Machine Learning
Sommersemester2020
Exercise 5

Ciheng Zhang (3472321) zch3183505@gmail.com
Gang Yu(3488292) HansVonCq@gmail.com
Huipanjun Tian (3471607) Thpjpyl5111217@gmail.com

June 2, 2020

Machine Learning
Sommersemester2020
Exercise 5

Ciheng Zhang(3472321)
Gang Yu(3488292)
Huibanjun Tian(3471607)

1 Inductive Construction

according to the table, At first we calculate the Entropy of the dataset.

$$P(-) = \frac{125}{240} P(+) = \frac{115}{240}$$

$$Ent(D) = -(P(-)log_2(P(-)) + P(+)log_2(p(+))) = 0.9987$$

Then we try try to divide the dataset by different features and calculate the Gain of each features, and choose one feature to divide the dataset.
BY F1:

F1=0		F1=1	
-	+	-	+
70	50	55	65

$$Ent(F1 = 0) = 0.9798$$

$$Ent(F1 = 1) = 0.9950$$

$$Gain(F1) = Ent(D) - \frac{D(F1 = 0)}{D} Ent(F1 = 0) - \frac{D(F1 = 1)}{D} Ent(F1 = 1) = 0.0113$$

BY F2:

F2=0		F2=1	
-	+	-	+
50	70	75	45

$$Ent(F2 = 0) = 0.9798$$

$$Ent(F2 = 1) = 0.9944$$

$$Gain(F2) = 0.0116$$

BY F3:

F3=0		F3=1		F3=2	
-	+	-	+	-	+
65	15	50	30	10	70

$$Ent(F3 = 0) = 0.6962$$

$$Ent(F3 = 1) = 0.9544$$

$$GEnt(F3 = 2) = 0.5436$$

$$Gain(F3) = 0.2673$$

Ciheng Zhang(3472321)
Gang Yu(3488292)
Huibanjun Tian(3471607)

BY F4:

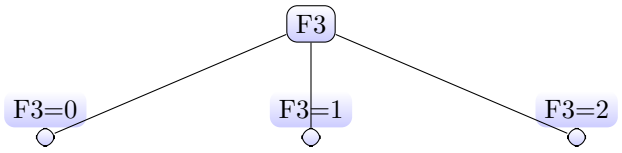
F4=0		F2=1	
-	+	-	+
70	50	65	65

$$Ent(F2 = 0) = 0.9798$$

$$Ent(F2 = 1) = 1$$

$$Gain(F2) = 0.008$$

because of $Gain(F3) > Gain(F2) > Gain(F1) > Gain(F4)$,we choose F3 as the feature:



Then we rebulid the table for each Note:

F3 =0				
F1	F2	F4	-	+
0	0	0	10	0
0	0	1	5	5
0	1	0	10	0
0	1	1	10	0
1	0	0	0	10
1	0	1	10	0
1	1	0	10	0
1	1	1	10	0

F3 =1				
F1	F2	F4	-	+
0	0	0	10	0
0	0	1	10	0
0	1	0	5	5
0	1	1	0	10
1	0	0	5	5
1	0	1	0	10
1	1	0	10	0
1	1	1	10	0

F3 =2				
F1	F2	F4	-	+
0	0	0	0	10
0	0	1	0	10
0	1	0	10	0
0	1	1	0	10
1	0	0	0	10
1	0	1	0	10
1	1	0	0	10
1	1	1	0	10

Then we calculate for the first Node(F3=0):

$$Ent(F3 = 0) = 0.6962$$

BY F1:

F1=0		F1=1	
-	+	-	+
35	5	30	10

$$Ent(F1 = 0) = 0.5435$$

$$Ent(F1 = 1) = 0.8112$$

$$Gain(F2) = 0.01885$$

Ciheng Zhang(3472321)
Gang Yu(3488292)
Huibanjun Tian(3471607)

BY F2:

