CIS 419/519: Homework 1

Ransford Antwi

Although the solutions are my own, I consulted with the following people while working on this homework: {Names here}

1. (a) Show your work:

$$P(\text{play outside} = \text{yes}) = \frac{60}{100} = 0.6$$

$$P(\text{play outside} = \text{no}) = \frac{40}{100} = 0.4$$

Entropy(S) = -p$_+ log_2 p_+ - p_- log_2 p_- = -0.6 log_2(0.6) - 0.4 log(0.4) = 0.971$

IG$_{Snow} = Entropy(S) - \sum \frac{|S_v|}{S}.Entropy(S_v)$
$For snow, p_+ = \frac{14}{18} = \frac{7}{9}$
$p_- = \frac{4}{18} = \frac{2}{9}$
Entropy(S$_v$) = 0.2819988 + 0.48220555 = 0.764
$IG = 0.971 - \frac{50}{100}.0.764$
= 0.589
$IG_{Sunny} = Entropy(S) - \sum \frac{|S_v|}{S}.Entropy(S_v)$
$For sunny, p_+ = \frac{25}{31}$
$p_- = \frac{6}{31} = \frac{2}{9}$
$Entropy(S_v) = 0.2502766 + 0.45856 = 0.7088$
$IG = 0.971 - \frac{50}{100}.0.7088$
= 0.617

Sunny is better because is has a higher Information Gain

(b) MinError(S) = min($\frac{11}{17}$, $\frac{6}{17}$) =
$\frac{6}{17} = 0.3529$
Information Gain: Color
Color = blue:
p$_+ = \frac{3}{12}, p_- = \frac{9}{12}.Min(Color = blue) = 0.25$
$Color = Red$:
$p_+ = \frac{3}{5}, p_- = \frac{2}{5}.Min(Color = red) = 0.4$
$Expectedentropy = (\frac{12}{17} * 0.25) + (\frac{5}{17} * 0.4) = 0.294$
IG = 0.3529 - 0.294 = 0.0589

Information Gain: Size
Size = Small: $p_+ = \frac{1}{9}, p_- = \frac{8}{9}. Min(Size = small) = 0.1111$
$Size = Large : p_+ = \frac{5}{8}, p_- = \frac{3}{8}. Min(Size = Large) = 0.375$
$ExpectedEntropy = (\frac{9}{17} * 0.1111) + (\frac{8}{17} * 0.375) = 0.23529$
IG = 0.3529 - 0.23529 = 0.1176

Information Gain: Act
Act = Stretch: $p_+ = \frac{5}{9}, p_- = \frac{4}{9}. Min(Act = Stretch) = 0.44444$
$Act = Dip : p_+ = \frac{1}{8}, p_- = \frac{7}{8}. Min(Act = Dip) = 0.125$
$ExpectedEntropy = (\frac{9}{17} * 0.44444) + (\frac{8}{17} * 0.125) = 0.294$
IG = 0.3529 - 0.23529 = 0.1176

Information Gain: Age
Age = Adult: $p_+ = \frac{1}{6}, p_- = \frac{5}{6}. Min(Age = Adult) = 0.16666 Age = Child : p_+ = \frac{5}{11}, p_- = \frac{6}{11}. Min(Age = Child) = 0.454545 ExpectedEntropy = (\frac{6}{17} * 0.16666) + (\frac{11}{17} * 0.454545) = 0.3529$ **IG = 0.3529 - 0.3529 = 0**

For **Root,** split on **Size** because it has the highest information gain.
IG for (Size = Small) = $\min(\frac{1}{9}, \frac{8}{9}) = 0.1111$

Information Gain: Size = Small ,Color
Color = Blue: $p_+ = \frac{1}{7}, p_- = \frac{6}{7}. Min(Color = Blue) = 0.142857$
$Color = Red : p_+ = \frac{0}{2}. Min(Color = Red) = 0; ExpectedEntropy = \frac{7}{9} * 0.142857 = 0.1111 IG = 0.1111 - 0.1111 = 0$

Information Gain: Size = Small, Act
Act = Stretch: $p_+ = \frac{0}{4}, p_- = \frac{4}{4}. Min(Act = Stretch) = 0$
$Act = Dip : p_+ = \frac{1}{5}, p_- = \frac{4}{5}. Min(Act = Dip) = 0.2 ExpectedEntropy = \frac{5}{9} * 0.2 = 0.11111 IG = 0.1111 - 0.1111 = 0$

