CS446: Machine Learning	**Fall 2014**
Problem Set 2	
Chen Zhang	*Handed In: 09/25/2014*

1 Problem 1

1.1

1.1.1 Calculation

The total entropy is $-\frac{7}{16}\log\frac{7}{16}-\frac{9}{16}\log\frac{9}{16}=0.99$
Entropy for yellow is $-\frac{5}{8}\log\frac{5}{8}-\frac{3}{8}\log\frac{3}{8}=0.95$. The entropy for purple is $-\frac{2}{8}\log\frac{2}{8}-\frac{6}{8}\log\frac{6}{8}=0.81$. Thus the entropy for color is $0.5\times 0.8+0.5\times 0.95=0.88$. The information gain is $0.99-0.88=0.11$. Similarly, for size,act and age, the information gains are all 0.11. Pick color as the first level.
Under Yellow, using the same method we have the information gain for size is 0.55. Under Yellow-Large, the information gain for act is 0.41. Under Yellow-Large-Stretch, we will end this branch.
Under Purple, using the same method, we will have the information gain for act is 0.41. Under Purple-Stretch, the information gain for age is 1. Thus we split the leaves and this branch is ended.

1.1.2 Result

```
If color=yellow
    if size=small
        class=T
    else
        if act=Dip
            class=F
        else
            if age=adult
                class=T
            else
                class=F
If color=purple
    if act=dip
        class=F
    else
        if age=adult
            class=T
        else
            class=F
```

1.2

1.2.1 Calculation

The new information gain can be defined as InformationGain= MajorityError(UpperLevel) - MajorityError(LowerLevel).
Using the above definition, the MajorityError is $\min(\frac{7}{16}, \frac{9}{16}) = \frac{7}{16}$. The MajorityError for yellow is $\min(\frac{3}{8}, \frac{5}{8}) = \frac{3}{8}$. The MajorityError for purple is $\min(\frac{2}{8}, \frac{6}{8}) = \frac{2}{8}$. Thus the information gain for color is $\frac{2}{16}$. Similarly, for size,act and age, the information gains are all $\frac{2}{16}$. Pick color as the first level.
Under Yellow, using the same method we have the information gain for size is $\frac{1}{4}$. Under Yellow-Large, the information gain for act is 0. Under Yellow-Large-Stretch, we will end this branch.
Under Purple, using the same method, we will have the information gain for act is 0. Under Purple-Stretch, the information gain for age is $\frac{1}{4}$. Thus we split the leaves and this branch is ended.

1.2.2 Result

```
If color=yellow
    if size=small
        class=T
    else
        if act=Dip
            class=F
        else
            if age=adult
                class=T
            else
                class=F
If color=purple
    if act=dip
        class=F
    else
        if age=adult
            class=T
        else
            class=F
```

Note that this tree is the same as the tree in (a).

1.3

1.3.1 Using (a)

Using the same method as in (a), we have the total entropy is 1. The information gain for color and size are 0.1. The information gain for act and age are 0.08. Then choose color as

the first level.
Under Yellow, the information gain for size is 0.55 which is the largest. Under Yellow-Large, majority is F and under yellow-small, everything is T. Thus end this branch to here.
Under purple, the information gain for act is 0.41 which is the same as that of age and is the largest. Under Purple-Dip, everything is F and under purple-stretch, the majority is T. Thus end this branch to here.

```
If color=yellow
    if size=small
        class=T
    else
        class=F
If color=purple
    if act=Dip
        class=F
    else
        class=T
```

Of the four test cases, 1 is predicted wrong. Thus the error rate is 0.25

1.3.2 Using (b)

Using the same method as in (b), we have the total entropy is $\frac{1}{2}$. The information gain for color and size are $\frac{1}{2}$. The information gain for act and age are still $\frac{1}{2}$. Then choose color as the first level.
Under Yellow, the information gain for size is $\frac{1}{4}$ which is the largest. Under Yellow-Large, majority is F and under yellow-small, everything is T. Thus end this branch to here.
Under purple, the information gain for act is 0 which is the same as that of age and size and is the largest. Under Purple-Dip, everything is F and under purple-stretch, the majority is T. Thus end this branch to here.

```
If color=yellow
    if size=small
        class=T
    else
        class=F
If color=purple
    if act=Dip
        class=F
    else
        class=T
```

