CIS 519: Homework 1

Ian MacDonald

10/04/2020

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

1. (a) Show your work:
Entropy(S) = $\frac{-35}{50} log_2 \frac{35}{50} - \frac{15}{50} log_2 \frac{15}{50}$
= .7(.515) + .3(1.737)
= .3605 + .5211
= .8816

Entropy gain from splitting on Sunny:
$\frac{34}{50}(\frac{-28}{34} log_2 \frac{28}{34} - \frac{6}{34} log_2 \frac{6}{34})$
$\frac{34}{50}(\frac{28}{34}(.28) + \frac{6}{34}(2.503))$
$\frac{28}{34}(.231 + .442)$
= .45764

$\frac{16}{50}(\frac{-7}{16} log_2 \frac{7}{16} - \frac{9}{16} log_2 \frac{9}{16})$
$\frac{16}{50}(\frac{7}{16}(1.193) + \frac{9}{16}(.83))$
$\frac{16}{50}(.522 + .467)$
= .31648

.45764 + .31648 = .77412
IG(Sunny) = .8816-.77412
= .10748

Entropy gain from splitting on Snow:
$\frac{18}{50}(\frac{-16}{18} log_2 \frac{16}{18} - \frac{2}{18} log_2 \frac{2}{18})$
$\frac{18}{50}(\frac{16}{18}(.17) + \frac{2}{18}(3.17))$
$\frac{18}{50}(.151 + .352)$
= .18108

$\frac{32}{50}(\frac{-19}{32} log_2 \frac{19}{32} - \frac{13}{32} log_2 \frac{13}{32})$
$\frac{32}{50}(\frac{19}{32}(.752) + \frac{13}{32}(1.3))$
$\frac{32}{50}(.4465 + .528)$
= .62368

.18108 + .62368 = .80476
IG(Sunny) = .8816-.80476
= .07684

$$P(\text{play outside = yes}) = \frac{35}{50}$$

$$P(\text{play outside = no}) = \frac{15}{50}$$

...

$$IG_{Snow} = .07684$$

$$IG_{Sunny} = .10748$$

Splitting on Sunny is better because the information gain is higher, .10748 compared to .07684

(b) MinError(S) = min(p_-, p_+)
= min($\frac{13}{24}$, $\frac{11}{24}$)
= $\frac{11}{24}$ = .4583

$Gain_{ME}$(S, Color) = $\frac{11}{24}$ - ($\frac{15}{24}$(min($\frac{10}{15}$, $\frac{5}{15}$) + $\frac{9}{24}$(min($\frac{6}{9}$, $\frac{3}{9}$))
= $\frac{11}{24}$ - ($\frac{15}{24}$*$\frac{5}{15}$ + $\frac{9}{24}$*$\frac{3}{9}$
= .125

$Gain_{ME}$(S, Size) = $\frac{11}{24}$ - ($\frac{18}{24}$(min($\frac{11}{18}$, $\frac{7}{18}$) + $\frac{6}{24}$(min($\frac{6}{24}$, $\frac{2}{6}$))
= $\frac{11}{24}$ - ($\frac{12}{24}$*$\frac{5}{12}$ + $\frac{12}{24}$*$\frac{6}{12}$
= .08

$Gain_{ME}$(S, Act) = $\frac{11}{24}$ - ($\frac{12}{24}$(min($\frac{5}{12}$, $\frac{7}{12}$) + $\frac{12}{24}$(min($\frac{6}{12}$, $\frac{6}{12}$))
= $\frac{11}{24}$ - ($\frac{12}{24}$*$\frac{5}{12}$ + $\frac{12}{24}$*$\frac{6}{12}$
= 0

$Gain_{ME}$(S, Age) = $\frac{11}{24}$ - ($\frac{12}{24}$(min($\frac{6}{12}$, $\frac{6}{12}$) + $\frac{12}{24}$(min($\frac{5}{912}$, $\frac{7}{12}$))
= $\frac{11}{24}$ - ($\frac{12}{24}$*$\frac{6}{12}$ + $\frac{12}{24}$*$\frac{6}{12}$
= 0
Since Information Gain is the highest, split on Color
S = (Color—Red)
MinError(S) = min($\frac{10}{15}$, $\frac{5}{15}$)
= .33

$Gain_{ME}$(S, Size) = $\frac{5}{15}$ - ($\frac{13}{15}$(min($\frac{8}{13}$, $\frac{5}{13}$) + $\frac{2}{15}$(min($\frac{2}{2}$, $\frac{0}{2}$))
= $\frac{5}{15}$ - ($\frac{13}{15}$*$\frac{5}{13}$ + $\frac{2}{15}$*$\frac{0}{2}$
= 0

$Gain_{ME}$(S, Act) = $\frac{5}{15}$ - ($\frac{11}{15}$(min($\frac{6}{11}$, $\frac{5}{11}$) + $\frac{4}{15}$(min($\frac{4}{15}$, $\frac{0}{15}$))
= $\frac{5}{15}$ - ($\frac{11}{15}$*$\frac{5}{11}$ + $\frac{4}{15}$*$\frac{0}{4}$
= 0

$Gain_{ME}$(S, Age) = $\frac{5}{15}$ - ($\frac{6}{15}$(min($\frac{6}{6}$, $\frac{0}{6}$) + $\frac{9}{15}$(min($\frac{5}{9}$, $\frac{4}{9}$))
= $\frac{5}{15}$ - ($\frac{6}{15}$*$\frac{0}{6}$ + $\frac{9}{15}$*$\frac{4}{9}$
= .067
Split on Age

