**CS446: Machine Learning** **Fall 2015**

# Problem Set 2

Boyu Li *Handed In: Hand in date*

1. Answer to problem 1

   (a) We use information gain to determine the attribute.
   First we calculate Entropy for this problem, $p = 35/50$ and $n = 15/50$. H(Y) = $-(35/50)\log_2(35/50) - (15/50)\log_2(15/50) = 0.88129$

      i. **Holiday:**
      $\{\text{Holidaty} = \text{yes}\} : p = 5/15, n = 10/15$,so $H_y = -(5/15)\log_2(5/15) - (10/15)log_2(10/15) = 0.9183$
      $\{\text{Holidaty} = \text{no}\} : p = 30/35, n = 5/35$,so $H_n = -(30/35)\log_2(30/35) - (5/35)log_2(5/35) = 0.59167$
      The information gain for the **Holiday** is :

      $$Gain(Holiday) = H(Y) - (15/50)H_y - (35/50)H_n = 0.19163$$

      ii. **Exam Tomorrow:**
      $\{\text{Exam Tomorrow} = \text{yes}\} : p = 15/16, n = 1/16$, so $H_y = -(15/16)\log_2(15/16) - (1/16)\log_2(1/16) = 0.33729$
      $\{\text{Exam Tomorrow} = \text{no}\} : p = 20/34, n = 14/34$, so $H_y = -(20/34)\log_2(20/34) - (14/34)\log_2(13/34) = 0.97742$
      The information gain for the **Exam Tomorrow** is :

      $$Gain(Holiday) = H(Y) - (16/50)H_y - (34/50)H_n = 0.10871$$

   So we know that Gain(Holiday) is grater than Gain(Exam Tomorrow),so we choose **Holiday** as the root attribute

   (b) the tree I make follows the following rule:

```
if color = yellow:
  if size = large:
    if age = Adult:
          if Act = Stretch
             Class = T;
          if Act != Stretch:
             Class = F;
    if age != Adult:
             Class = F;
  if size != Large:
             Class = T;
```

```
if color != yellow:
  if Age = Adult:
     if Act = Stretch:
             Class = T;
     if Act != Stretch;
             Class = F;
  if Age != Adult:
            Class = F;
```

(c) Because for the ID3 algorithm, it try to make sure the local optimal choice in every node, but in this way it may underestimate the with other nodes. So we can make sure that it is locally optimal in each node, but we can not make sure that it is not globally optimal choice.

2. Answer to problem 2

(a) For this problem I just expand the "FeatureGenerator.java" file, to expand the feature that contain at most 10 characters for the name, 5 comes from first name and 5 comes the last name. And for each instance we made, we add this features in.

(b) This one is the SGD function I choose the learning rate $\alpha$ with 0.01 and the error threshold with 50.

| p1 | p2 | p3 | p4 | p5 |
|---|---|---|---|---|
| 0.779661 | 0.627119 | 0.745763 | 0.59322 | 0.689655 |

$p_a = 0.68797$
S = 0.078145
CI = $(0.5261, 0.8479)$

(c) This one is for the ID3 depth = -1

| p1 | p2 | p3 | p4 | p5 |
|---|---|---|---|---|
| 0.694915 | 0.762712 | 0.728814 | 0.610169 | 0.62968 |

$p_a = 0.6836734$
S = 0.06464669
CI = $(0.55215, 0.8186)$

(d) This one is for the ID3 depth = 4

| p1 | p2 | p3 | p4 | p5 |
|---|---|---|---|---|
| 0.423729 | 0.711864 | 0.728814 | 0.525424 | 0.603448 |

$p_a = 0.598639$
S = 0.1282015
CI = $(0.33469, 0.862619)$

(e) This one is for the ID3 depth = 8

| p1 | p2 | p3 | p4 | p5 |
|---|---|---|---|---|
| 0.474576 | 0.745763 | 0.728814 | 0.59322 | 0.603448 |

$p_a = 0.629517$
S = 0.11109
CI = $(0.400413, 0.857903)$

(f) This one is for the Decision Stumps

| p1 | p2 | p3 | p4 | p5 |
|---|---|---|---|---|
| 0.779661 | 0.627119 | 0.745763 | 0.59322 | 0.689655 |

$p_a = 0.68707$
S = 0.078128
CI =$(0.5262, 0.84792)$

3. I get the following rank according to their accuracy in decreasing trend.
   1. **SGD**
   2. **Decision Stumps**
   3. **ID3 Without Max Depth**
   4. **ID3 with Max Depth(8)**
   5. **ID3 with Max Depth(4)**

   Then we need to compare the for each pair

   (a) **SGD** and **Decision Stumps**
   The average difference value of two pair is nearly 0. So the t-value is nearly 0 and the $t_\alpha/2$ for freedom equals 4 is 4.604.
   So t-value¡t-test. So this two pair are not statistically significant.

