Lab 2: Validation Analysis

Tshiamo Phaahla

19/11/2021

Please Note The Following.

The content contained in this document is from the second lab session in a series of lectures prepared for a two-week introductory course in Machine Learning at the University of Cape Town, South Africa. The course is aimed at students with some background in statistical modelling, computing, and linear algebra. It is recommended that you watch the third Key-Point Lecture in this series before tackling the lab session.

Link to the the relevant lab https://youtu.be/Lx0KsiwivKw

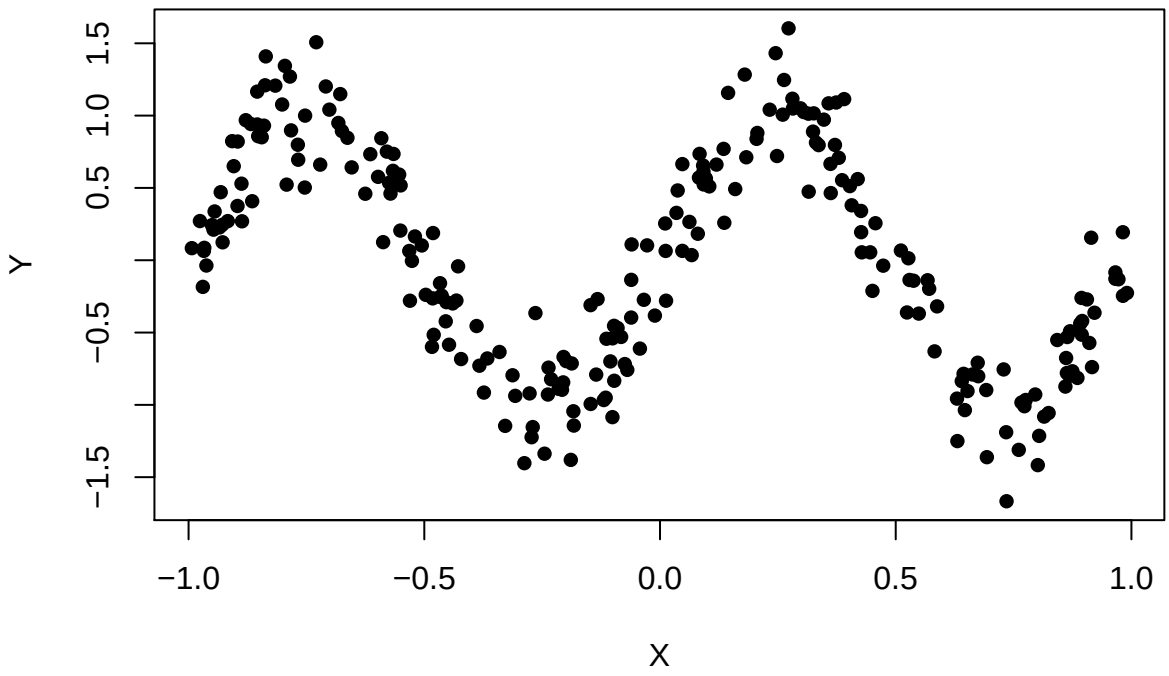
==========

Example 1

Goal: Fit a tree model to the simulated dataset, applying the appropriate complexity controls. Compare the resulting response curve to one fitted using an unconstrained tree.

Given: A simple sinusoidal pattern with noise.

