients.
- Based on an e-mail's content, e-mail providers also use classification to decide whether the incoming e-mail messages are spam.

Classification methods: Decision Trees

- A decision tree (also called prediction tree) uses a tree structure to specify sequences of decisions and consequences.
- Given a set of features $X = (x_1, x_2, ..., x_n)$, the goal is to predict a response or output variable Y
- Each member of the set $(x_1, x_2, ..., x_n)$ is called an input variable or feature.

label y/attribute x

Classification methods: Decision Trees



Example of a decision tree.

Classification methods: Decision Trees

- Prediction can be achieved by constructing a decision tree with test points and branches.
- At each test point, a decision is made to pick a specific branch and traverse down the tree.
- Eventually, a final point is reached, and a prediction can be made.

Classification methods: Decision Trees

- Each test point in a decision tree involves testing a particular input variable (or attribute), and each branch represents the decision being made.
- Due to its flexibility and easy visualization, decision trees are commonly deployed in data mining applications for classification purposes.

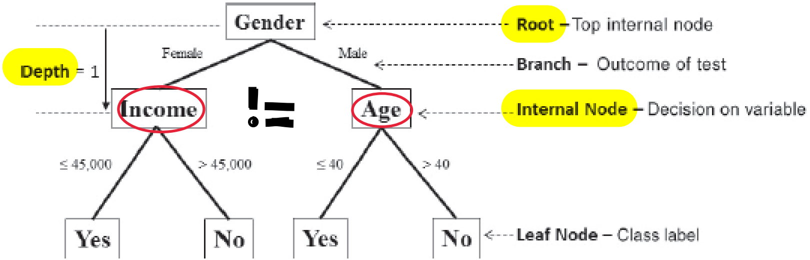
Classification methods: Decision Trees

- The input values of a decision tree can be categorical or continuous.
- A decision tree employs a structure of test points (called nodes) and branches, which represent the decision being made.
- A node without further branches is called a leaf node.
- The leaf nodes return class labels and, in some implementations, they return the probability scores.

Classification methods: Decision Trees

- Classification trees usually apply to output variables that are categorical (often binary) in nature, such as yes or no, purchase or not purchase, and so on.
- They can be easily represented in a visual way, and the corresponding decision rules are quite straightforward.
- We will start with an example with predicting whether customers will buy a product

Classification methods: Decision Trees



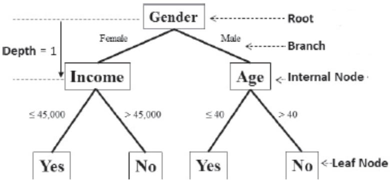
Example of a decision tree

Example of a decision tree. Source: *Data Science & Big Data Analytics*

Classification methods: Decision Trees

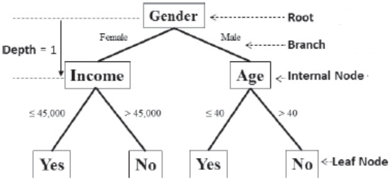
- 'Branch' refers to the outcome of a decision and is visualized as a line connecting two nodes.
- If a decision is numerical, the "greater than" branch is usually placed on the right, and the "less than" branch is placed on the left.
- Depending on the nature of the variable, one of the branches may need to include an "equal to" component.

Classification methods: Decision Trees



- Internal nodes are the decision or test points.
- Each internal node refers to an input variable or an attribute.
- The top internal node is called the root.

Classification methods: Decision Trees

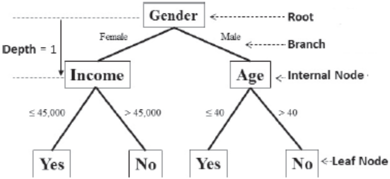


- The decision tree on the left is a binary tree in that each internal node has no more than two branches.
- The branching of a node is referred to as a split.
- The top internal node is called the root.

Classification methods: Decision Trees

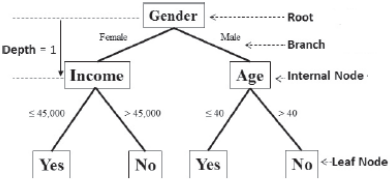
- Sometimes decision trees may have more than two branches stemming from a node.
- For example, suppose an input variable `Weather` is categorical and has three choices: Sunny, Rainy, and Snowy
- Then the corresponding node `Weather` in the decision tree may have three branches labeled as Sunny, Rainy, and Snowy, respectively.

Classification methods: Decision Trees



- The depth of a node is the minimum number of steps required to reach the node from the root.
- In the decision tree on the left, nodes `Income` and `Age` have a depth of one, and the four nodes on the bottom of the tree have a depth of two.