   (b) **Decision Stumps** and **ID3 without Max Depth**
   $diff_{avg} = 0.0018076$
   S(diff) = 0.0861963
   t-value = $diff_{avg}/(S(diff)/\sqrt{5}) = 0.04689$
   So t-value¡t-test. So this two pair are not statistically significant.

   (c) **ID3 without Max Depth** and **ID3 with Max Depth(8)**
   $diff_{avg} = 0.0560992$
   S(diff) = 0.0923011
   t-value = $diff_{avg}/(S(diff)/\sqrt{5}) = 1.359$
   So t-value¡t-test. So this two pair are not statistically significant.

   (d) **ID3 with Max Depth(8)** and **ID3 with Max Depth(4)**
   $diff_{avg} = 0.030503$
   S(diff) = 0.03031

t-value = $diff_{avg}/(S(diff)/\sqrt{5}) = 2.2496$
So t-value¡t-test. So this two pair are not statistically significant.

(e) **Evalutions:**
Just from the average accuracy,we may propose that the SGD is the best the algorithm. But we can see that there is overlap between the confidence interval for SGD and Max Depth without max depth and the t-value is really small for the difference between their accuracy, so we may think that SGD as same good as ID3 without max depth algorithm.

4. the following best accuracy decision tree shows like that:
**The Decision tree for the ID3 without max depth:**

```
firstName1=a = 1
|  lastName1=o = 1: +
|  lastName1=o = 0
|  |  firstName3=y = 1: -
|  |  firstName3=y = 0
|  |  |  lastName1=n = 1: -
|  |  |  lastName1=n = 0
|  |  |  |  lastName2=m = 1: -
|  |  |  |  lastName2=m = 0
|  |  |  |  |  lastName2=l = 1
|  |  |  |  |  |  firstName2=n = 1
|  |  |  |  |  |  |  firstName3=i = 1: -
|  |  |  |  |  |  |  firstName3=i = 0: +
|  |  |  |  |  |  firstName2=n = 0: -
|  |  |  |  |  lastName2=l = 0
|  |  |  |  |  |  firstName0=m = 1: +
|  |  |  |  |  |  firstName0=m = 0
|  |  |  |  |  |  |  firstName3=k = 1: -
|  |  |  |  |  |  |  firstName3=k = 0
|  |  |  |  |  |  |  |  lastName1=a = 1
|  |  |  |  |  |  |  |  |  firstName0=s = 1: -
|  |  |  |  |  |  |  |  |  firstName0=s = 0
|  |  |  |  |  |  |  |  |  |  firstName0=p = 1: -
|  |  |  |  |  |  |  |  |  |  firstName0=p = 0: +
|  |  |  |  |  |  |  |  lastName1=a = 0
|  |  |  |  |  |  |  |  |  firstName2=m = 1: -
|  |  |  |  |  |  |  |  |  firstName2=m = 0
|  |  |  |  |  |  |  |  |  |  lastName1=k = 1: -
|  |  |  |  |  |  |  |  |  |  lastName1=k = 0: +
firstName1=a = 0
|  firstName4=a = 1
|  |  lastName4=e = 1
```

```
|  |  |  firstName0=m = 1: +
|  |  |  firstName0=m = 0
|  |  |  |  firstName1=t = 1: +
|  |  |  |  firstName1=t = 0: -
|  |  lastName4=e = 0
|  |  |  lastName0=z = 1: -
|  |  |  lastName0=z = 0
|  |  |  |  lastName2=z = 1: -
|  |  |  |  lastName2=z = 0: +
