

Part1 Coding :

1. gini index and entropy

```
▶ print("Gini of data is ", gini(data))  
⇒ Gini of data is 0.4628099173553719  
  
[ ] print("Entropy of data is ", entropy(data))  
⇒ Entropy of data is 0.9456603046006401
```

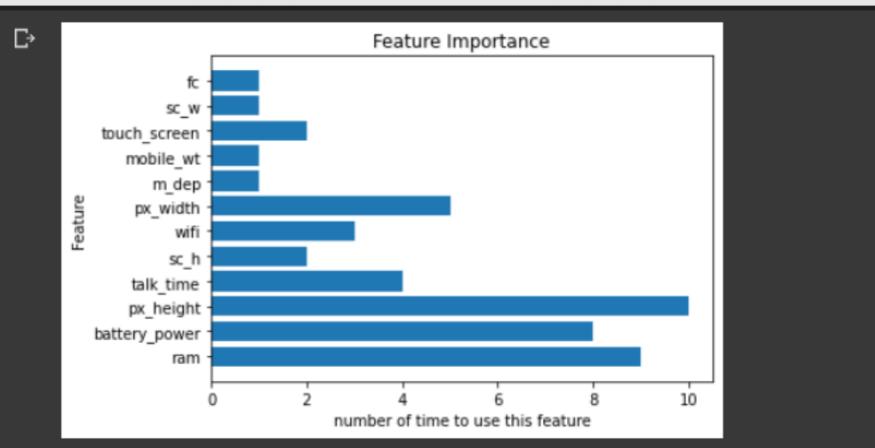
2.1

```
⇒ Accuracy using clf_depth3 on validation data : 0.92  
Accuracy using clf_depth10 on validation data : 0.9433333333333334
```

2.2

```
⇒ Accuracy using clf_gini on validation data : 0.92  
Accuracy using clf_entropy on validation data : 0.9333333333333333
```

3. feature importance(樹是clf depth 10的那一棵)



4.

```
⇒ Accuracy using clf_adaboost10 on validation data : 0.91  
Accuracy using clf_adaboost100 on validation data : 0.9666666666666667
```

5.1 (這裡名字取的不太好,抱歉,這兩個是random forest的)

```
⇒ Accuracy using clf_10 on validation data : 0.9333333333333333  
Accuracy using clf_100 on validation data : 0.9366666666666666
```

5.2

```
⇒ Accuracy using clf_random_features on validation data : 0.9133333333333333  
⇒ Accuracy using clf_all_features on validation data : 0.9566666666666667
```

6. 試跑下方do not modify code的部分

```
*** We will check your result for Question 3 manually *** (5 points)  
*** We will check your result for Question 6 manually *** (20 points)  
Approximate score range: 45.0 ~ 70.0  
*** This score is only for reference ***
```

Part 2 Questions:

所以試想

1.

因為我們在訓練的過程中總是想要找到 1 個分類方式使得子樹兩端的 labels 皆一致。當我們在某個父節點要將極少的 columns 分成 2 個子樹時，我們會因為上述核心思想而傾向於介由某幾行的某個特徵去拆分，但這通常無法 generalized

under 某個 threshold 下

到任一資料集的表現上，因此 decision tree 容易出現 overfitting.

decision tree 是有機會可以百分之百 fit training data 的，因為當 2 筆不同 label 的資料要被分成 2 類時，我們一定可以找到這 2 筆資料的 ^{feature value} 不同之處並以此作分類依據，所以我們可以想見任意 N 筆資料若不限制樹深度的作任意拆分，最壞情況就是怎麼拆分都不乾淨 直到父節點剩 2 筆不同 (只為了分出 label 的資料，
分) 此時我們就可以套用上述狀況去做到百分之百正確率的拆分 (當然，這個論述不適用於 2 筆 features 皆相同的 data 谷得有著不同 labels 的情況) (此時生成的 2 個葉子節點僅包含 1 筆資料，此分類依據是很不 general 的)

Reduce risk of overfitting a decision tree:

① 限制樹的深度

② 限制 minimum number

of data per leaf^③ 在分離節點時，最多考慮幾種特徵值：

樹深限制以防為了拆分出 pure 的子樹而過度

生長；葉節點資料量限制以防為了分出少數幾

筆資料而無限制的拆分下去；特徵值限制

也如同前述 2 者的概念，ex：為了分出 2 款飲料是否

熱銷，最後竟以瓶身顏色做為判斷，很明顯此 feature 是分無可分的情況下才給的一個不合理判

斷準則 \Rightarrow 這 3 者核心概念皆相似，都是為了使

decision tree 不要為了完美 fit training data 而

使得分類不 generalized 所給出的限制。

方法

2.

a.

$$\text{True, } D_{t+1}(i) = \frac{D_t(i) \cdot \exp(-\alpha_t \cdot y_i \cdot h_t(x_i))}{z_t},$$

→ 這些 missclassified 的 weight 都乘上 $\frac{e^{(\alpha_t)}}{z_t}$ 作更新

b.

