Part. 1, Coding (70%):

Q1:

```
Gini of data is  0.4628099173553719


Entropy of data is  0.9456603046006402
```

Q2:

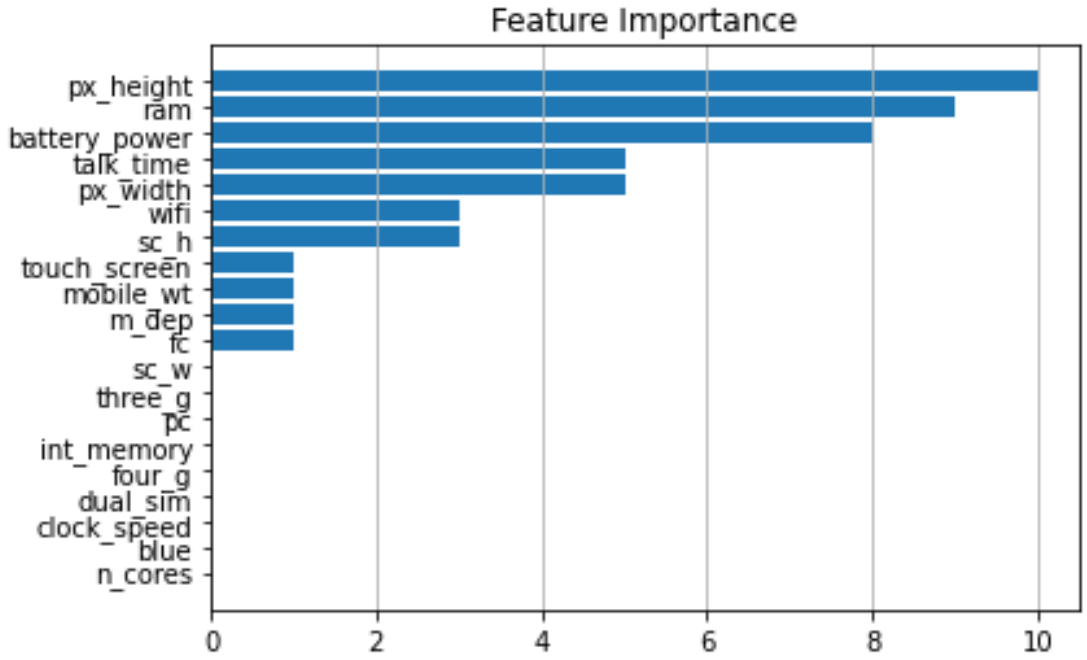
2.1

```
criterion=gini
max_depth=3: 0.9166666666666666
max_depth=10: 0.94
```

2.2

```
criterion=gini: 0.9166666666666666
criterion=entropy: 0.93
```

Q3:



Q4:

4.1

```
n_estimators=10: 0.94
n_estimators=100: 0.9733333333333334
```

Q5:

5.1

```
n_estimators=10: 0.91
n_estimators=100: 0.9466666666666667
```

5.2

```
max_features=sqrt(n_features): 0.9333333333333333
max_features=n_features: 0.9533333333333334
```

Part. 2, Questions (30%):

Q1:

因為 decision tree 會想辦法將所有 training set 的 sample 都分成功，所以有可能會為了將一兩個特定的 sample 分出來，而導致問一些多餘的問題 (一直切一直切)。