Classification methods: Decision Trees



- Leaf nodes are at the end of the last branches on the tree.
- They represent class labels: the outcome of all the prior decisions.
- The path from the root to a leaf node contains a series of decisions made at various internal nodes.

Classification methods: Decision Trees

- Decision trees are widely used in practice.
- For example, to classify animals, questions (like cold-blooded or warm-blooded, mammal or not mammal) are answered to arrive at a certain classification.
- Another example is a checklist of symptoms during medical evaluation of a patient.
- The artificial intelligence engine of a video game commonly uses decision trees to control the autonomous actions of a character in response to various scenarios.

Classification methods: Decision Trees

- Retailers can use decision trees to segment customers or predict response rates to marketing and promotions.
- Financial institutions can use decision trees to help decide if a loan application should be approved or denied. In the case of loan approval, computers can use the logical if-then statements to predict whether the customer will default on the loan.
- For customers with a clear (strong) outcome, no human interaction is required; for observations that may not generate a clear response, a human is needed for the decision.

Decision Trees: example in R

- Our first example of decision trees in in R concerns a bank that wants to market its term deposit products (such as Certificates of Deposit) to the appropriate customers.
- Given the demographics of clients and their reactions to previous campaign phone calls, the bank's goal is to predict which clients would subscribe to a term deposit.

Decision Trees: example in R

- The dataset 'bank-sample.csv' which has been posted to IVLE contains records of 2000 customers
- The variables include (1) job, (2) marital status, (3) education level, (4) if the credit is in default, (5) if there is a housing loan, (6) if the customer currently has a personal loan, (7) contact type, (8) result of the previous marketing campaign contact (poutcome), and finally (9) if the client actually subscribed to the term deposit.

Decision Trees: example in R

- Attributes (1) through (8) are the input variables or features
- (9) is considered the (binary) outcome: The outcome subscribed is either yes (meaning the customer will subscribe to the term deposit) or no (meaning the customer won't subscribe).
- All the variables listed earlier are categorical.

✗ KNN

Decision Trees: example in R

- Preliminary look at the dataset
- ```
  > bankdata = read.csv("bank-sample.csv", header=TRUE)
  > head(bankdata)
    age          job  marital education default balance
  1  31   management   single  tertiary      no       0
  2  45 entrepreneur  married  tertiary      no    1752
  3  46     services divorced secondary      no    4329
  4  35   management  married  tertiary      no    1108
  5  39   management  married secondary      no    1410
  6  31   management   single  tertiary      no     499
  ```
  ```

Decision Trees: example in R

- Preliminary look at the dataset
- ```
  housing loan  contact day month duration campaign
1   yes   no cellular 15   apr   185   2
2   yes yes cellular 20   nov   56   2
3   no   no cellular 21   nov   534   2
4   yes   no cellular 17   nov   52   1
5   yes   no unknown 23   may   55   1
6   yes   no unknown   9   jun   122   2
  pdays previous poutcome subscribed
1   -1   0 unknown   no
2   -1   0 unknown   no
3   -1   0 unknown   yes
4   -1   0 unknown   no
5   -1   0 unknown   no
6   -1   0 unknown   no
```
```

Decision Trees: example in R

- In R, the package `rpart` contains functions for modeling decision trees
- The optional package `rpart.plot` enables the plotting of a tree.
- We will show how to use decision trees in R to predict which clients would subscribe to a term deposit.
-
```
install.packages("rpart")
install.packages("rpart.plot")
library("rpart")
library("rpart.plot")
```

Decision Trees: example in R

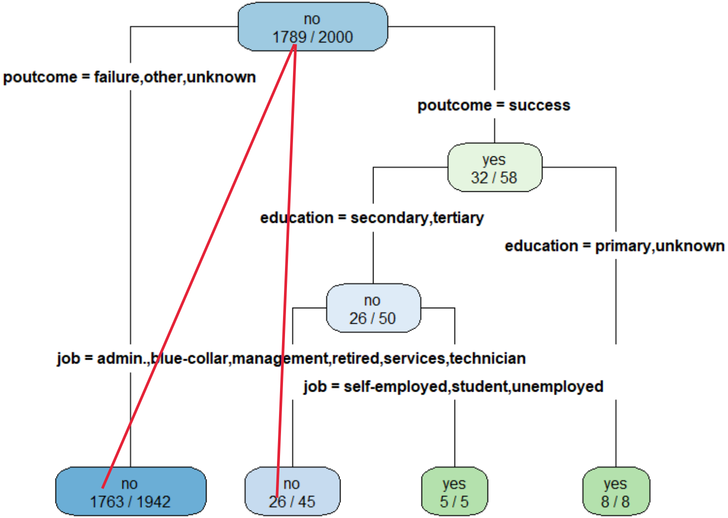
- We will build a decision tree to predict `subscribed` based on the features: `job`, `marital`, `education`, `default`, `housing`, `loan`, `contact` and `poutcome`.
- We will study how the decision tree is fitted in more detail after the recess week
-
```r
fit <- rpart(subscribed ~job + marital + education
+ default + housing + loan + contact + poutcome,
method="class",
data=bankdata,
control=rpart.control(minsplit=1),
parms=list(split='information'))
```