|  firstName4=a = 0
|  |  firstName3=a = 1
|  |  |  lastName0=s = 1
|  |  |  |  firstName0=m = 1: +
|  |  |  |  firstName0=m = 0: -
|  |  |  lastName0=s = 0
|  |  |  |  firstName0=b = 1
|  |  |  |  |  lastName0=d = 1: -
|  |  |  |  |  lastName0=d = 0: +
|  |  |  |  firstName0=b = 0: +
|  |  firstName3=a = 0
|  |  |  firstName3=d = 1: +
|  |  |  firstName3=d = 0
|  |  |  |  firstName2=a = 1
|  |  |  |  |  lastName3=t = 1: +
|  |  |  |  |  lastName3=t = 0
|  |  |  |  |  |  lastName0=c = 1: +
|  |  |  |  |  |  lastName0=c = 0
|  |  |  |  |  |  |  lastName0=m = 1: +
|  |  |  |  |  |  |  lastName0=m = 0
|  |  |  |  |  |  |  |  lastName0=k = 1
|  |  |  |  |  |  |  |  |  firstName1=m = 1: -
|  |  |  |  |  |  |  |  |  firstName1=m = 0: +
|  |  |  |  |  |  |  |  lastName0=k = 0
|  |  |  |  |  |  |  |  |  firstName0=f = 1: +
|  |  |  |  |  |  |  |  |  firstName0=f = 0: -
|  |  |  |  firstName2=a = 0
|  |  |  |  |  lastName3=r = 1: +
|  |  |  |  |  lastName3=r = 0
|  |  |  |  |  |  firstName4=n = 1
|  |  |  |  |  |  |  firstName0=g = 1: -
|  |  |  |  |  |  |  firstName0=g = 0: +
|  |  |  |  |  |  firstName4=n = 0
|  |  |  |  |  |  |  firstName3=n = 1
|  |  |  |  |  |  |  |  lastName0=r = 1: +
|  |  |  |  |  |  |  |  lastName0=r = 0
```

```
|  |  |  |  |  |  |  |  |  lastName2=a = 1
|  |  |  |  |  |  |  |  |  |  firstName0=r = 1: -
|  |  |  |  |  |  |  |  |  |  firstName0=r = 0: +
|  |  |  |  |  |  |  |  |  lastName2=a = 0: -
|  |  |  |  |  |  |  firstName3=n = 0
|  |  |  |  |  |  |  |  firstName2=d = 1
|  |  |  |  |  |  |  |  |  firstName0=a = 1
|  |  |  |  |  |  |  |  |  |  lastName0=b = 1: +
|  |  |  |  |  |  |  |  |  |  lastName0=b = 0: -
|  |  |  |  |  |  |  |  |  firstName0=a = 0: +
|  |  |  |  |  |  |  |  firstName2=d = 0
|  |  |  |  |  |  |  |  |  firstName2=n = 1
|  |  |  |  |  |  |  |  |  |  firstName1=o = 1
|  |  |  |  |  |  |  |  |  |  |  firstName0=l = 1: -
|  |  |  |  |  |  |  |  |  |  |  firstName0=l = 0: +
|  |  |  |  |  |  |  |  |  |  firstName1=o = 0: -
|  |  |  |  |  |  |  |  |  firstName2=n = 0
|  |  |  |  |  |  |  |  |  |  firstName3=o = 1
|  |  |  |  |  |  |  |  |  |  |  firstName0=a = 1: +
|  |  |  |  |  |  |  |  |  |  |  firstName0=a = 0
|  |  |  |  |  |  |  |  |  |  |  |  firstName2=k = 1: +
|  |  |  |  |  |  |  |  |  |  |  |  firstName2=k = 0: -
|  |  |  |  |  |  |  |  |  |  firstName3=o = 0
|  |  |  |  |  |  |  |  |  |  |  firstName0=n = 1: +
|  |  |  |  |  |  |  |  |  |  |  firstName0=n = 0: -
```

| | | |
|---|---|---|
| Correctly Classified Instances | 45 | 76.2712 % |
| Incorrectly Classified Instances | 14 | 23.7288 % |

