Ford Tang

46564602

CS 178

Homework 4

1.

   a. Y = +1 : 4/10
   Y = -1 : 6/10
   Entropy: - (4/10) log (4/10) – (6/10) log (6/10) = 0.970950594

   b. P(X1 = 0) = 4/10
   P(X1 = 1) = 6/10
   P(X2 = 0) = 5/10
   P(X2 = 1) = 5/10
   P(X3 = 0) = 3/10
   P(X3 = 1) = 7/10
   P(X4 = 0) = 3/10
   P(X4 = 1) = 7/10
   P(X5 = 0) = 7/10
   P(X5 = 1) = 3/10
   Entropy (y = +1 | X1 = 0) = 0.970950594
   Entropy (y = +1 | X1 = 1) = 0.970950594
   Entropy (y = +1 | X2 = 0) = 1
   Entropy (y = +1 | X2 = 1) = 1
   Entropy (y = +1 | X3 = 0) = 0.881290899
   Entropy (y = +1 | X3 = 1) = 0.881290899
   Entropy (y = +1 | X4 = 0) = 0.881290899
   Entropy (y = +1 | X4 = 1) = 0.881290899
   Entropy (y = +1 | X5 = 0) = 0.881290899
   Entropy (y = +1 | X5 = 1) = 0.881290899

   Split on feature X2.

2.

   a.
   ```
   >> X = load('data/kaggle.X1.train.txt');
   >> Y = load('data/kaggle.Y.train.txt');
   >> [Xtr Xte Ytr Yte] = splitData(X,Y, .75);
   >> dt = treeRegress(Xtr, Ytr, 'maxDepth', 20);
   >> mse(dt, Xte, Yte)
   ans = 0.7344
   ```

   b.
   ```
   >> for i = 0:15;
   i,
   dt = treeRegress(Xtr, Ytr, 'maxDepth', i);
   mse(dt, Xtr, Ytr),
   mse(dt, Xte, Yte),
   ```

```
end;
```

As we go deeper, the complexity increases. After the 7th level, overfitting occurs. The 7th level depth is best.

c.

```
>> for i=2.^[3:12],
dt = treeRegress(Xtr, Ytr, 'minParent', i);
i,
mse(dt, Xtr, Ytr),
mse(dt, Xte, Yte),
end;
```

Complexity is decreasing as I increases. After 512 overfitting occurs. 512 is the best depth.

d.

```
>> dt = treeRegress(Xtr, Ytr, 'minParent', 512);
Xpredict = load('data/kaggle.X1.test.txt');
Ypredict = predict(dt, Xpredict);
file_name = fopen('FordTang.csv', 'w');
fprintf(file_name,'ID,Prediction\n');
for i = 1:length(Ypredict),
fprintf(file_name, '%d,%d\n', i, Ypredict(i));
end;
fclose(file_name);
```