CS446: Machine Learning **Spring 2017**

Problem Set 2

Bangqi Wang *Handed In: February 17, 2017*

1. Answer to problem 1

   a. The root of the decision tree is the attribute that has largest information gain.

Attribute	Value	Study Today = yes	Study Today = no
Holiday	yes	20	1
Holiday	no	15	14
Exam Tomorrow	yes	10	5
Exam Tomorrow	no	25	10

Table 1: The `Study Pattern` data set

- $Gain(S, a) \equiv Entropy(S) \ - \sum_{v \in Values(a)} \frac{|S_v|}{|S|} Entropy(S_v)$
- $Entropy(S)$:
  - $Entropy(S) = -\frac{35}{50} log(\frac{35}{50}) - \frac{15}{50} log(\frac{15}{50}) \approx 0.611$
  - $Entropy(S_{Holidy=yes}) = -\frac{20}{21} log(\frac{20}{21}) - \frac{1}{21} log(\frac{1}{21}) \approx 0.191$
  - $Entropy(S_{Holidy=no}) = -\frac{15}{29} log(\frac{15}{29}) - \frac{14}{29} log(\frac{14}{29}) \approx 0.693$
  - $Entropy(S_{Exam=yes}) = -\frac{10}{15} log(\frac{10}{15}) - \frac{5}{15} log(\frac{5}{15}) \approx 0.637$
  - $Entropy(S_{Exam=no}) = -\frac{25}{35} log(\frac{25}{35}) - \frac{10}{35} log(\frac{10}{35}) \approx 0.598$
- $Gain(S, a)$:
  - $Gain(S, Holiday)$
    $= Entropy(S) - \frac{|S_{Holiday=yes}|}{|S|} Entropy(S_{Holiday=yes}) - \frac{|S_{Holiday=no}|}{|S|} Entropy(S_{Holiday=no})$
    $= 0.611 - \frac{21}{50} \times 0.191 - \frac{29}{50} \times 0.693 \approx 0.129$
  - $Gain(S, Exam)$
    $= Entropy(S) - \frac{|S_{Exam=yes}|}{|S|} Entropy(S_{Exam=yes}) - \frac{|S_{Exam=no}|}{|S|} Entropy(S_{Exam=no})$
    $= 0.611 - \frac{15}{50} \times 0.637 - \frac{35}{50} \times 0.598 \approx 0.0013$
- **Holiday** is the attribute that will be the root of the decision tree because it has largest information gain.

b.
```
if Color = Blue:
    if Size = Small:
        Inflated = F
    if Size = Large:
        if Act = Stretch:
            if Age = Adult:
                Inflated = F
```

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

c. Finding the optimal decision tree is NP-Complete. The ID3 algorithm are based on greedy heuristics that split the attribute based on locally optimal decisions, information gain. Therefore, ID3 cannot guanrantee a globally optimal decision tree because there is no backtracking after greedily selecting locally optimal decisions.

2. Answer to problem 2

a. FeatureGenerator.java: extract 10 features from firstname and lastname.

```
features = new String[] { "firstName0", "firstName1", "firstName2", "firstName3",
    "firstName4", "lastName0", "lastName1", "lastName2", "lastName3", "lastName4" };

// feats.add("firstName0=" + firstName.charAt(0));
// feats.add("firstNameN=" + firstName.charAt(firstName.length() - 1));
for(int i = 0; i < 5; i++) {
    if (firstName.length() > i) {
        feats.add("firstName" + Integer.toString(i) + "=" + firstName.charAt(i));
    }
    if (lastName.length() > i) {
        feats.add("lastName" + Integer.toString(i) + "=" + lastName.charAt(i));
    }
}
```

For question a, I generated ten feature types. the features *firstName0-4* stand for the first five characters in first name, and features *lastName0-4* stand for the first five characters in last name.

b. • The table below shows the accuracy of algorithms in descending order.