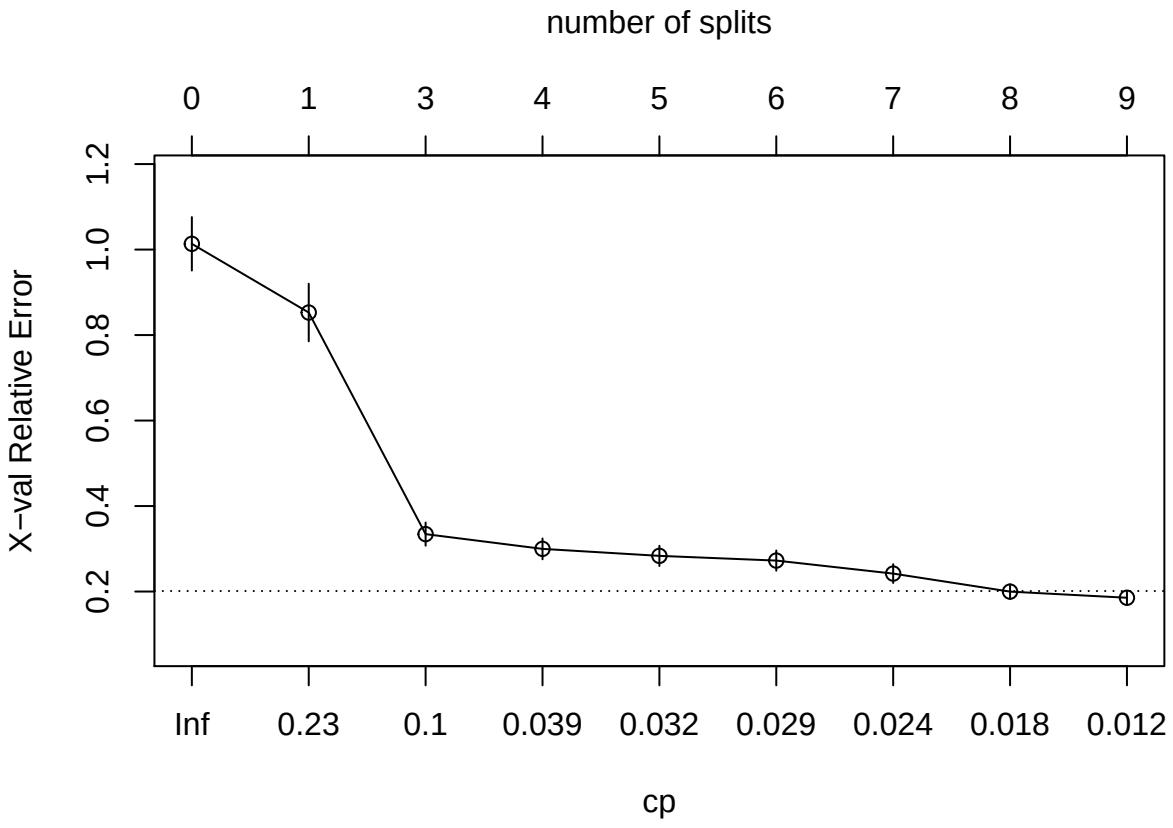
```
set.seed(2021)
N = 250
X = runif(N, -1, 1)
e = rnorm(N, 0, 0.25)
Y = sin(2*pi*X) + e
plot(Y~X, pch=16)
```



Task 1: Fit a tree model and perform a validation analysis in order to select a level of pruning to apply. See ?printcp and ?plotcp.

```
res = rpart(Y~X, control = rpart.control(cp=0.01))

plotcp(res, upper=c("splits"))
```




```
printcp(res)
```

```
##
## Regression tree:
## rpart(formula = Y ~ X, control = rpart.control(cp = 0.01))
##
## Variables actually used in tree construction:
## [1] X
##
## Root node error: 139.54/250 = 0.55814
##
## n= 250
##
##         CP nsplit rel error  xerror     xstd
## 1 0.247447      0   1.00000 1.01344 0.062525
## 2 0.223098      1   0.75255 0.85274 0.067377
## 3 0.047922      3   0.30636 0.33434 0.027106
## 4 0.032071      4   0.25844 0.29980 0.024357
## 5 0.031146      5   0.22636 0.28349 0.023710
## 6 0.027163      6   0.19522 0.27246 0.023673
## 7 0.021877      7   0.16806 0.24203 0.022015
## 8 0.014944      8   0.14618 0.19983 0.017096
## 9 0.010000      9   0.13123 0.18547 0.015745
```

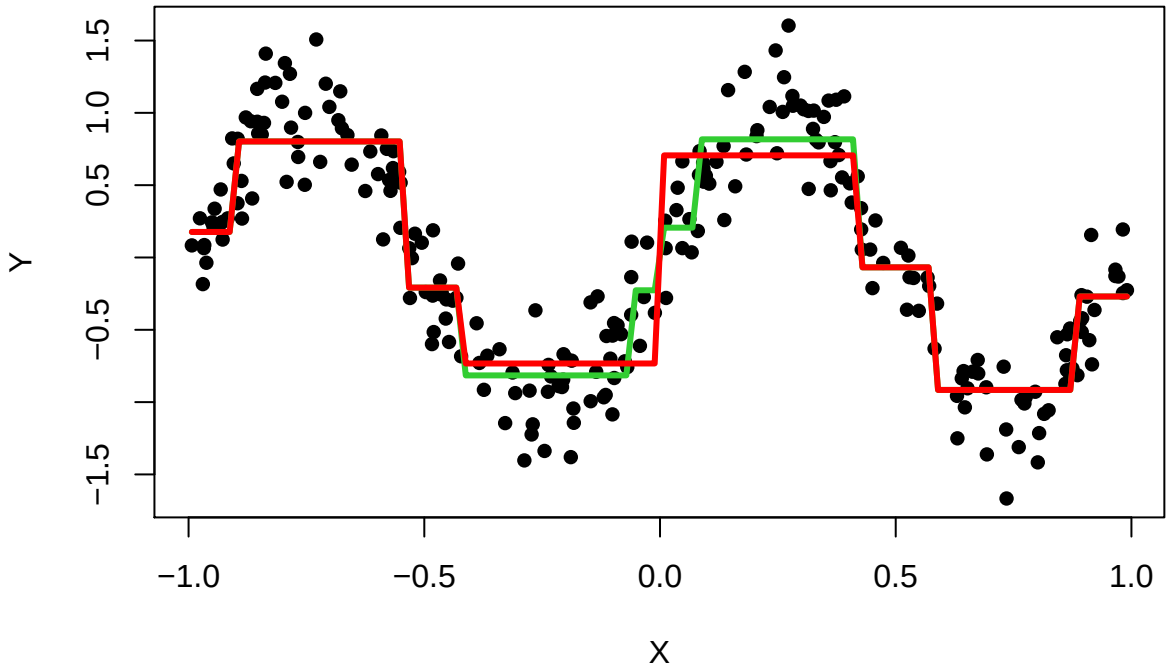
```
res_pruned = prune.rpart(res, cp=0.024)
```

