NYCU Pattern Recognition, HW3

310551079 梁友誠

Part. 1, Coding:

1. Entropy and Gini Index of the array

```
print("Gini of data is ", gini(data))
✓ 0.4s
Gini of data is  0.4628099173553719
```

```
print("Entropy of data is ", entropy(data))
✓ 0.3s
Entropy of data is  0.9456603046006401
```

2. Decision Tree:

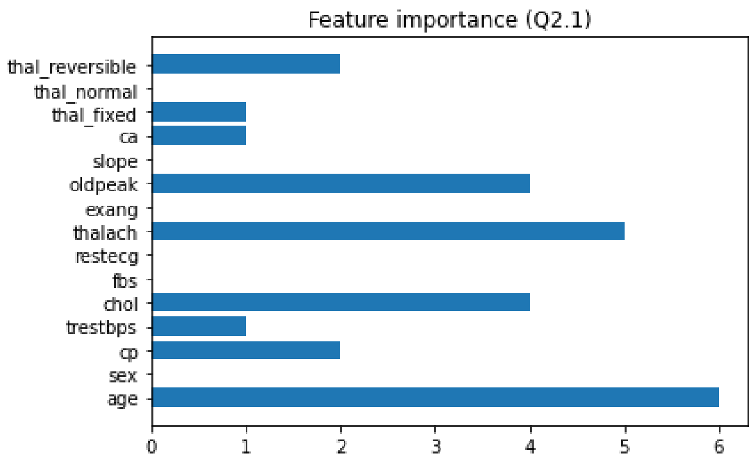
2.1: Criterion = 'gini', Max_depth=3 and Max_depth=10:

```
criterion=gini, max_depth=3, accuracy: 0.78
criterion=gini, max_depth=10, accuracy: 0.69
```

2.2 Max_depth=3, Criterion='gini' and Criterion='entropy':

```
criterion=gini, max_depth=3, accuracy: 0.78
criterion=entropy, max_depth=3, accuracy: 0.75
```

3. Feature importance of my decision tree (max_depth = 10):



4. AdaBoost (n_estimators=10 and n_estimators=100):

```
n_estimators = 10, criterion=entropy, max_depth=1, accuracy: 0.81
```

```
n_estimators = 100, criterion=entropy, max_depth=1, accuracy: 0.81
```

5. Random Forest:
   5.1: Criterion='gini', Max_depth = None, Max_features = sqrt(n_features), Bootstrap=True, n_estimators=10 and n_estimators=100:

```
n_estimators=10, max_features=sqrt(n_features), accuracy: 0.78
n_estimators=100, max_features=sqrt(n_features), accuracy: 0.84
```

   5.2: Criterion = 'gini', Max_depth = None, N_estimators = 10, Bootstrap = True, Max_features=sqrt(n_features) and Max_features=n_features

```
n_estimators=10, max_features=sqrt(n_features), accuracy: 0.84
n_estimators=10, max_features=n_features, accuracy: 0.79
```

6. I used the upsampling of class0 as the new training dataset to increase the accuracy of the model. The ratio of class0 and class1 is 63:135 at first, after upsampling twice of class0, the ratio of class0 and class1 will become 126:135. And then I use this new dataset to train Random Forest. The hyperparameter of the model is *n_estimators = 10~20, max_features = sqrt(n_features)*2, decision tree depth = 2, criterion = "entropy"*. Due to the randomness of Random Forest, we will not get the same accuracy score on the testing dataset after each training. Therefore, I need to train the model several times and then choose the best model with the best accuracy > 0.85. The best accuracy I can get at last is 0.86.

```
Test-set accuarcy score:  0.86
```

Part. 2, Questions:

Q1.

Model A misclassification rate : $\frac{100+100}{800} = 25\%$

Model B misclassification rate : $\frac{200}{800} = 25\%$

$\therefore$ misclassification rate is same in Model A and B.

Model A cross entropy :

first node : $-(\underbrace{\frac{3}{4}\cdot\log_2(\frac{3}{4})}_{-0.31} + \underbrace{\frac{1}{4}\cdot\log_2(\frac{1}{4})}_{-0.5}) \doteqdot 0.81$

second node : $-(\frac{3}{4}\cdot\log_2(\frac{3}{4}) + \frac{1}{4}\cdot\log_2(\frac{1}{4})) \doteqdot 0.81$

$\therefore$ total entropy $= \frac{1}{2}\cdot 0.81 + \frac{1}{2}\times 0.81 = 0.81$

Model A gini index :

first node: $1-(\frac{3}{4})^2-(\frac{1}{4})^2 = 0.375$

second node : $1-(\frac{3}{4})^2-(\frac{1}{4})^2 : 0.375$

$\therefore$ total gini $= \frac{1}{2}\times 0.375 + \frac{1}{2}\times 0.375 : 0.375$

Mode B cross entropy :

first node : $-(\underbrace{\frac{1}{3}\cdot\log_2(\frac{1}{3})}_{-0.528} + \underbrace{\frac{2}{3}\cdot\log_2(\frac{2}{3})}_{-0.389}) \doteqdot 0.918$

second node : 0

total entropy : $\frac{3}{4}\cdot 0.918 \doteqdot 0.688$

Model B gini index :

First node : $1-(\frac{1}{3})^2-(\frac{2}{3})^2 = \frac{4}{9}$

second node : 0

total gini : $\frac{3}{4}\cdot\frac{4}{9} = \frac{1}{3} \doteqdot 0.33$

$\therefore$ $0.688 < 0.81$, $0.33 < 0.375$ both entropy and gini index in tree B are lower than tree A

Q2:

$$E_{x,t}[e^{-ty(x)}] = \sum_t \int e^{-ty(x)} p(t|x)\, p(x)\, dx$$

we know that t is defined as target value -1 or 1

so we can rewrite $E_{x,t}[e^{-ty(x)}]$ to

$$\int [e^{-y(x)} p(t=1|x) + e^{y(x)} p(t=-1|x)]\, p(x)\, dx$$

To minimize the error function, we set the derivative of the error function E w.r.t $y(x)$ to 0.

$$\therefore \frac{dE}{dy(x)} = \frac{d}{dy(x)} \int [e^{-y(x)} p(t=1|x) + e^{y(x)} p(t=-1|x)]\, p(x)\, dx = 0$$

$$\Rightarrow y'(x)\, e^{y(x)} p(t=-1|x) - y'(x)\, e^{-y(x)} p(t=1|x) = 0$$

$$\Rightarrow y'(x)\, e^{y(x)} p(t=-1|x) = y'(x)\, e^{-y(x)} p(t=1|x)$$

$$\Rightarrow e^{y(x)} p(t=-1|x) = e^{-y(x)} p(t=1|x)$$

$$\Rightarrow \frac{e^{y(x)}}{e^{-y(x)}} = \frac{p(t=1|x)}{p(t=-1|x)}$$

$$\Rightarrow e^{2y(x)} = \frac{p(t=1|x)}{p(t=-1|x)}$$

$$\Rightarrow 2y(x) = \ln\left(\frac{p(t=1|x)}{p(t=-1|x)}\right)$$

$$\Rightarrow y(x) = \frac{1}{2} \ln\left(\frac{p(t=1|x)}{p(t=-1|x)}\right)$$

$\therefore y(x) = \frac{1}{2} \ln\left(\frac{p(t=1|x)}{p(t=-1|x)}\right)$ is the minimizing function