`Algorithm`	`fold1`	`fold2`	`fold3`	`fold4`	`fold5`	p_A	σ	99% `interval`
DT	72.31%	70.18%	76.09%	74.24%	68.33%	72.23%	0.028	[64.25%, 80.21%]
Stumps	76.92%	64.91%	69.57%	77.27%	68.33%	71.40%	0.049	[57.30%, 85.50%]
DT_8	73.85%	71.93%	63.04%	71.21%	68.33%	69.67%	0.038	[58.84%, 80.50%]
DT_4	60.00%	68.42%	58.70%	72.73%	68.33%	65.64%	0.054	[50.12%, 81.16%]
SGD	67.69%	71.93%	56.52%	75.76%	51.67%	64.65%	0.092	[38.31%, 91.11%]

Table 2: The `Accuracy` for five algorithms

- Calculating 99% confidence intervals:

$$I = [p_A - \frac{t \times \sigma}{\sqrt{N}},\ p_A + \frac{t \times \sigma}{\sqrt{N}}],\ t_{0.995} = 2.576 \tag{1}$$

$$I_{DT} = 72.23\% \pm \frac{2.576 \times 0.028}{\sqrt{5}} = [64.25\%, 80.21\%] \tag{2}$$

$$I_{Stumps} = 71.40\% \pm \frac{2.576 \times 0.049}{\sqrt{5}} = [57.30\%, 85.50\%] \tag{3}$$

$$I_{DT_8} = 79.67\% \pm \frac{2.576 \times 0.038}{\sqrt{5}} = [58.84\%, 80.50\%] \tag{4}$$

$$I_{DT_4} = 65.64\% \pm \frac{2.576 \times 0.054}{\sqrt{5}} = [50.12\%, 81.16\%] \tag{5}$$

$$I_{SGD} = 64.65\% \pm \frac{2.576 \times 0.092}{\sqrt{5}} = [38.31\%, 91.11\%] \tag{6}$$

- To show if the difference between the two consecutive algorithm's performance is or is not statistically significant:
  - $DT \rightarrow Stumps$: No
  - $Stumps \rightarrow DT_8$: No
  - $DT_8 \rightarrow DT_4$: No
  - $DT_4 \rightarrow SGD$: No

- **Implement**:
  - **Decision Tree**:
    The implementation of decision tree is in folder $/bwang34{-}hw2/data/decision{-}trees/src/$. The tree structure is in $Id3.java$ and the decision tree classifier is in $WekaTester.java$.

  - **SGD**:
    The implementation of SGD is in the same folder. The SGD structure is in

Gradient.java. This SGD structure is similar to Id3 structure in $Classifier$ class. The classifier is also in $WekaTester.java$ and has similar usage as $Id3$.

- **Decision Stumps**:
  The implementation of Decision Stumps is also in the same folder. The strucutre is in $WekaTester.java$. I first construct 100 decision tree with depth 4, and then use the labels as new features. Finally, I run SGD classifier on those new features with five-fold cross validation.

- **Parameter Tuning**:

(rate,threshold)	fold1	fold2	fold3	fold4	fold5	p_A
$(0.001, 20)$	61.54%	68.42%	65.22%	77.27%	65.00%	67.49%
$(0.0001, 20)$	70.77%	63.16%	60.87%	75.76%	66.67%	67.45%
$(0.00001, 20)$	75.38%	64.91%	63.04%	78.79%	71.67%	70.77%
$(0.00001, 50)$	64.62%	70.18%	67.39%	84.85%	68.33%	71.01%
$(0.00001, 100)$	76.92%	64.91%	69.57%	77.27%	68.33%	71.40%

Table 3: The Accuracy for Decision Stumps with Depth 4 Decision Tree

  - The final learning rate is 0.00001, and the threshold is 100.

- **Conclusion**:
  - Decision tree with more depth is more accurarte.
  - SGD is unstable with large fluctuation. The confidence interval is wide.
  - Decision Stumps is much more acurate than decision tree with same depth and SGD. Decision Stumps constructed by decision tree with depth 4 is much acurate than decision tree with depth 4 and 8.
  - High learning rate might cause earlier saturation so the accuracy is lower.
  - High thredhold can increase the accuracy for Decision Stumps because it avoids the overfitting of each decision tree.

- **Code**:
  Folder: $/bwang34 - hw2/data/decision - trees/src/$
  Files:
  - SGD - $Gradient.java$
  - Decision Stumps - $WekaTester.java$
  - Decision Tree - $Id3.java$
  - Feature Generator - $FeatureGenerator.java$

  .