**The decision tree for ID3 with max depth 4:**

```
firstName1=a = 1
|  lastName1=n = 1: -
|  lastName1=n = 0
|  |  firstName0=g = 1: -
|  |  firstName0=g = 0
|  |  |  lastName1=k = 1: -
|  |  |  lastName1=k = 0
|  |  |  |  lastName3=g = 1: -
|  |  |  |  lastName3=g = 0: +
firstName1=a = 0
|  firstName4=a = 1
|  |  lastName3=g = 1: -
|  |  lastName3=g = 0
```

```
|  |  |  firstName2=n = 1
|  |  |  |  firstName3=n = 1: +
|  |  |  |  firstName3=n = 0: -
|  |  |  firstName2=n = 0
|  |  |  |  lastName4=a = 1: -
|  |  |  |  lastName4=a = 0: +
|  firstName4=a = 0
|  |  firstName3=a = 1
|  |  |  firstName0=y = 1: -
|  |  |  firstName0=y = 0
|  |  |  |  lastName1=a = 1: -
|  |  |  |  lastName1=a = 0: +
|  |  firstName3=a = 0
|  |  |  firstName0=a = 1
|  |  |  |  lastName3=m = 1: -
|  |  |  |  lastName3=m = 0: +
|  |  |  firstName0=a = 0
|  |  |  |  firstName2=a = 1: +
|  |  |  |  firstName2=a = 0: -
```

| | | |
|---|---|---|
| Correctly Classified Instances | 43 | 72.8814 % |
| Incorrectly Classified Instances | 16 | 27.1186 % |

**The decision tree for ID3 with max depth 4:**

```
firstName1=a = 1
|  lastName1=o = 1: +
|  lastName1=o = 0
|  |  firstName3=y = 1: -
|  |  firstName3=y = 0
|  |  |  lastName1=n = 1: -
|  |  |  lastName1=n = 0
|  |  |  |  lastName2=m = 1: -
|  |  |  |  lastName2=m = 0
|  |  |  |  |  lastName2=l = 1
|  |  |  |  |  |  firstName2=n = 1
|  |  |  |  |  |  |  firstName3=i = 1: -
|  |  |  |  |  |  |  firstName3=i = 0: +
|  |  |  |  |  |  firstName2=n = 0: -
|  |  |  |  |  lastName2=l = 0
|  |  |  |  |  |  firstName0=m = 1: +
|  |  |  |  |  |  firstName0=m = 0
|  |  |  |  |  |  |  firstName3=k = 1: -
|  |  |  |  |  |  |  firstName3=k = 0
```

```
|  |  |  |  |  |  |  |  lastName1=a = 1: +
|  |  |  |  |  |  |  |  lastName1=a = 0: +
firstName1=a = 0
|  firstName4=a = 1
|  |  lastName4=e = 1
|  |  |  firstName0=m = 1: +
|  |  |  firstName0=m = 0
|  |  |  |  firstName1=t = 1: +
|  |  |  |  firstName1=t = 0: -
|  |  lastName4=e = 0
|  |  |  lastName0=z = 1: -
|  |  |  lastName0=z = 0
|  |  |  |  lastName2=z = 1: -
|  |  |  |  lastName2=z = 0: +
|  firstName4=a = 0
|  |  firstName3=a = 1
|  |  |  lastName0=s = 1
|  |  |  |  firstName0=m = 1: +
|  |  |  |  firstName0=m = 0: -
|  |  |  lastName0=s = 0
|  |  |  |  firstName0=b = 1
|  |  |  |  |  lastName0=d = 1: -
|  |  |  |  |  lastName0=d = 0: +
|  |  |  |  firstName0=b = 0: +
|  |  firstName3=a = 0
|  |  |  firstName3=d = 1: +
|  |  |  firstName3=d = 0
|  |  |  |  firstName2=a = 1
|  |  |  |  |  lastName3=t = 1: +
|  |  |  |  |  lastName3=t = 0
|  |  |  |  |  |  lastName0=c = 1: +
|  |  |  |  |  |  lastName0=c = 0
|  |  |  |  |  |  |  lastName0=m = 1: +
|  |  |  |  |  |  |  lastName0=m = 0
|  |  |  |  |  |  |  |  lastName0=k = 1: +
|  |  |  |  |  |  |  |  lastName0=k = 0: -
|  |  |  |  firstName2=a = 0
|  |  |  |  |  lastName3=r = 1: +
|  |  |  |  |  lastName3=r = 0
|  |  |  |  |  |  firstName4=n = 1
|  |  |  |  |  |  |  firstName0=g = 1: -
|  |  |  |  |  |  |  firstName0=g = 0: +
|  |  |  |  |  |  firstName4=n = 0
|  |  |  |  |  |  |  firstName3=n = 1
|  |  |  |  |  |  |  |  lastName0=r = 1: +
```

```
|  |  |  |  |  |  |  |  lastName0=r = 0: -
|  |  |  |  |  |  |  firstName3=n = 0
|  |  |  |  |  |  |  |  firstName2=d = 1: +
|  |  |  |  |  |  |  |  firstName2=d = 0: -


Correctly Classified Instances          44               74.5763 %
Incorrectly Classified Instances        15               25.4237 %
```