Information Gain: Size = Small, Age
Age = Adult: $p_+ = \frac{1}{5}, p_- = \frac{4}{5}. Min(Age = Adult) = 0.2$
$Age = Child : p_+ = \frac{0}{4}, p_- = \frac{4}{4}. Min(Age = Child) = 0 ExpectedEntropy = \frac{5}{9} * 0.2 = 0.11111 IG = 0.1111 - 0.1111 = 0$

For Size = **small**, Split on **Color** because all IGs are equal and color comes first in the table.
IG for (Size = Large) = $\min(\frac{5}{8}, \frac{3}{8}) = 0.375$

Information Gain: Size = Large ,Color
Color = Blue: $p_+ = \frac{2}{5}, p_- = \frac{3}{5}. Min(Color = Blue) = 0.4$
$Color = Red : p_+ = \frac{3}{3}. Min(Color = Red) = 0; ExpectedEntropy = \frac{5}{8} * 0.4 = 0.25 IG = 0.375 - 0.25 = 0.125$

Information Gain: Size = Large, Act
Act = Stretch: $p_+ = \frac{5}{5}, p_- = \frac{0}{5}. Min(Act = Stretch) = 0$
$Act = Dip : p_+ = \frac{0}{3}, p_- = \frac{3}{3}. Min(Act = Dip) = 0 ExpectedEntropy = 0 IG = 0.375 - 0 = 0.375$

Information Gain: Size = Large, Age
Age = Adult: $p_+ = \frac{0}{1}, Min(Age = Adult) = 0$

$Age = Child : p_+ = \frac{5}{7}, p_- = \frac{2}{7}. Min(Age = Child) = 0.286 Expected Entropy = \frac{7}{8} * 0.286 = 0.25 IG = 0.375 - 0.25 = 1.25$
$For Size =$ **large**, $Split on$ **Color** $because it has the highest IG and comes before Age$
IG for (Size = Small, Color = Blue) = min($\frac{1}{7}$, $\frac{6}{7}$) = 0.142
Information Gain: (Size = Small,Color = Blue), Age
Age = Adult: $p_+ = \frac{1}{5}, p_- = \frac{4}{5}. Min(Age = Adult) = 0.2$
$Age = Child : p_+ = \frac{0}{2}. Min(Age = Child) = 0; Expected Entropy = \frac{5}{7} * 0.2 = 0.142 IG = 0.142 - 0.142 = 0$

Information Gain: (Size = Small,Color = Blue), Act
Act = Stretch: $p_+ = \frac{0}{4}, p_- = \frac{4}{4}. Min(Act = Stretch) = 0$
$Act = Dip : p_+ = \frac{1}{3}. p_- = \frac{2}{3}. Min(Act = Dip) = 0.3333; Expected Entropy = \frac{3}{7} * 0.3333 = 0.142 IG = 0.142 - 0.142 = 0$

$For Size =$ **Small** $and Color =$ **Blue**, $Split on$ **Act** $because it comes before Age$
IG for (Size = Large, Color = Blue) = min($\frac{2}{5}$, $\frac{3}{5}$) = 0.4
Information Gain: (Size = Large,Color = Blue), Age
Age = Adult: $p_+ = \frac{0}{1}, Min(Age = Adult) = 0$
$Age = Child : p_+ = \frac{2}{4}. p_- = \frac{2}{4}. Min(Age = Child) = 0.5; Expected Entropy = \frac{4}{5} * 0.5 = 0.4 IG = 0.4 - 0.4 = 0$

Information Gain: (Size = Large,Color = Blue), Act
Act = Stretch: $p_+ = \frac{2}{2}, p_- = \frac{0}{2}. Min(Act = Stretch) = 0$
$Act = Dip : p_+ = \frac{0}{3}. p_- = \frac{3}{3}. Min(Act = Dip) = 0; Expected Entropy = 0 IG = 0.4$
For Size = **Large** and Color = **Blue**, Split on **Act** because it has the higher information gain.