Task 2: Compare the response curves under a unconstrained tree and that of the pruned model.

```
M =100
X_dummy = seq(min(X), max(X), length=M)
Lat = data.frame(X = X_dummy)

plot(Y~X, pch=16)
predY = predict(res, Lat)
predY_pruned = predict(res_pruned, Lat)

lines(predY~Lat$X, col="limegreen", lwd=3)
lines(predY_pruned~Lat$X, col="red", lwd=3)
```



Some Interpretation The unconstrained model follows the data more closely. This means that is accounts for some of the noise and is replicating characteristics which are unique to this data set. When this happens we are said to be overfitting our model. What we want is the constrained model which follows the underlying pattern of the data.

Example 2:

Goal: Analyse the ptitanic data from the rpart.plot package using an appropriately chosen tree-based model.

Task 1: Calculate the relative frequency of binary-encoded response conditional on the sex and pclass variables (Intersections as well). What do these reveal about the data?

We will start by creating a boolean vector of our response variable.

```
data(ptitanic)
attach(ptitanic)

Y = (survived == 'survived')

mean(Y[sex=='male'])
```

```
## [1] 0.1909846
```

```
mean(Y[sex=='female'])
```

```
## [1] 0.7274678
```

```
mean(Y[pclass=='1st'])
```

```
## [1] 0.619195
```

```
mean(Y[pclass=='2nd'])
```

```
## [1] 0.4296029
```

```
mean(Y[pclass=='3rd'])
```

```
## [1] 0.2552891
```

```
mean(Y[(pclass=='1st')&(sex=='male')])
```

```
## [1] 0.3407821
```

```
mean(Y[(pclass=='2nd')&(sex=='male')])
```

```
## [1] 0.1461988
```

```
mean(Y[(pclass=='3rd')&(sex=='male')])
```

```
## [1] 0.1521298
```

```
mean(Y[(pclass=='1st')&(sex=='female')])
```

```
## [1] 0.9652778
```

```
mean(Y[(pclass=='2nd')&(sex=='female')])
```

```
## [1] 0.8867925
```

```
mean(Y[(pclass=='3rd')&(sex=='female')])
```