Decision Trees: example in R

- We can visualize the resulting fitted decision tree using `rpart.plot`:

```
rpart.plot(fit, type=4, extra=2, clip.right.labs=
    FALSE, varlen=0, faclen=0)
```

Classification methods: Decision Trees



Decision Trees: example in R

- We will illustrate decision trees using another example: Iris classification dataset

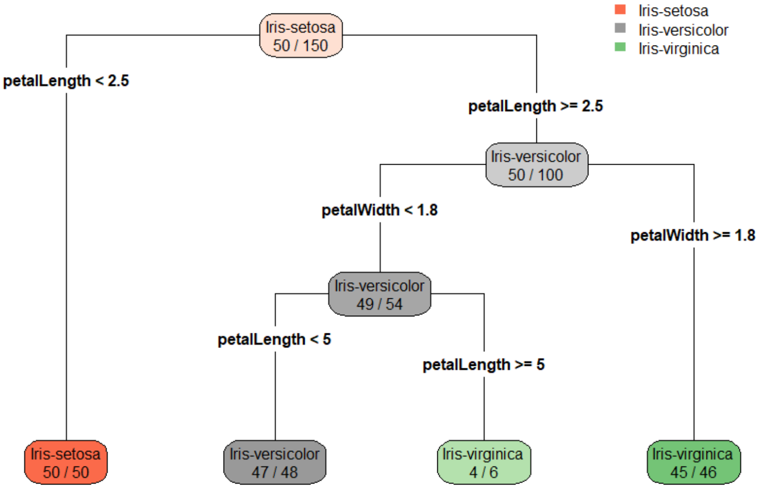
```
iris= read.csv("iris.csv",header=FALSE)
names(iris)= c("sepalLength","sepalWidth",
                 "petalLength","petalWidth","Species
                    ")
```

Decision Trees: example in R

- The task is to predict Iris species based on sepal length / width as well as petal length/width:

```
fit.iris <- rpart(Species ~sepalLength
+sepalWidth+petalLength+petalWidth,
method="class",
data=iris,
control=rpart.control(minsplit=1),
parms=list(split='information'))
rpart.plot(fit.iris, type=4, extra=2, clip.right.
    labs=FALSE, varlen=0, faclen=0)
```

Decision Trees: example in R

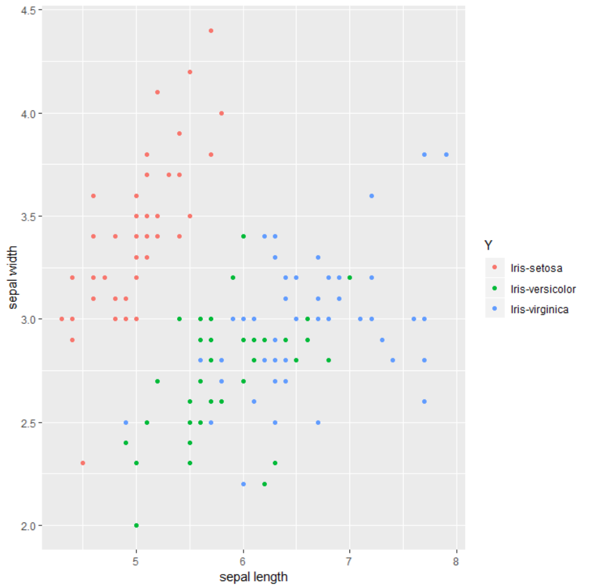


Decision Trees: example in R

- Compare the fitted decision tree with visual plots:
-

```r
library(ggplot2)
library(magrittr)
# sepal width vs. sepal length
ggplot(iris, aes(x=X1, y=X2, color=Y)) +
geom_point()+
labs(x = "sepal length")+  labs(y = "sepal width")
# petal width vs. petal length
ggplot(iris, aes(x=X3, y=X4, color=Y)) +
geom_point()+
labs(x = "petal length")+  labs(y = "petal width")
```

Decision Trees: example in R



Decision Trees: example in R