F2=0		F2=1	
-	+	-	+
25	15	40	0

$$Ent(F2 = 0) = 0.9544$$

$$Ent(F1 = 1) = 0$$

$$Gain(F2) = 0.219$$

BY F4:

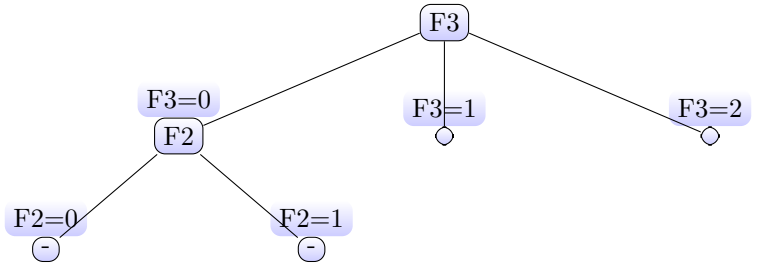
F4=0		F4=1	
-	+	-	+
30	10	35	5

$$Ent(F4 = 0) = 0.8112$$

$$Ent(F4 = 1) = 0.5435$$

$$Gain(F4) = 0.01885$$

because of $Gain(F2) > Gain(F1) = Gain(F4)$,we choose F2 as the feature:



Then we calculate for the second Node(F3=1):

$$Ent(F3 = 1) = 0.9544$$

BY F1:

F1=0		F1=1	
-	+	-	+
25	15	25	15

Ciheng Zhang(3472321)
Gang Yu(3488292)
Huibanjun Tian(3471607)

$$Ent(F1 = 0) = 0.9544$$
$$Ent(F1 = 1) = 0.9544$$
$$Gain(F1) = 0$$

BY F2:

F2=0		F2=1	
-	+	-	+
25	15	25	15

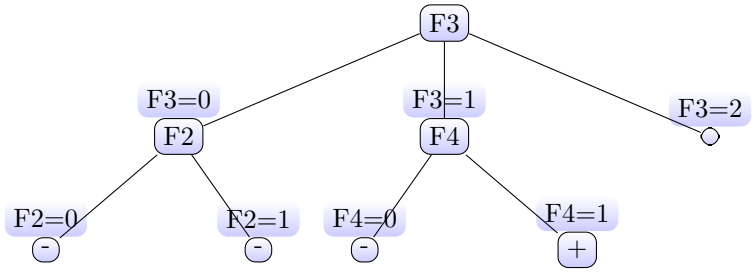
$$Ent(F2 = 0) = 0.9544$$
$$Ent(F2 = 1) = 0.9544$$
$$Gain(F2) = 0$$

BY F4:

F4=0		F4=1	
-	+	-	+
30	10	20	20

$$Ent(F4 = 0) = 0.8112$$
$$Ent(F4 = 1) = 1$$
$$Gain(F4) = 0.0488$$

because of $Gain(F4) > Gain(F1) = Gain(F2)$,we choose F4 as the feature:



Then we calculate for the second Node(F3=2):

$$Ent(F3 = 2) = 0.5436$$

BY F1:

Ciheng Zhang(3472321)
Gang Yu(3488292)
Huibanjun Tian(3471607)

F1=0		F1=1	
-	+	-	+
10	30	0	40

$$Ent(F1 = 0) = 0.8113$$

$$Ent(F1 = 1) = 0$$

$$Gain(F1) = 0.1379$$

BY F2:

F2=0		F2=1	
-	+	-	+
0	40	10	30

$$Ent(F2 = 0) = 0.8113$$

$$Ent(F2 = 1) = 0$$

$$Gain(F2) = 0.1379$$

BY F4:

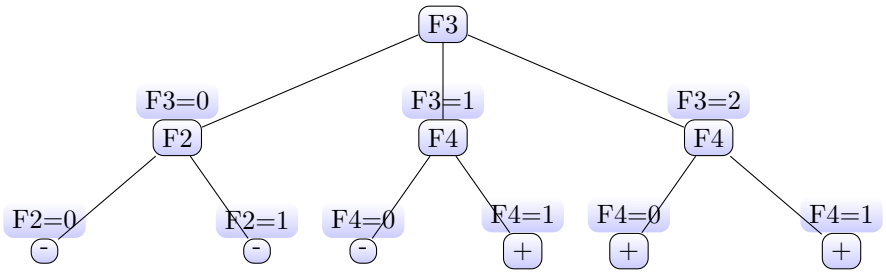
F4=0		F4=1	
-	+	-	+
10	30	0	40

$$Ent(F4 = 0) = 0.8113$$

$$Ent(F4 = 1) = 0$$

$$Gain(F4) = 0.1379$$

because of $Gain(F4) = Gain(F1) = Gain(F2)$,we choose F4 as the feature:



So we calculate the error rate for this tree:

$$E = \frac{55}{240} = 0.229$$

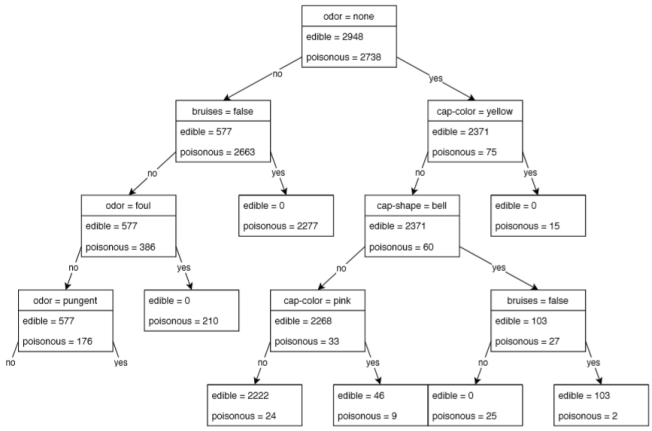
Ciheng Zhang(3472321)
Gang Yu(3488292)
Huibanjun Tian(3471607)

2 Minimal Error Pruning

First we calculate the error rate of original tree:

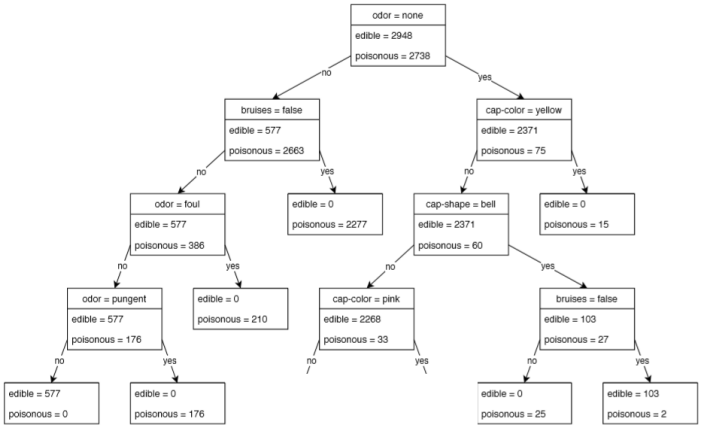
$$E(origal) = \frac{1}{n(T)}\Sigma_{t\in T}e(t) = 0.00615$$

Then first time Pruning:



Then we calculate the Error rate:

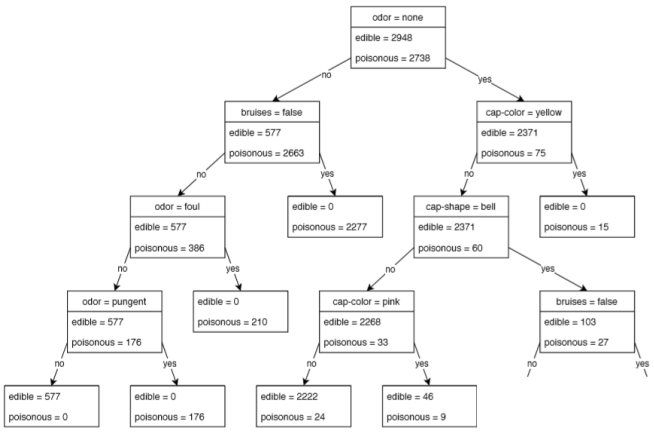
$$E(P1) = 0.0371$$



Then we calculate the Error rate:

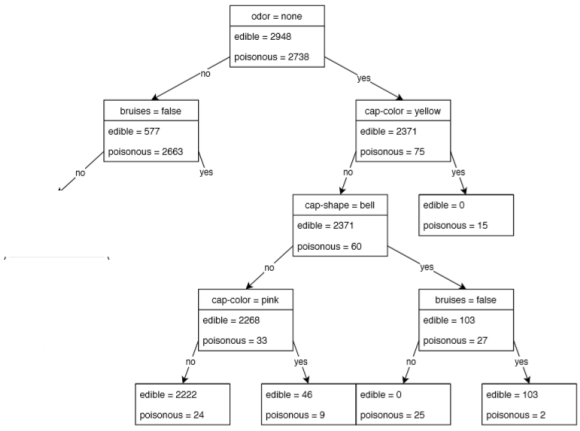
Ciheng Zhang(3472321)
Gang Yu(3488292)
Huibanjun Tian(3471607)

$$E(P2) = 0.0061$$



Then we calculate the Error rate:

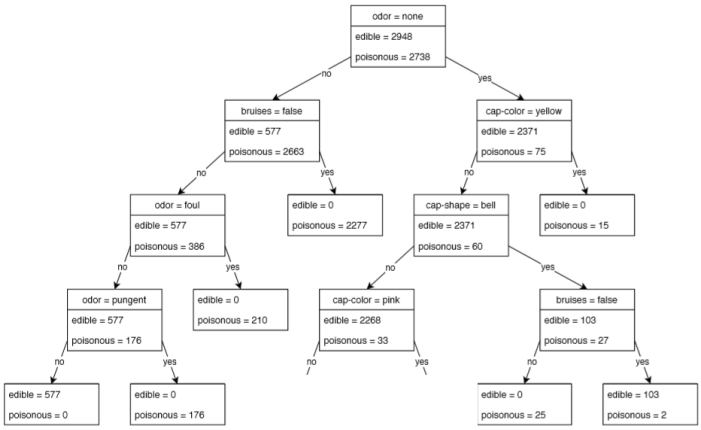
$$E(P3) = 0.0106$$



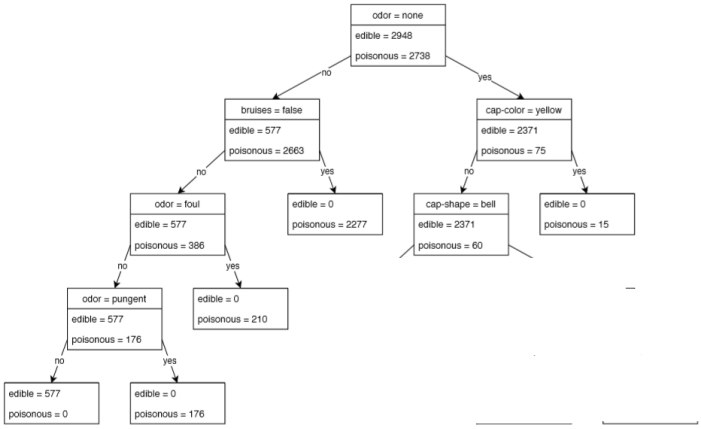
$$E(P4) = 0.107$$

because of$E(P2) < E(P3) < E(P1)$ we Pruning the second viable node,and get the new tree:

Ciheng Zhang(3472321)
Gang Yu(3488292)
Huibanjun Tian(3471607)