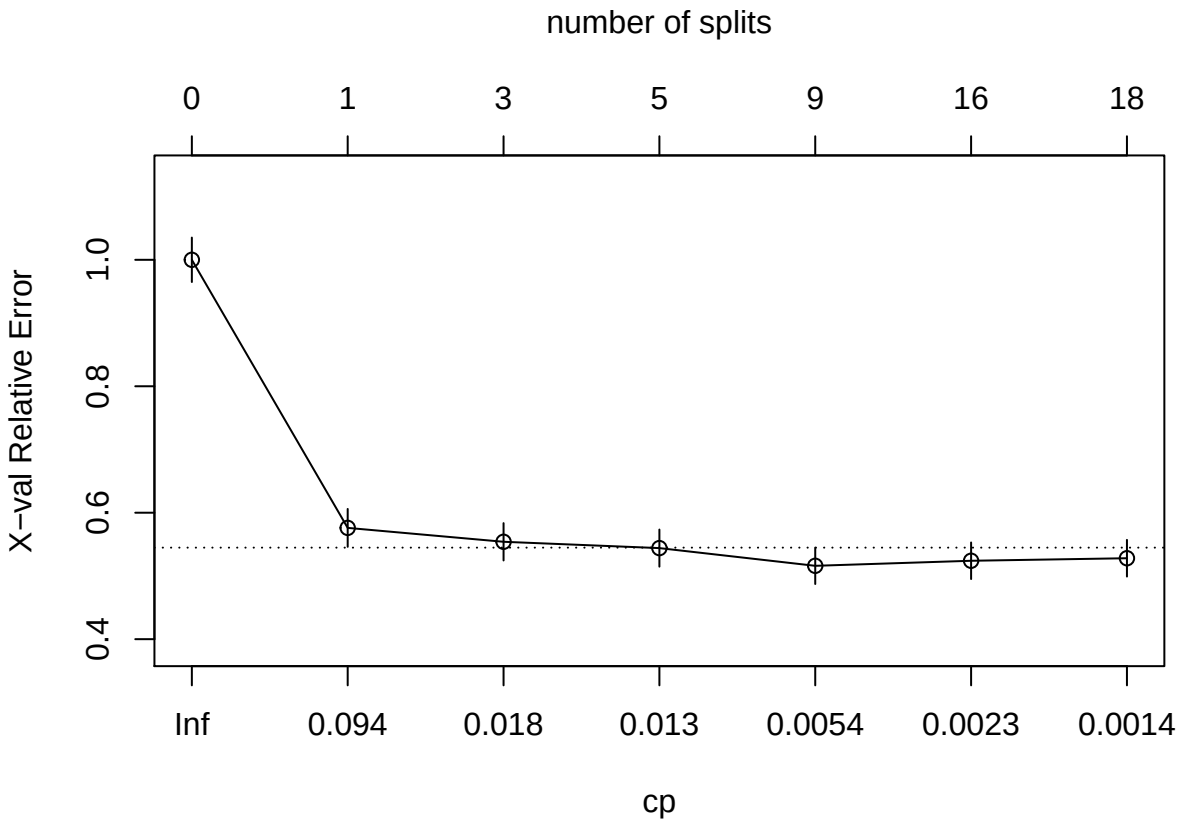
```
## [1] 0.4907407
```

Initial Comments The survival rate of males in significantly lower (19%) than of females (73%). The first class passengers have the highest survival rate at 62%, compared to the second and third class passengers at 43% and 26% respectively.

The intersections show that males in first class had a higher survival rate than those in second and third class. The same can be said for females although the general trend of males having a lower survival rate is still evident event when taking into account intersections.

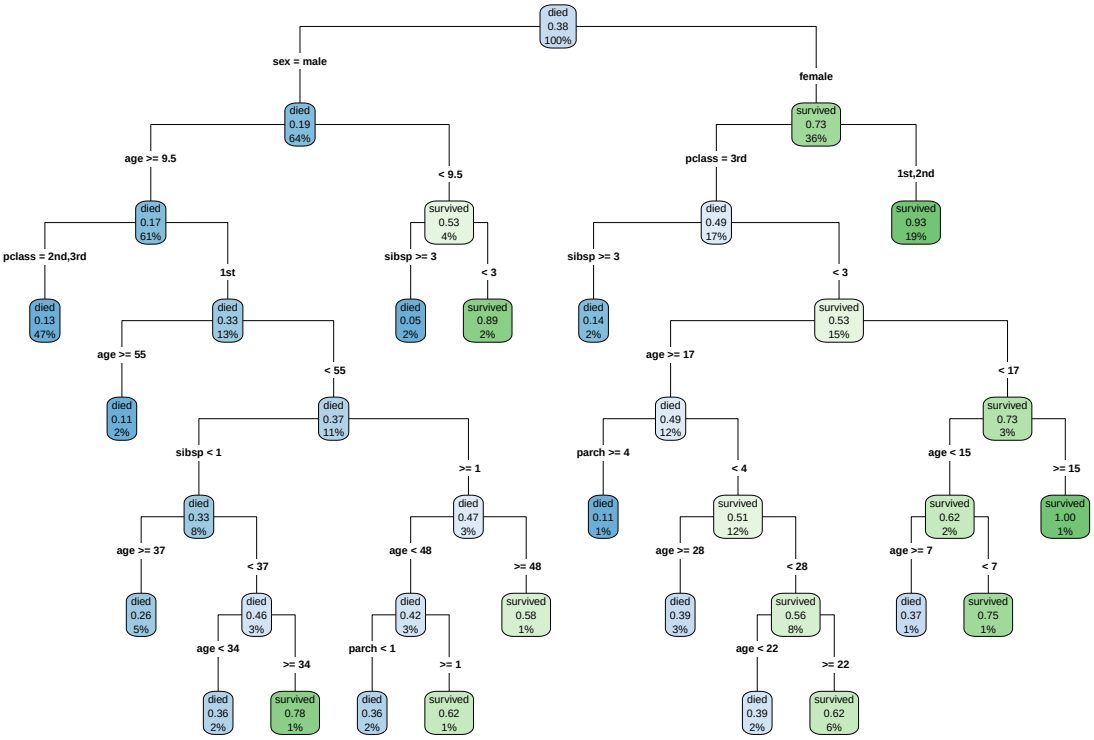
Task 2: Fit a tree model using all of the available inputs and interpret the resulting model.

```
set.seed(2021)
model = rpart(survived~., data = ptitanic, control= rpart.control(cp=0.001))
plotcp(model, upper=c("splits"))
```