- **Tree Display**:
  - **Decision Tree**:

```
================================================
ID3, unlimited, fold3

ID3

lastName0=m = 1
|  firstName2=e = 1: +
|  firstName2=e = 0
|  |  lastName1=o = 1: +
|  |  lastName1=o = 0
|  |  |  firstName0=p = 1: -
|  |  |  firstName0=p = 0
|  |  |  |  firstName0=r = 1: -
|  |  |  |  firstName0=r = 0
|  |  |  |  |  firstName0=y = 1: -
|  |  |  |  |  firstName0=y = 0
|  |  |  |  |  |  firstName1=u = 1: -
|  |  |  |  |  |  firstName1=u = 0
|  |  |  |  |  |  |  lastName3=a = 1
|  |  |  |  |  |  |  |  firstName3=r = 1: +
|  |  |  |  |  |  |  |  firstName3=r = 0: -
|  |  |  |  |  |  |  lastName3=a = 0: +
lastName0=m = 0
|  lastName1=l = 1
|  |  firstName0=d = 1: -
|  |  firstName0=d = 0: +
|  lastName1=l = 0
|  |  lastName2=l = 1
|  |  |  firstName2=r = 1: -
|  |  |  firstName2=r = 0
|  |  |  |  lastName4=n = 1: -
|  |  |  |  lastName4=n = 0
|  |  |  |  |  firstName2=h = 1: -
|  |  |  |  |  firstName2=h = 0: +
|  |  lastName2=l = 0
|  |  |  lastName2=o = 1
|  |  |  |  firstName0=b = 1: -
|  |  |  |  firstName0=b = 0: +
|  |  |  lastName2=o = 0
|  |  |  |  firstName3=f = 1: +
|  |  |  |  firstName3=f = 0
```

```
|  |  |  |  |  lastName4=l = 1
|  |  |  |  |  |  firstName2=h = 1: -
|  |  |  |  |  |  firstName2=h = 0
|  |  |  |  |  |  |  lastName0=l = 1: -
|  |  |  |  |  |  |  lastName0=l = 0
|  |  |  |  |  |  |  |  lastName0=q = 1: -
|  |  |  |  |  |  |  |  lastName0=q = 0: +
|  |  |  |  |  lastName4=l = 0
|  |  |  |  |  |  firstName1=o = 1
|  |  |  |  |  |  |  lastName0=f = 1: -
|  |  |  |  |  |  |  lastName0=f = 0
|  |  |  |  |  |  |  |  firstName2=e = 1: -
|  |  |  |  |  |  |  |  firstName2=e = 0
|  |  |  |  |  |  |  |  |  firstName3=a = 1
|  |  |  |  |  |  |  |  |  |  lastName0=h = 1: +
|  |  |  |  |  |  |  |  |  |  lastName0=h = 0: -
|  |  |  |  |  |  |  |  |  firstName3=a = 0
|  |  |  |  |  |  |  |  |  |  firstName2=n = 1: +
|  |  |  |  |  |  |  |  |  |  firstName2=n = 0
|  |  |  |  |  |  |  |  |  |  |  lastName2=n = 1: +
|  |  |  |  |  |  |  |  |  |  |  lastName2=n = 0
|  |  |  |  |  |  |  |  |  |  |  |  lastName1=a = 1: -
|  |  |  |  |  |  |  |  |  |  |  |  lastName1=a = 0
|  |  |  |  |  |  |  |  |  |  |  |  |  firstName3=g = 1: -
|  |  |  |  |  |  |  |  |  |  |  |  |  firstName3=g = 0: +
|  |  |  |  |  |  firstName1=o = 0
|  |  |  |  |  |  |  lastName0=l = 1
|  |  |  |  |  |  |  |  firstName1=a = 1
|  |  |  |  |  |  |  |  |  firstName0=d = 1: +
|  |  |  |  |  |  |  |  |  firstName0=d = 0: -
|  |  |  |  |  |  |  |  firstName1=a = 0: +
|  |  |  |  |  |  |  lastName0=l = 0
|  |  |  |  |  |  |  |  lastName3=m = 1
|  |  |  |  |  |  |  |  |  firstName2=r = 1: -
|  |  |  |  |  |  |  |  |  firstName2=r = 0: +
|  |  |  |  |  |  |  |  lastName3=m = 0