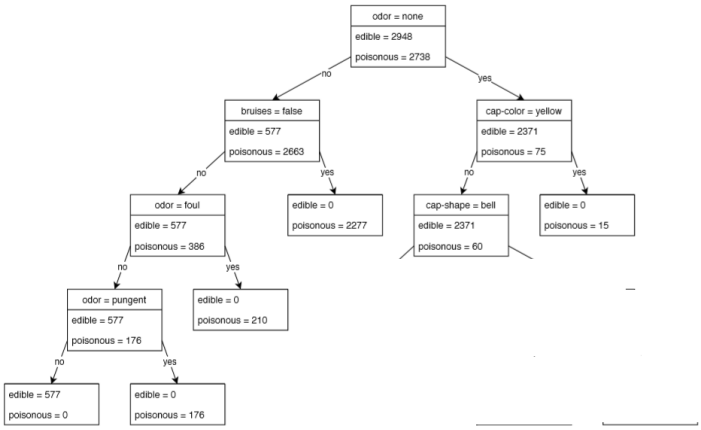
Then we calculate for new viable node:



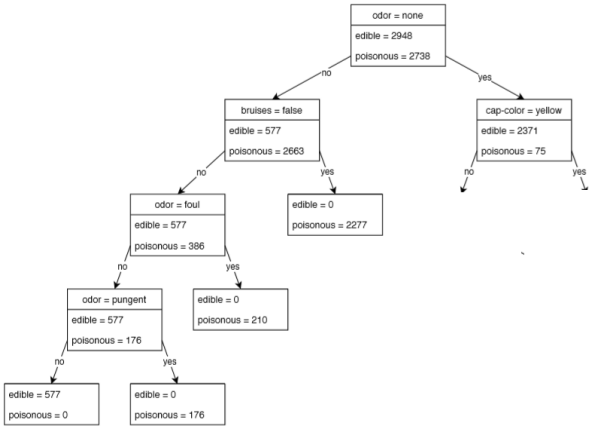
$$E(P5) = 0.0106$$

because of $E(P5) = E(P3) < others$ we prun the viable node of last time. and become a new tree:

Ciheng Zhang(3472321)
Gang Yu(3488292)
Huibanjun Tian(3471607)



Then we repeat the uper method:



$$E(P6) = 0.0132$$

then we need to remove the node of last time. because $E(P6) < others$. And we get the final tree:

Ciheng Zhang(3472321)
Gang Yu(3488292)
Huibanjun Tian(3471607)

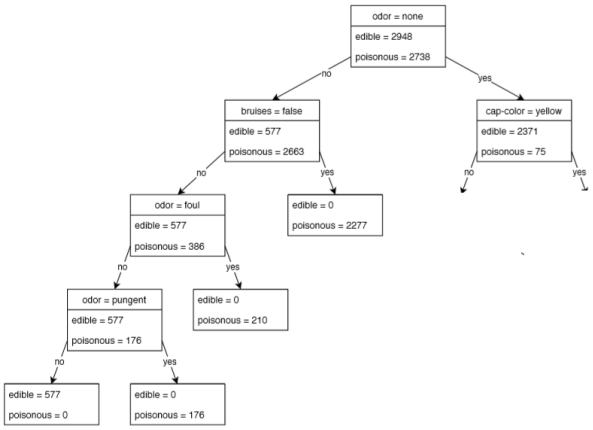


Figure 1: Final Tree

3 Regression with Decision Tress and kNN

A regression tree is aim to one regression space. The classification tree's Node is features, and the viable node of classification tree is Tags of classification. But all of the regression tree'node is a regression vlaue, and the viable node of the regression tree is the best regression value. Then when we want to bulid a regression tree, we need to transver all the input values and calculate the distance.And we use the distance to bulid a loss function. Then we Minimal the loss function and get the vlaue of next node. But for a classification tree we calculate and compare the Entropy Gain to decide the next Node.

A kNN use for a regression problem, at first we need to calculate a point and get the nearest k points. Then the output values is the mean value of those k nearest points. Then we fit all of the output points and solve a regression problem.

References

[1] https://blog.csdn.net/u012328159/article/details/70184415

[2] https://www.jianshu.com/p/7c385f268bf9

[3] https://blog.csdn.net/on2way/article/details/88673075