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

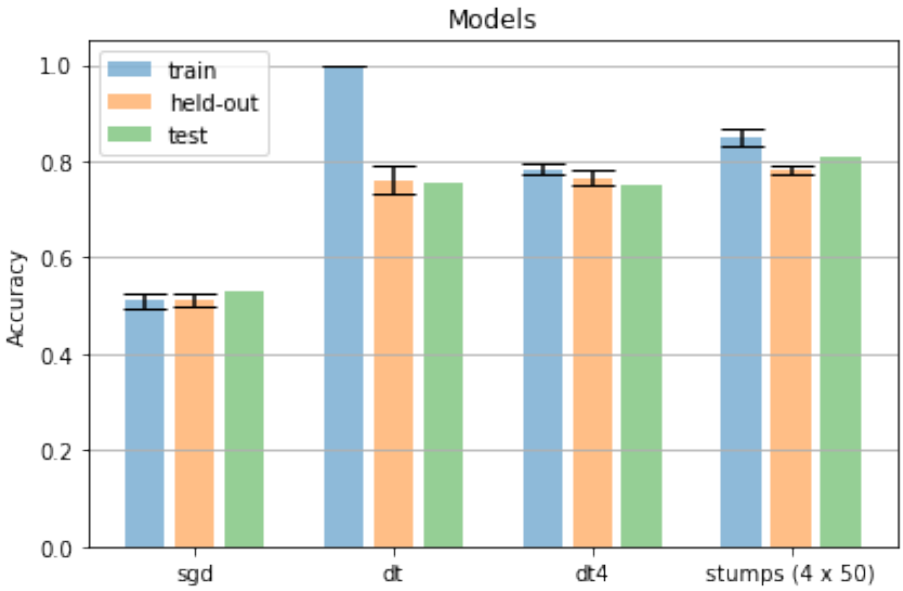

Figure 1: Model performance on the Madelon dataset

```
if Color = Red:
  Inflated = T
```

(c) ID3 cannot guarantee a globally optimal tree. Finding a minimal decision tree consistent with a set of data is an NP-hard problem. One attribute of ID3 is that is uses a greedy strategy by selecting the locally best attribute to split the dataset on each iteration and the algorithm is harder to use on continuous data. This is because if the values of any given attribute are continuous, then there are many more places to split the data on the attribute, which can be time consuming, ID3 like most decision tree algorithms is also prone to overfitting.

2. (a) See Figure 1 for the model performance in the Madelon dataset.

    1. Ranking by the held out and test performances: the best was the decision stumps as features model, followed closely by the decision stumps model, followed by the decision tree model and the worst was the SGD model.
    My testing performance estimate $\text{acc}_{heldout}$ was similar to acc_{test} for the SGD, decision tree and decision stumps since for all these models, $\text{acc}_{heldout}$ was within a standard deviation of acc_{test}. For the decision stumps as feature model, acc_{test} was greater than a standard deviation away from $\text{acc}_{heldout}$, however, I will stay say that it was similar, since $\text{acc}_{heldout}$ was a good prediction of how

well the model will do on test data as the relative performance of all the models on the test data was similar to how the models performed on the heldout data, relative to one another. Interestingly, the accuracy of the real testing performance(acc_{test}) was higher than $\text{acc}_{heldout}$ for the decision stumps as features model and the SGD model. For the decision tree and decision stumps model, $\text{acc}_{heldout}$ was higher.

2. The decision tree model had the highest training accuracy. This was because the model is based on constructing a tree from the training data and since when evaluating training accuracy we predict on the training data, the decision tree will return the exact same labels that were used to train it and hence it has 100% training accuracy. A decision tree is very prone to overfitting.
3. The confidence intervals were as follows:
   SGD training: +- 0.036 SGD heldout: +- 0.016
   DT training: +- 0 DT heldout: +- 0.03
   DT4 training: 0.0145 DT4 heldout: +- 0.02
   stumps training: +- 0.0187 Stumps heldout: +- 0.0347
   The results for the SGD classifier and DT4 classifier I would say are statisitically significant because the boundaries of the confidence intervalss overlap with each other. One wway to have a tighter confidence interval will be to increase the sample size, i.e. increase the size of the datasets and increase the number of folds we used in the CV evaluation.
4. Repeatedly running the classifier on testing data will only give us an accuracy for that particular arrangement and is prone to overfitting, whilst with k fold cross validation, we loop through different partitions of the data and then average the results to get an overall accuracy. This reduces bias because in k fold CV, most of the data is used for training and it reduces variance because most of the data is also used in the validation set.

(b) The models' Training accuracies on the Badges dataset were as follows:

Algorithm	Accuracy
Decision Tree	1.0
SGD	0.74
SGD + Decision Stump Features	0.74
Decision Stump	0.677

The models' Testing accuracies on the Badges dataset were as follows:

Algorithm	Accuracy
SGD + Decision Stump Features	0.661
Decision Stump	0.66
SGD	0.591
Decision Tree	0.587

The order of the training and testing accuracies of the models are not the same. The decision tree performed the best on the training sets in both cases but the decision stumps as features SGD model performed best on the Madelon test dataset as well as the Badges test dataset.