|  |  |  |  |  |  |  |  |  firstName1=e = 1
|  |  |  |  |  |  |  |  |  |  firstName2=n = 1: +
|  |  |  |  |  |  |  |  |  |  firstName2=n = 0
|  |  |  |  |  |  |  |  |  |  |  firstName2=o = 1
|  |  |  |  |  |  |  |  |  |  |  |  lastName0=b = 1: -
|  |  |  |  |  |  |  |  |  |  |  |  lastName0=b = 0: +
|  |  |  |  |  |  |  |  |  |  |  firstName2=o = 0
|  |  |  |  |  |  |  |  |  |  |  |  lastName2=r = 1
|  |  |  |  |  |  |  |  |  |  |  |  |  firstName0=m = 1: -
```

```
|  |  |  |  |  |  |  |  |  |  |  |  |  firstName0=m = 0: +
|  |  |  |  |  |  |  |  |  |  |  |  lastName2=r = 0: -
|  |  |  |  |  |  |  |  |  firstName1=e = 0
|  |  |  |  |  |  |  |  |  |  firstName0=t = 1
|  |  |  |  |  |  |  |  |  |  |  lastName4=e = 1: -
|  |  |  |  |  |  |  |  |  |  |  lastName4=e = 0
|  |  |  |  |  |  |  |  |  |  |  |  lastName0=s = 1: -
|  |  |  |  |  |  |  |  |  |  |  |  lastName0=s = 0: +
|  |  |  |  |  |  |  |  |  |  firstName0=t = 0
|  |  |  |  |  |  |  |  |  |  |  firstName3=o = 1
|  |  |  |  |  |  |  |  |  |  |  |  firstName0=a = 1: +
|  |  |  |  |  |  |  |  |  |  |  |  firstName0=a = 0
|  |  |  |  |  |  |  |  |  |  |  |  |  firstName0=m = 1: +
|  |  |  |  |  |  |  |  |  |  |  |  |  firstName0=m = 0: -
|  |  |  |  |  |  |  |  |  |  |  firstName3=o = 0
|  |  |  |  |  |  |  |  |  |  |  |  firstName4=o = 1
|  |  |  |  |  |  |  |  |  |  |  |  |  firstName0=s = 1: -
|  |  |  |  |  |  |  |  |  |  |  |  |  firstName0=s = 0: +
|  |  |  |  |  |  |  |  |  |  |  |  firstName4=o = 0
|  |  |  |  |  |  |  |  |  |  |  |  |  lastName0=s = 1
|  |  |  |  |  |  |  |  |  |  |  |  |  |  firstName0=d = 1: +
|  |  |  |  |  |  |  |  |  |  |  |  |  |  firstName0=d = 0
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  lastName4=h = 1
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  firstName0=s = 1: -
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  firstName0=s = 0: +
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  lastName4=h = 0: -
|  |  |  |  |  |  |  |  |  |  |  |  |  lastName0=s = 0
|  |  |  |  |  |  |  |  |  |  |  |  |  |  lastName3=l = 1
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  firstName0=d = 1: -
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  firstName0=d = 0: +
|  |  |  |  |  |  |  |  |  |  |  |  |  |  lastName3=l = 0: -
```

```
Correctly Classified Instances          35               76.087  %
Incorrectly Classified Instances        11               23.913  %
Kappa statistic                          0.5199
Mean absolute error                      0.2391
Root mean squared error                  0.489
Relative absolute error                 48.1752 %
Root relative squared error             98.1732 %
Total Number of Instances               46
```

