CS446: Machine Learning **Spring 2017**

Problem Set 2

Zhuo Li(zhuol2) *Handed In: February 16, 2017*

1. Answer to problem 1

   (a) We can first calculate the overall entropy:

$$\begin{aligned} Entropy(S) &= Entropy([35+, 15-]) \\ &= -\frac{35}{50}\log_2(35/50) - \frac{15}{50}\log_2(15/50) \\ &= 0.881 \end{aligned}$$

   And we have,

$$\begin{aligned} S_{Holiday\_yes} &\leftarrow [20+, 1-] \\ S_{Holiday\_no} &\leftarrow [15+, 14-] \\ Entropy(S_{Holiday\_yes}) &= -(20/21)\log_2(20/21) - (1/21)\log_2(1/21) = 0.276 \\ Entropy(S_{Holiday\_no}) &= -(15/29)\log_2(15/29) - (14/29)\log_2(14/29) = 0.999 \\ Gain(S, Holiday) &= Entropy(S) - \frac{21}{50}Entropy(S_{Holiday\_yes}) - \frac{29}{50}Entropy(S_{Holiday\_no}) \\ &= 0.881 - \frac{21}{50} \times 0.276 - \frac{29}{50} \times 0.999 \\ &= 0.186 \end{aligned}$$

   Similarly,

$$\begin{aligned} S_{Exam\_yes} &\leftarrow [10+, 5-] \\ S_{Exam\_no} &\leftarrow [25+, 10-] \\ Entropy(S_{Exam\_yes}) &= -(10/15)\log_2(10/15) - (5/15)\log_2(5/15) = 0.918 \\ Entropy(S_{Exam\_no}) &= -(25/35)\log_2(25/35) - (10/35)\log_2(10/35) = 0.863 \\ Gain(S, Exam) &= Entropy(S) - \frac{15}{50}Entropy(S_{Holiday\_yes}) - \frac{35}{50}Entropy(S_{Holiday\_no}) \\ &= 0.881 - \frac{15}{50} \times 0.918 - \frac{35}{50} \times 0.863 \\ &= 0.0015 \end{aligned}$$

   So $Gain(S, Holiday) > Gain(S, Exam)$, we need to choose **Holiday** as the root attribute.

   (b) By the definition of *MajorityError*, information gain can be defined as:

$$Gain(S, A) = MajortiyError(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} MajorityError(S_v)$$

Where $Values(A)$ is the set of all possible values of attribute A, S_v is the subset of S for which attribute A has value of v.
For the root of the tree, all the 4 attributes have the same information gain:

$$\begin{aligned} Gain(S, Color) &= Gain(S, Size) = Gain(S, Act) = Gain(S, Age) \\ &= ME(S) - \frac{1}{2}ME([3+, 5-]) - \frac{1}{2}ME([6+, 2-]) \\ &= \frac{7}{16} - \frac{1}{2} \times \frac{3}{8} - \frac{1}{2} \times \frac{2}{8} \\ &= 0.125 \end{aligned}$$

So we could choose either attribute as the root. Here we just choose *Color*.
The following procedure is similar, we just show the final decision tree below:

```
if Color = Blue :
  if Size = Small :
      class = F
  if Size = Large :
    if Act = Dip :
      class = T
    if Act = Stretch :
      if Age = Adult :
        class = F
      if Age = Child :
        class = T
if Color = Red :
  if Act = Dip :
    class = T
  if Act = Stretch :
    if Age = Adult :
      class = F
    if Age = Child :
      class = T
```

(c) ID3 algorithm can not guarantee a globally optimal decision tree. Since finding a minimal decision tree consistent with a set of data is NP-hard and the ID3 algorithm is a greedy algorithm which only makes locally optimal decision without backtracking. In this way, the result may not be a optimal one.

