

4.2. Tree based methods

Tree-based modeling is an exploratory technique for uncovering structure in data, increasingly used for devising prediction rules and summarizing large multivariate datasets. Tree based models are useful for both classification and regression problems. In these problems, there is a set of predictor variable (X), and a single response variable (Y). If Y is a factor, for example, $Y \in \{A, B, C\}$, classification rules are of the form:

$$\begin{aligned}X &\leq 5 \Rightarrow Y = A \\5 < X &\leq 10 \Rightarrow Y = B \\X &> 10 \Rightarrow Y = C\end{aligned}$$

If Y is numeric, regression rules are of the form:

$$\begin{aligned}X &\leq 3 \Rightarrow Y = 3.4 \\3 < X &\leq 5 \Rightarrow Y = 5.6 \\X &> 5 \Rightarrow Y = 7.8\end{aligned}$$

1. Regression tree:

Example (Splus):

```
car.test.frame
help(car.test.frame)
auto.tree1=tree(Mileage~Weight,car.test.frame)
plot(auto.tree1,type="u")
text(auto.tree1)
auto.tree2=tree(Price~Type,car.test.frame)
plot(auto.tree2,type="u")
text(auto.tree2)
auto.tree3=tree(Price~Reliability,car.test.frame,na.action=na.exclude)
plot(auto.tree3,type="u")
text(auto.tree3)
```

2. Classification tree:

Example (Splus):

```
auto.tree4=tree(Country~Price,car.test.frame)
plot(auto.tree4,type="u")
text(auto.tree4)
kyphosis
```

```
help(kyphosis)
auto.tree5=tree(Kyphosis~Age+Number+Start,kyphosis)
plot(auto.tree5,type="u")
text(auto.tree5)
```