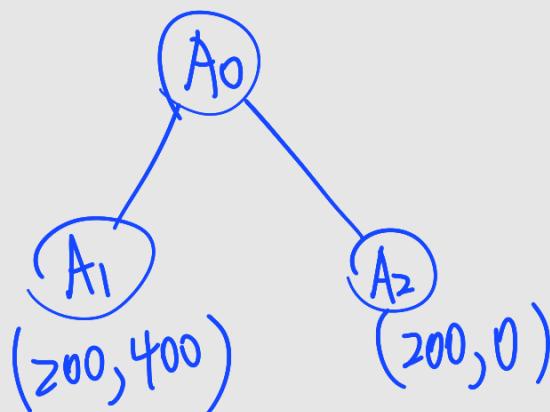
True, In the course of boosting iterations, the weights will increase for datas that are repeatedly missclassified by the weak classifiers. The weighted training error E_t of the t^{th} weak classifier on the training data therefore tends to increase.

c.

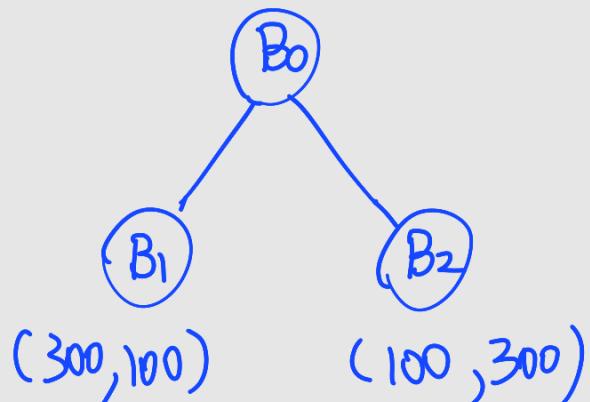
False, If the data is not linearly separable, and the weak classifier we use is linear classifier, then the linear combination of such weak classifiers cannot give zero training error.

3.

Tree model A:



Tree model B:



miss classification rate:

$$\frac{200}{800} = 25\%$$

miss classification rate:

$$\frac{100+100}{800} = 25\%$$

Cross-entropy A

$$\begin{aligned} & \frac{600}{800} \left(-\frac{1}{3} \log_2 \frac{1}{3} + -\frac{2}{3} \log_2 \frac{2}{3} \right) \\ & + \frac{200}{800} \left(-1 \log_2 1 \right) \end{aligned}$$

$$\begin{aligned} & = \frac{3}{4} \left(-\frac{1}{3} (0 - \log_2 \frac{3}{3}) + \frac{2}{3} (1 - \log_2 \frac{3}{3}) \right) + \frac{1}{4} \times 0 \\ & = \frac{3}{4} \left(\frac{1}{3} \log_2 3 - \frac{2}{3} + \frac{2}{3} \log_2 \frac{3}{3} \right) = \frac{1}{2} + \frac{3}{4} \log_2 \frac{3}{3} \# \end{aligned}$$

Gini A

$$\begin{aligned} & \frac{3}{4} \left(1 - \left(\frac{1}{3} \right)^2 - \left(\frac{2}{3} \right)^2 \right) + \frac{1}{4} (1 - 1^2) \\ & = \frac{1}{3} \# \end{aligned}$$

cross-entropy B

$$\begin{aligned}& \frac{400}{800} \left(-\frac{3}{4} \log_2 \frac{3}{4} + \frac{1}{4} \log_2 \frac{1}{4} \right) \times 2 \\&= -\frac{3}{4} (\log_2 3 - 2) - \frac{1}{4} (0 - 2) \\&= 2 - \frac{3}{4} \log_2 3 \quad \#\end{aligned}$$

Gini B

$$\begin{aligned}& \frac{400}{800} \left(1 - \left(\frac{1}{4} \right)^2 - \left(\frac{3}{4} \right)^2 \right) \times 2 \\&= 1 - \frac{1}{16} - \frac{9}{16} \\&= \frac{3}{8} \quad \#\end{aligned}$$