**The decision vector for the SGD:**

```
0.191
0.048
-0.049
-0.086
0.027
-0.037
-0.411
-0.248
0.019
-0.026
0.055
0.048
0.217
0.333
0.056
0.105
-0.001
-0.062
-0.064
-0.070
-0.079
0.143
-0.169
-0.003
0.040
-0.001
0.420
-0.001
-0.001
0.064
-0.285
-0.039
-0.001
-0.015
```

```
0.121
-0.001
-0.001
0.104
0.018
-0.057
-0.220
-0.001
-0.001
0.049
-0.079
0.003
-0.015
-0.001
-0.001
-0.001
-0.031
-0.001
0.216
-0.319
0.124
0.172
-0.141
-0.086
-0.001
0.226
-0.127
0.096
0.073
-0.123
-0.139
0.214
-0.095
0.099
-0.001
-0.121
-0.160
0.171
0.210
0.053
-0.158
-0.001
-0.016
-0.093
0.314
```

```
-0.001
0.021
0.167
-0.073
0.034
-0.091
0.007
0.301
-0.218
0.110
0.057
-0.108
0.317
0.026
-0.115
-0.001
0.055
-0.159
-0.132
0.079
-0.080
0.068
0.014
-0.118
-0.001
0.480
-0.001
-0.092
0.016
-0.558
-0.105
-0.038
0.140
-0.338
-0.027
0.055
0.284
-0.011
0.338
0.008
-0.081
-0.001
-0.246
-0.280
0.189
```

```
-0.146
0.044
-0.001
-0.001
0.029
-0.001
-0.001
-0.001
-0.001
0.083
0.098
0.120
0.012
0.075
-0.237
0.083
0.053
-0.194
0.000
-0.084
0.251
-0.083
-0.053
0.276
-0.098
0.066
-0.109
0.096
-0.183
-0.001
-0.035
-0.158
-0.239
0.109
-0.001
-0.001
-0.152
0.056
-0.010
0.405
-0.036
-0.035
-0.196
0.030
0.044
```

```
-0.393
0.323
0.140
-0.001
0.223
-0.019
0.093
-0.217
-0.051
-0.095
-0.001
-0.001
-0.001
0.077
-0.190
-0.012
0.071
-0.126
-0.001
0.042
0.207
-0.127
-0.001
0.011
-0.287
-0.301
-0.106
0.398
-0.073
-0.001
0.258
0.140
-0.027
0.080
-0.099
-0.043
-0.001
0.031
-0.090
-0.265
0.070
0.048
-0.158
0.291
-0.001
```

```
-0.375
0.260
-0.061
-0.076
-0.069
-0.149
-0.159
-0.169
0.125
0.017
-0.001
0.324
-0.100
0.318
0.042
0.051
0.207
-0.001
-0.146
-0.030
-0.411
0.089
0.074
0.111
-0.174
0.006
-0.034
0.322
0.037
-0.025
0.180
0.053
-0.054
0.170
-0.021
0.019
-0.001
0.067
-0.132
0.155
-0.132
0.116
-0.001
-0.001
0.055
```

```
0.005663043017077229]

Correctly Classified Instances          46               77.9661 %
Incorrectly Classified Instances        13               22.0339 %
```

From the accuracy of three ID3 tree with different max depth. ID3 without specifying any max depth has the highest accuracy, with the pruned level, the accuracy goes down. so we may think that it has the chance that overfitting the data so that do predict test data no very well. For the decision stumps algorithm,we first generate 100 decision stumps with ID3 algorithm with max depth in 4 and then use SGD algorithm to calculate accuracy.This is more accurate solution because it performs well on the test data.Also, because we randomize the training smaples and then doing a sample over the training data, it may get better result for decision stumps than the decision tree.