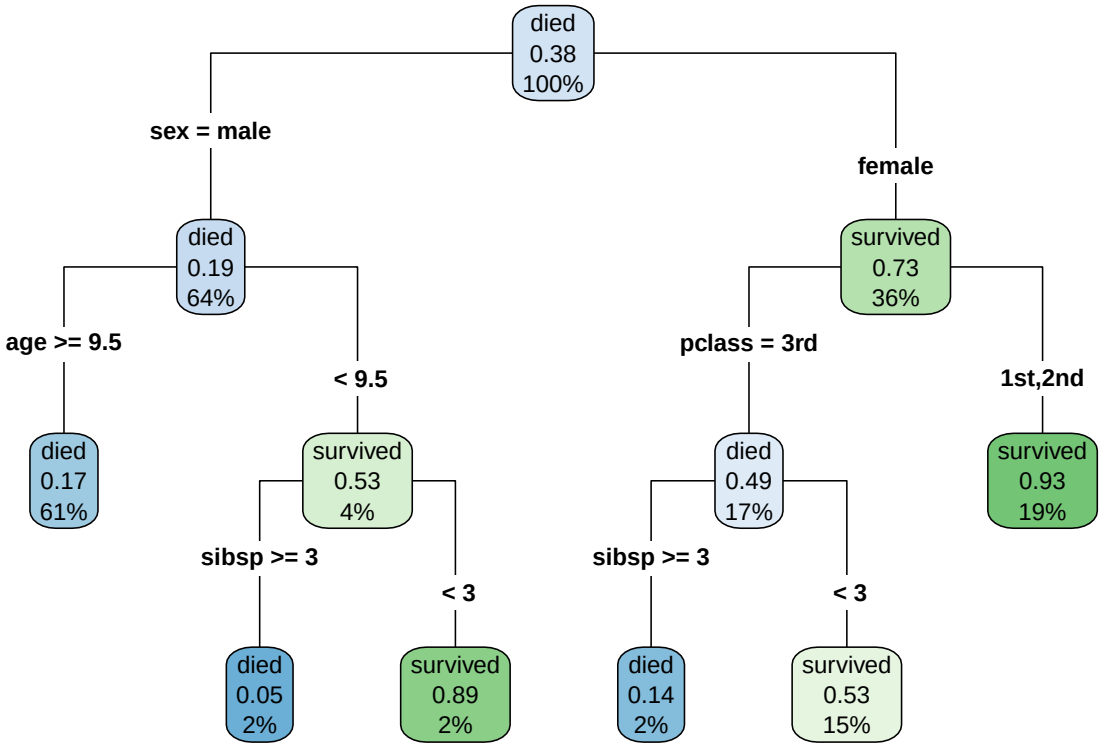



```
model_pruned = prune.rpart(model, cp=0.013)

rpart.plot(model, type=4, fallen.leaves=F, branch=1)
```




```
rpart.plot(model_pruned, type=4, fallen.leaves = F, branch=1)
```



Some Interpretation

In root node we see that you are more likely to have not survived. If you are a male then survival rate drops significantly. Females were more likely to survive then not. Shouldn't be a surprise given the major discrepancies between the empirical probabilities of survival based on sex. From our pruned model we can see that the most significant predictor in determining survival is `sex`. The females had a higher survival rate than males.

`Age` was a significant predictor, but only for males. For females the most significant predictor in the class. If you were a younger male then your survival probability increases. On the female side of the tree, we see that you're more likely to survive than not. If you are in 1st and 2nd class, then you're much more likely to have survived. Unfortunately if you are a third class passenger then the same cannot be said.

This more or less follows from what we saw in the EDA. Here, because we have an algorithm that is tied to an objective function and we've conducted a validation analysis with a predictive approach in mind. We can see that these variables are indeed predictive in whether you have survived or not.

We also know that there are class discrepancies but these discrepancies are drowned out if you are male simply because the survival rate is just so low anyways. On the female side they are predictive.