Comparing Predictive Models Report

Tingjun Li

li400@purdue.edu

1. Algorithm Details

(a)

Y: Riskiness/ Symboling

M: make; FT: fuel-type; A: aspiration; ND: num-of-doors; B: body-style;

D: drive-wheels; EL: engine-location; ET: engine-type; NC: num-of-cylinders; FS: fuel-system;

P(Y|M,FT,A,ND,B,D,EL,ET,NC,FS) = P(M|Y)P(FT|Y)P(A|Y)P(ND|Y)P(B|Y)P(D|Y)P(EL|Y) P(ET|Y)P(NC|Y)P(FS|Y)/P(M,FT,A,ND,B,D,EL,ET,NC,FS)

(b)In orther to solve this formula, we need to solve P(X|Y) and P(X), X is an attribute.

for P(X|Y): we need to know **total high risk(positive) number of Y** and **the number of X that leads to high risk(positive) Y**.

for P(X): **number of cases containing X, size of the training se**t.

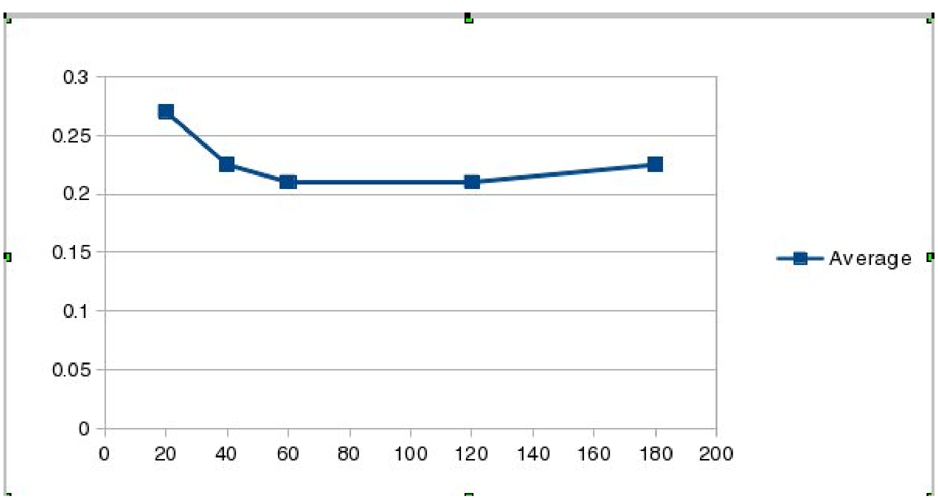
Negative Y probability can be obtained through 1-P(X|Y)

(c)Maximum likelihood estimate for V8(drive-wheels)

3. Evaluation

Training set size	20	40	60	120	180
#1	0.25	0.35	0.15	0.3	0.25
#2	0.3	0.2	0.1	0.2	0.3
#3	0.4	0.1	0.15	0.3	0.2
#4	0.25	0.1	0.35	0.3	0.25
#5	0.25	0.2	0.2	0.2	0.25
#6	0.3	0.3	0.1	0.1	0.1
#7	0.25	0.25	0.2	0.1	0.1
#8	0.25	0.2	0.25	0.1	0.1
#9	0.15	0.2	0.35	0.05	0.35
#10	0.3	0.35	0.25	0.45	0.35
Average	0.27	0.225	0.21	0.21	0.225

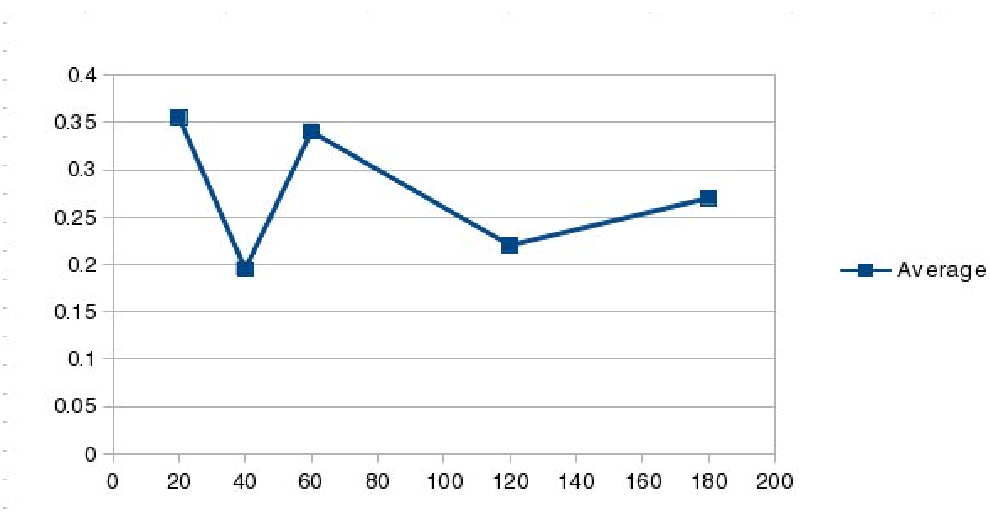
Learning Curve for NBC:



4. Compare to the regression tree

Training Set Size	20	40	60	120	180
#1	0.25	0.35	0.35	0.0	0.05
#2	0.5	0.4	0.3	0.05	0.3
#3	0.25	0.25	0.3	0.1	0.4
#4	0.3	0.1	0.65	0.3	0.45
#5	0.55	0.35	0.5	0.25	0.05
#6	0.05	0.05	0.6	0.15	0.6
#7	0.6	0.0	0.6	0.05	0.7
#8	0.55	0.0	0.05	0.25	0.05
#9	0.0	0.45	0.0	0.3	0.05
#10	0.5	0.0	0.05	0.75	0.05
Average	0.355	0.195	0.34	0.22	0.27

Learning Curve for regression tree:



Analysis:

Comparing the learning curve from two different algorithms, we can see that Bayes classifier has better performance(lower zero-one loss and more stable performance).

Besides the general performance of the two algorithms, we can also see the effect of changing the training set size. Generally speaking, increasing the size of training set will lead to a higher accuracy(lower zero-one loss), but we can see there will be problem when training set size is too big. In both algorithm, we can see the largest size(180) cannot produce optimal result. I think it is the overfitting problem here.