.

– **Decision Stumps**:

```
Decision Stumps, fold4

cs446.homework2.Gradient@681a9515


Correctly Classified Instances          51               77.2727 %
Incorrectly Classified Instances        15               22.7273 %
Kappa statistic                          0.5326
Mean absolute error                      0.2273
Root mean squared error                  0.4767
Relative absolute error                 46.9613 %
Root relative squared error             96.961  %
Total Number of Instances               66
.
```

– **Decision Tree - 8**:

```
==================================================
ID3, depth 8, fold1

ID3

firstName3=f = 1: +
firstName3=f = 0
|  lastName0=c = 1: -
|  lastName0=c = 0
|  |  lastName4=l = 1
|  |  |  lastName0=q = 1: -
|  |  |  lastName0=q = 0: +
|  |  lastName4=l = 0
|  |  |  firstName0=r = 1
|  |  |  |  firstName1=o = 1: +
|  |  |  |  firstName1=o = 0
|  |  |  |  |  firstName1=a = 1: +
|  |  |  |  |  firstName1=a = 0
|  |  |  |  |  |  firstName1=e = 1: +
|  |  |  |  |  |  firstName1=e = 0: -
|  |  |  firstName0=r = 0
|  |  |  |  lastName0=m = 1
|  |  |  |  |  firstName2=n = 1: -
|  |  |  |  |  firstName2=n = 0
```

```
|  |  |  |  |  |  firstName0=p = 1: -
|  |  |  |  |  |  firstName0=p = 0
|  |  |  |  |  |  |  lastName2=t = 1
|  |  |  |  |  |  |  |  firstName0=t = 1: +
|  |  |  |  |  |  |  |  firstName0=t = 0: -
|  |  |  |  |  |  |  lastName2=t = 0: +
|  |  |  |  lastName0=m = 0
|  |  |  |  |  lastName0=l = 1
|  |  |  |  |  |  firstName1=a = 1
|  |  |  |  |  |  |  firstName0=d = 1: +
|  |  |  |  |  |  |  firstName0=d = 0: -
|  |  |  |  |  |  firstName1=a = 0: +
|  |  |  |  |  lastName0=l = 0
|  |  |  |  |  |  lastName3=m = 1
|  |  |  |  |  |  |  firstName2=r = 1: -
|  |  |  |  |  |  |  firstName2=r = 0: +
|  |  |  |  |  |  lastName3=m = 0
|  |  |  |  |  |  |  lastName2=l = 1
|  |  |  |  |  |  |  |  firstName2=r = 1: -
|  |  |  |  |  |  |  |  firstName2=r = 0: +
|  |  |  |  |  |  |  lastName2=l = 0
|  |  |  |  |  |  |  |  lastName3=l = 1: +
|  |  |  |  |  |  |  |  lastName3=l = 0: -
```

```
Correctly Classified Instances          48               73.8462 %
Incorrectly Classified Instances        17               26.1538 %
Kappa statistic                          0.4785
Mean absolute error                      0.3371
Root mean squared error                  0.4722
Relative absolute error                 67.4285 %
Root relative squared error             94.4514 %
Total Number of Instances               65
.
```

– **Decision Tree - 4**:

```
==================================================
ID3, depth 4, fold4

ID3

lastName2=l = 1
```

```
|  firstName2=r = 1: -
|  firstName2=r = 0
|  |  firstName2=m = 1: -
|  |  firstName2=m = 0: +
lastName2=l = 0
|  lastName2=o = 1
|  |  firstName0=d = 1: -
|  |  firstName0=d = 0
|  |  |  firstName2=l = 1: -
|  |  |  firstName2=l = 0: +
|  lastName2=o = 0
|  |  firstName3=f = 1: +
|  |  firstName3=f = 0
|  |  |  lastName0=m = 1
|  |  |  |  firstName0=n = 1: -
|  |  |  |  firstName0=n = 0: +
|  |  |  lastName0=m = 0
|  |  |  |  lastName1=l = 1: +
|  |  |  |  lastName1=l = 0: -
```

```
Correctly Classified Instances          48               72.7273 %
Incorrectly Classified Instances        18               27.2727 %
Kappa statistic                          0.409
Mean absolute error                      0.3812
Root mean squared error                  0.4582
Relative absolute error                 78.7652 %
Root relative squared error             93.1911 %
Total Number of Instances               66
.
```

– **SGD**:

```
==================================================
SGD, fold4

cs446.homework2.Gradient@6b884d57


Correctly Classified Instances          50               75.7576 %
Incorrectly Classified Instances        16               24.2424 %
Kappa statistic                          0.5152
Mean absolute error                      0.2424
```

```
Root mean squared error                  0.4924
Relative absolute error                 50.0921 %
Root relative squared error            100.1409 %
Total Number of Instances               66
.
```