Of the four test cases, 1 is predicted wrong. Thus the error rate is 0.25.
Note that this prediction is also the same as using the prediction by (a)

2 Problem 2

The equation to calculate the 0.99 confidence interval is (According to the table $t_\alpha = 4.604$)

$$\bar{X} \pm t_\alpha \times \frac{S}{\sqrt{N}} \tag{1}$$

2.1 SGD

2.1.1 R=0.00002

Experiment Number	1	2	3	4	5	Ave
Accuracy	0.7125	0.67	0.695	0.705	0.685	0.6935

Its' 0.99 confidence interval is 0.659 - 0.728

2.1.2 R=0.002

Experiment Number	1	2	3	4	5	Ave
Accuracy	0.595	0.685	0.57	0.61	0.625	0.617

Its' 0.99 confidence interval is 0.528 - 0.705

From the above values we could see that SGD is very sensitive to the learning rate. Especially in the case that only one example is used at one time to update the weights. In this case, the learning process is not very stable, it is sensitive to noises and the resulting weights are not very good compared with the cases where a batch of examples are used at one step to update the weights.

2.2 Id3

2.2.1 Depth=Maximum

Experiment Number	1	2	3	4	5	Ave
Accuracy	0.9	0.885	0.8575	0.8775	0.8875	0.8815

Its' 0.99 confidence interval is 0.849 - 0.912

2.2.2 Depth=4

Experiment Number	1	2	3	4	5	Ave
Accuracy	0.7775	0.79	0.76	0.775	0.775	0.7755

Its' 0.99 confidence interval is 0.753 - 0.798

2.2.3 Depth=6

Experiment Number	1	2	3	4	5	Ave
Accuracy	0.82	0.835	0.815	0.795	0.81	0.815

Its' 0.99 confidence interval is 0.785 - 0.845

2.2.4 Depth=8

Experiment Number	1	2	3	4	5	Ave
Accuracy	0.87	0.81	0.81	0.8375	0.8325	0.832

Its' 0.99 confidence interval is 0.781 - 0.883
From the above values we could see that in principal, bigger depth will result in better trees. But since their confidence interval overlaps a lot, the differences are not very significant.

2.3 Stumps as features

Experiment Number	1	2	3	4	5	Ave
Accuracy	0.6925	0.65	0.6675	0.685	0.695	0.678

Its' 0.99 confidence interval is 0.639 - 0.717
I also tried using very few stumps (e.g 5 and 10) and sometimes the classifier can not be built because the stump feature space is not consistent. This makes sense since if we use very few stumps as feature space and if the depth of the tree is not big enough, we may result in two identical feature set with different labels.
From the result we could also see that the result from (e) seems to be the worse of (b) and (d). This makes sense since with more combinations, we are adding more fluctuations , variables and noises to the system and the resulting separator will be worse than either two component in the combination.

2.4 Comparison of the algorithms

From the average accuracy we could see that the accuracy with descending order is : Id3 with max depth, Id3 with depth=8, Id3 with depth=4, SGD baseline, SGD with stumps as features.
From the confidence interval we could see that Id3 with max depth only overlaps with Id3 with depth=8, meaning that these two don't have significant difference but Id3 with max depth has significant difference compared with others. Also, Id3 with depth=8 overlaps with Id3 with depth=4, meaning these two don't have significant difference, but is significantly different from SGD and SGD with stumps. Id3 with depth=4 also has significant difference compared with SGD and SGD with stumps. The SGD and SGD with stumps' confidence interval overlaps a lot, meaning that these two algorithms used don't have significant difference.
As a summary, I think the Id3 algorithm is the best compared with others. This is based on that appropriate depth is used and the feature space is of the size similar to the problem set and the function space is not complicated. Id3 method is stable compared with SGD baseline, which is sensitive to learning rate and noises and other things. Compared with the SGD with stumps as features, Id3 use much less time to compute and will use less memory, while SGD with stumps use a lot of memory and will be affected by the behavior of SGD, which sometimes is not very stalbe.
However, if our feature space and function space changes, Id3 may not be the best, because if feather space is extremely large, we may need to generate a huge tree, which take a lot

of time and memory. Also if the function space is complicated, the Id3 can not handle this case, and we may need more sophisticated tree algorithms to tackle the problem.