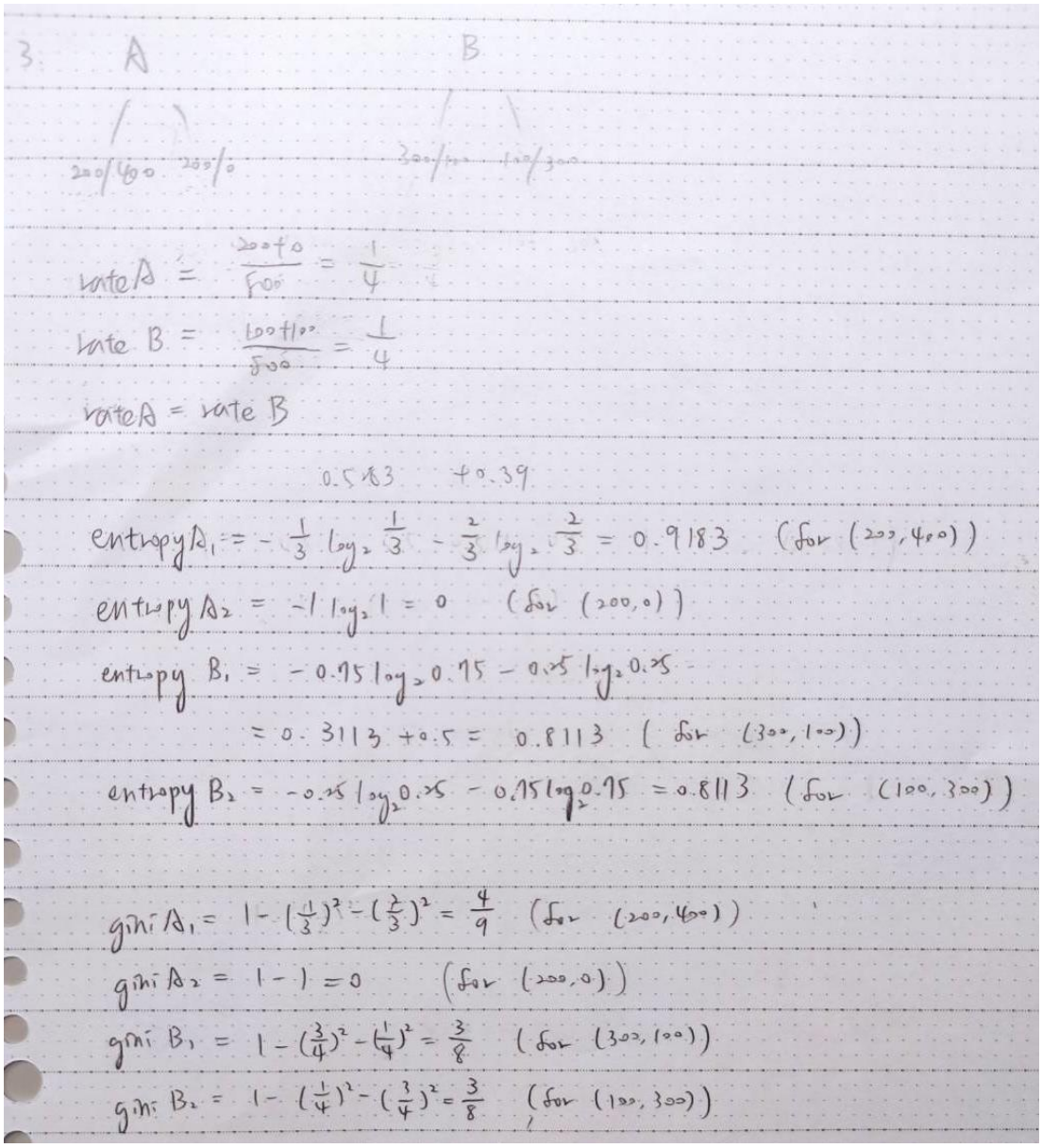
不一定，看 training set 的情況，如果有某兩個 sample 的各種 feature 都一樣，但他們卻屬於不同的 class 的話，就可能會分不出來，降低 accuracy。

Determine max depth of decision tree, cross-validation, use random forest

Q2:

a. True, follows from the update equation.
b. True. During boosting iterations, the weak classifiers try to classify more difficult samples. The weights will increase for samples that are repeatedly misclassified by the weak classifiers. The weighted training error will thus to increase.
c. False, if the data in the training set cannot be separated by a linear combination of the specific type of weak classifiers. No matter how many iterations are performed, there will still exist training error.

Q3:


3.
A
B
200/400
200/0
300/100
100/300
rate A = 200+0 / 600 = 1/4
rate B = 100+100 / 600 = 1/4
rate A = rate B
0.583
+0.39
entropy A1 = -1/3 log2 1/3 - 2/3 log2 2/3 = 0.9183 (for (200, 400))
entropy A2 = -1 log2 1 = 0 (for (200, 0))
entropy B1 = -0.75 log2 0.75 - 0.25 log2 0.25
= 0.3113 + 0.5 = 0.8113 (for (300, 100))
entropy B2 = -0.25 log2 0.25 - 0.75 log2 0.75 = 0.8113 (for (100, 300))
gini A1 = 1 - (1/3)^2 - (2/3)^2 = 4/9 (for (200, 400))
gini A2 = 1 - 1 = 0 (for (200, 0))
gini B1 = 1 - (3/4)^2 - (1/4)^2 = 3/8 (for (300, 100))
gini B2 = 1 - (1/4)^2 - (3/4)^2 = 3/8 (for (100, 300))