Homework 4: Decision Trees

1 Wisconsin Breast Cancer Data (10pts)

You can download the data session4_homework.txt on Piazza/Resources.

Step1: Read the data

```
##        Id V1 V2 V3 V4 V5 V6 V7 V8 V9 Class
## 1 1000025  5  1  1  1  2  1  3  1  1     2
## 2 1002945  5  4  4  5  7 10  3  2  1     2
## 3 1015425  3  1  1  1  2  2  3  1  1     2
## 4 1016277  6  8  8  1  3  4  3  7  1     2
## 5 1017023  4  1  1  3  2  1  3  1  1     2
## 6 1017122  8 10 10  8  7 10  9  7  1     4
```

Columns of the data frame are:

- `Id` - subject Id
- `V1` to `V9` - attributes
- `Class` - diagnosis: 4 = malignant, 2 = benign

Some observations have missing values “?”.

```
sum(datBreastCancer=="?")
```

```
## [1] 16
```

Total dimension of the data set and the number of malignant diagnoses:

```
dim(datBreastCancer)
```

```
## [1] 699  11
```

```
(p4<-sum(datBreastCancer$Class==4)/dim(datBreastCancer)[1])
```

```
## [1] 0.3447783
```

Step2: Fit classification tree.

Step3: Prune the tree to make analysis easier.

Step4: Comment the trees and calculate probabilities

1. Observe the tree.
2. Define event E_1E1 as `E1={(V2>=2.5),(V3>=2.5),V6!=<?,1,2>}`. Calculate probability $P(E_1)$P(E1)
3. Calculate probability $P(4|E_1)$P(4|E1)
4. Define event $E(V6=2)$E(V6=2) as all combinations of events (paths on the tree) including event $V6=2$V6=2. Calculate probability $P(4|E(V6=2))$P(4|E(V6=2))
5. Define event $E(V3=3)$E(V3=3) as all combinations of events (paths on the tree) satisfying $V3=3$V3=3. Calculate probability $P(4|E(V3=3))$P(4|E(V3=3))
6. Calculate probability $P(E_1|4)$P(E1|4) using Bayes Theorem and directly from the observed data

2 Time series of stock prices (10pts)

Predict returns of exchange traded fund SPY representing S&P 500 with a group of stock returns of companies in the index.

Select year 2014.

Download the file session4_spyPortfolio.csv on Piazza/Resources.

Step1: Read the file

```
##     SPLS.A    MTB.A    UNM.A    VLO.A AMZN.A ADBE.A    CSX.A     PG.A
## 1 13.35805 106.3561 31.95808 44.67762 397.97  59.29 26.06945 71.50668
## 2 13.52942 106.4948 31.96739 44.21176 396.44  59.16 26.23555 71.42678
## 3 13.13528 106.1620 31.92084 44.64179 393.63  58.12 26.04176 71.59547
##      CMA.A    PEP.A    DNB.A   MDLZ.A    SYY.A   VIAB.A    MDT.A      T.A
## 1 44.10851 74.31774 113.9543 32.85969 32.90575 79.78655 53.40803 28.78814
## 2 44.32504 74.44447 115.0964 32.80304 33.02484 79.46537 54.43439 28.66459
## 3 44.24972 74.48068 113.7090 32.59536 32.97904 78.66707 55.25548 28.79637
##      CHK.A    MRO.A    TMK.A  MU.A AABA.A    MON.A AKAM.A    HSY.A
## 1 24.55435 32.30710 51.62000 21.66  39.59 108.2317  46.53 88.48643
## 2 24.36987 31.94617 51.72000 20.97  40.12 108.2690  46.45 88.37573
```

```
## 3 24.16694 31.86289 51.53333 20.67  39.93 107.6455  46.11 88.12668
##       BAC.A    WDC.A     EL.A    SPY.A
## 1 15.44962 75.43743 70.42357 170.5143
## 2 15.74710 75.98487 70.36625 170.4863
## 3 15.98700 75.62904 70.77708 169.9923
```

Step2: Create daily log returns of all stocks and SPY.

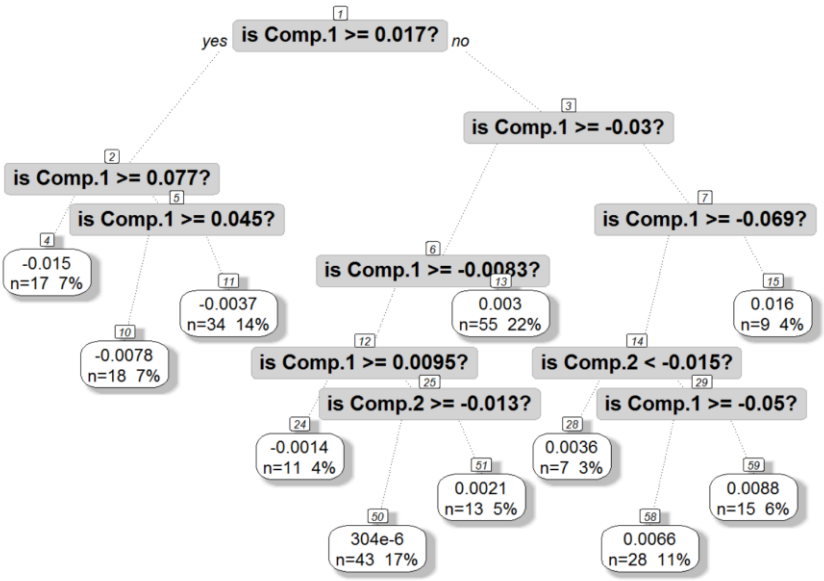
Step3: Make daily log returns of all stock prices lagged one day relative to the daily log returns of SPY.

```
#R Equivalent
SPYPortf<-log(SPYPortf)
SPYPortf<-apply(SPYPortf,2,diff)
```

Step4: Apply PCA

Step5: Grow regression tree from the PCA components

You will get something which looks like that:



Step6: Prune the tree.

Step7: Interpret the tree.

Step8: Create vector of predictions by pruned tree and plot the graph