Applied Machine Learning Homework 1

Abhinav Malik

February 10, 2018

Solution 1 a):
Let the initial entropy be E(S).

E(S)=-35/50 log_2(35/50) - 15/50 log_2(15/50)
E(S)=0.882
Information gain for Sunny weather

Entropy(Sunny=yes)=-20/21 log_2(20/21) -1/21 log_2(1/21)
Entropy(Sunny=yes)=0.2737

Entropy(Sunny=no)=-15/29 log_2(15/29) -14/29 log_2(14/29)
Entropy(Sunny=yes)=0.99

The information gain for splitting on Weather=Sunny is

Gain=0.882 -35/50(0.2737) -15/50(0.99)
Gain=0.39071

Information gain for Snowy weather

Entropy(Snow=yes)=-25/35 log_2(25/35) -10/35 log_2(10/35)
Entropy(Snow=yes)=0.8642

Entropy(Snow=no)=-10/15 log_2(10/15) -5/15 log_2(5/15)
Entropy(Snow=no)=0.92834

The information gain for splitting on Weather=Snow is

Gain=0.882 -35/50(0.8642) -15/50(0.92834)
Gain=0.002

We should split the data on Weather=Sunny for maximum information gain.

Solution 1 b): The decision tree for the given dataset is-

```
if color=blue:
    if size=small:
        inflated=F
    if size!=small:
        if act=dip:
            inflated=T
        if act!=dip:
            if age=adult:
                inflated=F
            if age!=adult:
                inflated=T
if color!=blue
    if act=dip:
        inflated=T
    if act!=dip
        if age=adult
            inflated=F
        if age!=adult
            inflated=T
```

Solution 1 c): No, ID3 doesn't guarantee a globally optimal decision tree. The ID3 algorithm looks for the optimal solution locally and not globally. An attribute which gives the maximum information gain locally, might not be the best solution globally. Sometimes it makes sense to select an attribute which works well in the long term but not well enough at the current level which ID3 fails to implement.

Solution 2:

1) SGDClassifier(loss="log",penalty="l2",alpha=0.001,learning_rate="optimal",eta0=0.73)
The values of p_a are-

- 0.63571
- 0.6642
- 0.5928
- 0.5857
- 0.5857

The average p_a is- 0.612822
The values of tr_a are-

- 0.62
- 0.62

- 0.6071
- 0.6242
- 0.6171

The average tr_a is- 0.61768

The 99% interval is-[0.5467,0.6773]

2) DecisionTreeClassifier
The values of p_a are-

- 0.5214
- 0.5214
- 0.5
- 0.5642
- 0.5071

The average p_a is- 0.52282
The values of tr_a are-

- 0.64
- 0.6485
- 0.63285
- 0.64571
- 0.63428

The average tr_a is- 0.640268

The 99% interval is-[0.4767,0.5673]

3) DecisionTreeClassifier with max depth 4
The values of p_a are-

- 0.4785
- 0.45
- 0.4785
- 0.5357
- 0.5285

The average p_a is- 0.49424
The values of tr_a are-

- 0.5357
- 0.5542
- 0.5342
- 0.5457
- 0.5614

The average tr_a is- 0.54624

The 99% interval is-[0.4281,0.5599]

4) DecisionTreeClassifier with max depth 8
The values of p_a are-

- 0.48571
- 0.4571
- 0.55
- 0.5714
- 0.4928

The average p_a is- 0.5114
The values of tr_a are-

- 0.56285
- 0.5914
- 0.57857
- 0.5842
- 0.57571

The average tr_a is- 0.5785

The 99% interval is-[0.4423,0.5917]

5) Decision stumps as features
The values of p_a are-

- 0.528571
- 0.45714

- 0.528571
- 0.578571
- 0.54285

The average p_a is- 0.52714

The 99% interval is-[0.4459,0.6081]

Conclusion: According to the computed p_a values, SGDClassifier is the best classifier. The order of the classifier is-

- 0.612822(SGDClassifier)
- 0.52714(Decision stumps as features)
- 0.52282(DecisionTreeClassifier)
- 0.5114(DecisionTreeClassifier with max depth 8)
- 0.49424(DecisionTreeClassifier with max depth 4)

A decision tree with no restriction on the maximum depth should be better than a decision tree with restrictions and this is clearly reflected in the results obtained. A decision tree with maximum depth of 4 will not be as effective which is why it's the poorest of them all. For the given data set a linear classifier would've worked best which is backed up by the result wherein SGD had the best p_a.