S = (Color—Blue)
MinError(S) = min($\frac{6}{9}$, $\frac{3}{9}$)
= .33

$Gain_{ME}$(S, Age) = $\frac{3}{9}$ - ($\frac{6}{9}$(min($\frac{6}{6}$, $\frac{0}{6}$) + $\frac{3}{9}$(min($\frac{3}{3}$, $\frac{0}{3}$))
= $\frac{3}{9}$ - ($\frac{6}{9}$*$\frac{0}{6}$ + $\frac{3}{9}$*$\frac{0}{3}$
= .33
Split on Age, no need to calculate farther since IG = MinError(S)

S = (Color—Red, Age—Adult)
All false, no need to split

S = (Color—Red, Age—Child)
MinError(S) = min($\frac{4}{9}$, $\frac{5}{9}$)

$Gain_{ME}$(S, Act) = $\frac{4}{9}$ - ($\frac{5}{9}$(min($\frac{5}{5}$, $\frac{0}{5}$) + $\frac{4}{9}$(min($\frac{5}{9}$, $\frac{0}{5}$))
= $\frac{4}{9}$ - ($\frac{5}{9}$*$\frac{0}{5}$ + $\frac{4}{9}$*$\frac{0}{4}$
= .444

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

(c) ID3 does not guarantee a globally optimized tree. While it will create a decision tree that perfectly fits the training data, since it will continuously add nodes until all of the data is accounted for, it will not always have the minimum depth, since it is a greedy algorithm that creates splits based on only the locally optimal choice, which does not take into account the rest of the tree as a whole.

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

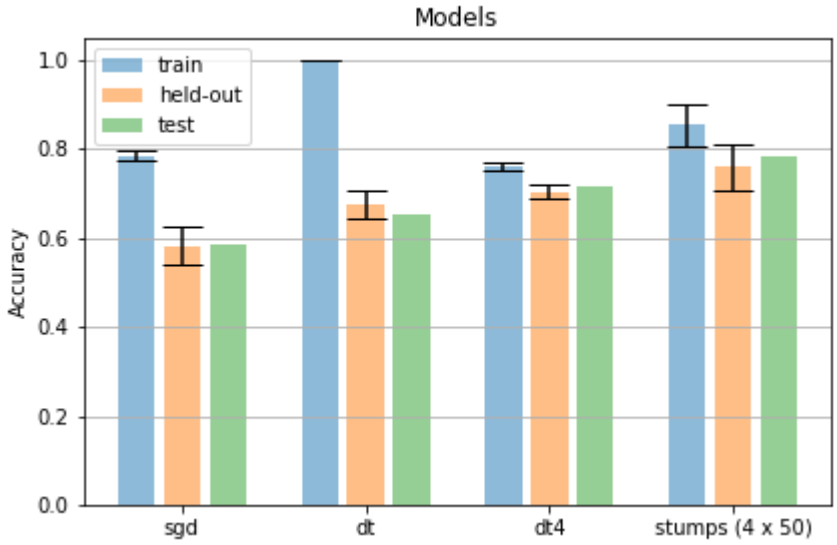

Figure 1: Model performance on the Madelon dataset

1. Classifiers ranked from best to worst in terms of held-out performance - Decision stumps as features, decision stumps, decision tree, sgd. Classifiers had the same ranking for the testing performances as well. In both cases, the classifier that used decision stumps as features performed the best, and the classifier that just used sgd performed the worst. In all cases, the actual testing performance fell within the confidence intervals calculated for the estimated testing performance.
2. The classifier with the highest training accuracy was the decision tree, with 100 percent training accuracy. Since the decision tree has no max depth, it will continuously create nodes until all of the training data is accounted for, resulting in perfect training accuracy for this classifier.
3. 95 Percent Confidence Intervals for accuracy of held out data:
   sgd - [.545, .620]
   dt - [.649, .702]
   ds - [.693, .717]
   combined - [.714, .805]
   The sgd confidence interval does not overlap with any other interval, so we can say that it is significantly worse than the other models. However, the decision tree and decision stump models both overlap, so we cannot say that one is significantly bet-

ter than the other. Additionally, the same applies to the decision stump and combined models. Therefore, our result that the combined model is the best model is not statistically significant (when compared to the stumps model, but it is significant compared to the other two models). To tighten the confidence intervals, we could have increased the number of cross validation folds, which would have increased our observations and decreased the variance, leading to a tighter confidence interval.

4. Cross validation is an important technique to use because it allows you to estimate how well your classifier will do on an unknown data set. By using cross-validation, you prevent your classifier from being too specific to or biased by the testing set, which will help it perform better when predicting unknown data sets.

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

Algorithm	Accuracy
SGD	.575
Decision Tree	.455
Decision Stump	.5
SGD + Decision Stump Features	.57

1. For the testing accuracy, from best to worst, the models were - SGD, Decision stumps as features, decision stump, decision tree. For the training accuracy, from best to worst, the models were - decision tree, sgd, decision stumps as features, decision stumps. The order for these was not the same, and were actually pretty different. It was expected that the decision tree would have the highest training accuracy at 100 percent, but sgd had the second highest training accuracy and the best testing accuracy, so the sgd wasn't too far off in terms of training vs testing rank. Additionally, the model that performed the best on the madelon test set (decision stumps as features) was not the same as the one that performed the best on the badges test set (sgd). However, the decision stumps as features model was very close behind the sgd model for the badges test set, with an accuracy of .57 compared to .575.

3. **Extra Credit**
For the extra credit, I added one new feature column to the badges data set. This feature was 1 if the total letters in the name (first and last together as one name) was an odd number, and 0 if the total letters in the name was an even number. Adding this feature increased the accuracy of my decision tree model on the test and training sets to 100 percent. Because of this, I chose the decision tree model to predict the leaderboard and hidden data sets.