2. Answer to problem 2

(a) The generated 260 features files in Weka ARFF format are:

- *badges.fold1.arff*
- *badges.fold2.arff*
- *badges.fold3.arff*

- *badges.fold4.arff*
- *badges.fold5.arff*

Combinations of 4 datasets are also concatenated as below for CV later:

- *badges.fold2345.arff*
- *badges.fold1345.arff*
- *badges.fold1245.arff*
- *badges.fold1235.arff*
- *badges.fold1234.arff*

(b) Using the 260 features generated in (a), we basically get the results below:

Algorithms	Acc. (p_A)	Std. Dev.	99% Conf. Inter.	Params
SGD	68.83	5.93	[56.62, 81.04]	LearningRate = 0.0355 Threshold = 0.000001
Unlimited DT	72.23%	2.77	[66.53, 77.93]	
DT (depth 4)	65.64%	5.39	[54.54, 76.74]	Depth=4
DT (depth 8)	69.67%	3.76	[61.93, 77.41]	Depth=8
DT + SGD	69.10%	5.73	[57.30, 80.90]	LearningRate = 0.01 Threshold = 0.00001

The decision trees with best performance in CV are presented/displayed in the following files:

- *tree_unlimited.txt*
- *tree_depth_4.txt*
- *tree_depth_8.txt*

For SGD we tune the parameters learningRate λ from 0.0001 to 0.1 with a step of 0.0003, and $threshold = 0.000001, 0.00001, 0.0001$. The best combination is shown on the above table.
Algorithm performs ranks: Unlimited DT ¿ DT(depth 8) ¿ DT + SGD ¿ SGD ¿ DT(depth 4). There are no statistical significant differences between the two algorithms.
In the table shown above, the **Unlimited DT** seems to perform the best, however theoretically, the **SGD over Decision Stump** algorithm should perform the best, because it takes advantage of Decision tree expressiveness to generate datasets with better features. When we say decision tree expressiveness, we usually means it can represent not only linear functions, while SGD we use can only work on this. The reason why we do not get the best performance may be the lack of tuning on SGD parameters.
One thing need to be pointed out is the last algorithm may fluctuate on results. It depends on the randomly sampled 50% training instances for building the stump. The fluctuation may converge with more iterations.
We can also find that, in the results of DT algorithms (i.e. 2 to 4), the DT with no limits of depth performs the best. This may be the case that the maximum

depth generated is still not complex enough to cause a overfit on the dataset. In this case, the more complex the tree is, the higher the accuraccy is.
Program running results are as shown below.

Figure 1: SGD

```
----- Summary -----
Optimal Learning Rate: 0.0355
Optimal Threshold: 1.0E-6
Average Accuracy: 68.83%
Standard Deviation: 5.93
```

Figure 2: Unlimited DT

```
========== ID3 with Unlimited Depth ==========
----- Fold 1/5 -----
Fold Accuracy: 68.33333333333333
----- Fold 2/5 -----
Fold Accuracy: 74.24242424242425
----- Fold 3/5 -----
Fold Accuracy: 76.08695652173913
----- Fold 4/5 -----
Fold Accuracy: 70.17543859649123
----- Fold 5/5 -----
Fold Accuracy: 72.3076923076923
----- Summary -----
Average Accuracy: 72.23%
Standard Deviation: 2.77
```

Figure 3: DT with depth 4

```
========== ID3 with Depth of 4 ==========
----- Fold 1/5 -----
Fold Accuracy: 68.33333333333333
----- Fold 2/5 -----
Fold Accuracy: 72.72727272727273
----- Fold 3/5 -----
Fold Accuracy: 58.69565217391305
----- Fold 4/5 -----
Fold Accuracy: 68.42105263157895
----- Fold 5/5 -----
Fold Accuracy: 60.0
----- Summary -----
Average Accuracy: 65.64%
Standard Deviation: 5.39
```

Figure 4: DT with depth 8

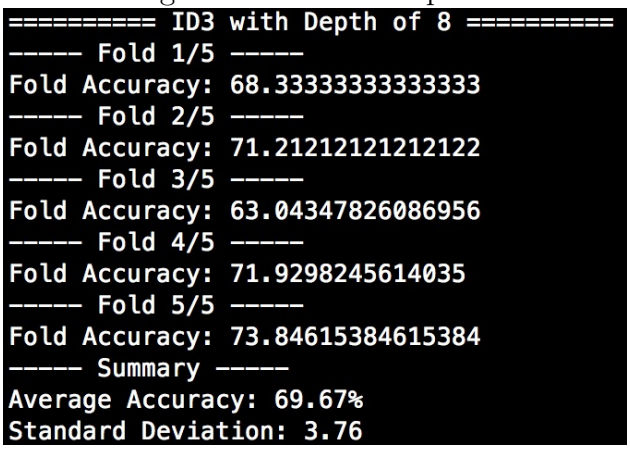

Figure 5: DT + SGD