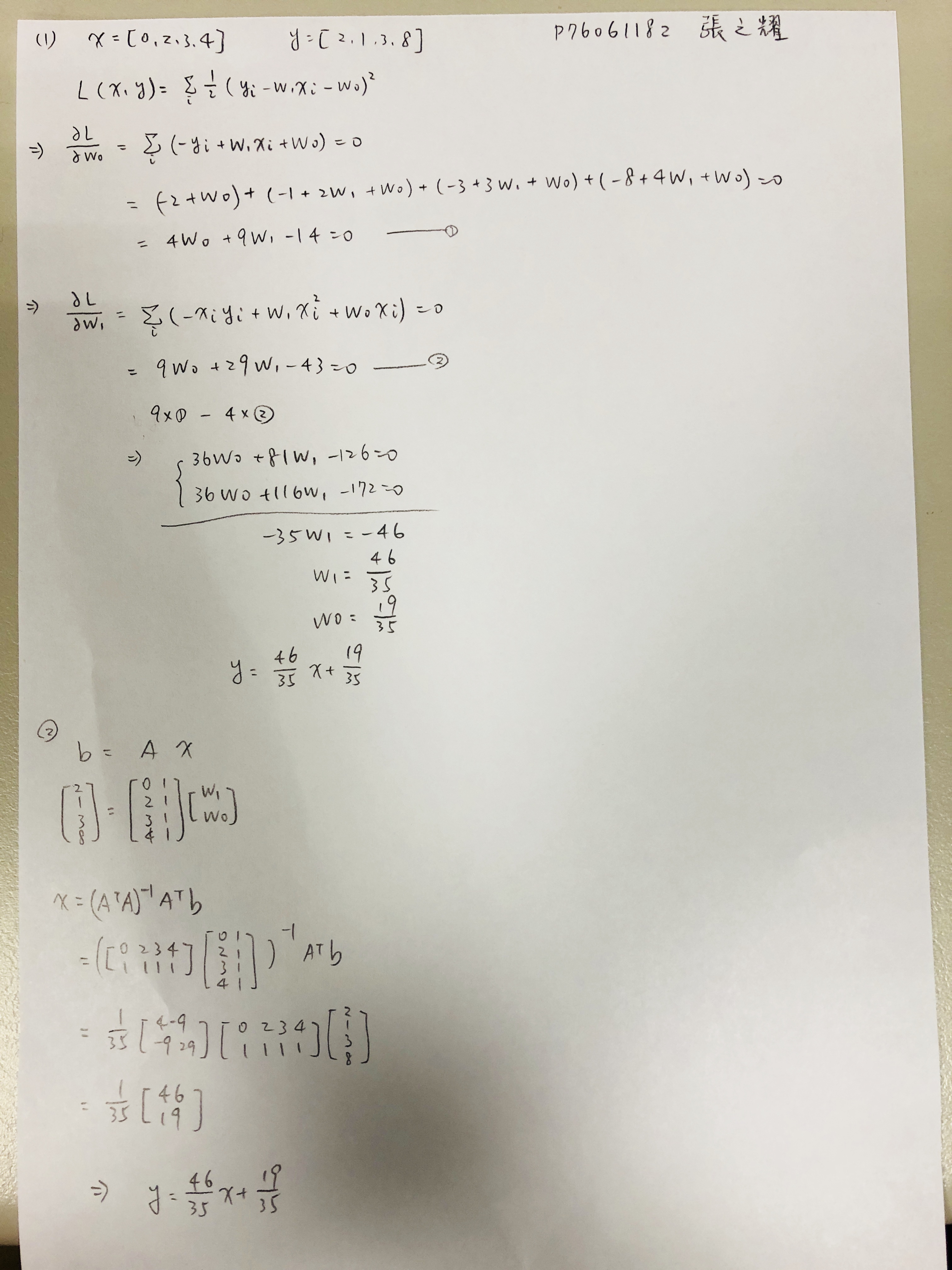
Homework 1

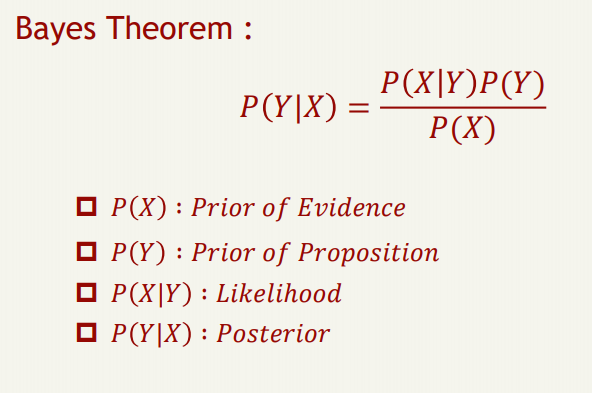
P76061182 張之耀

1.



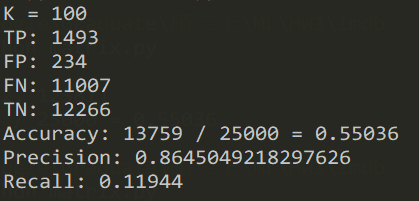
2.

用IMDB dataset來練習 Naïve Bayes Classifier

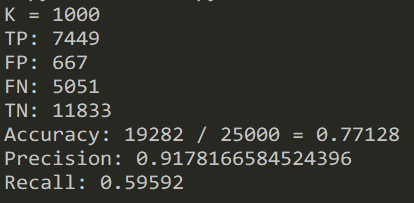


藉由這個理論來完成判斷一段文字是正面還是負面。

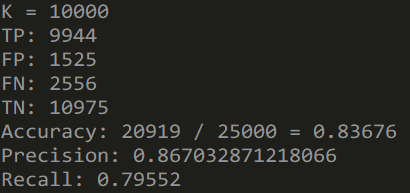
K=100的情況下：



K=1000的情況下：



K=10000的情況下：



從這邊可以發現隨著尋找top K frequency 愈多的強況下，準確率逐漸提升。

但由於我們是假設每個word發生機率互不影響(互相獨立)，所以才使用

Naïve Bayes 來處理。

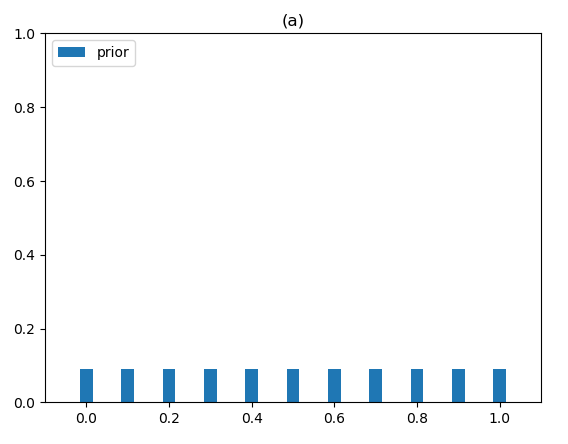
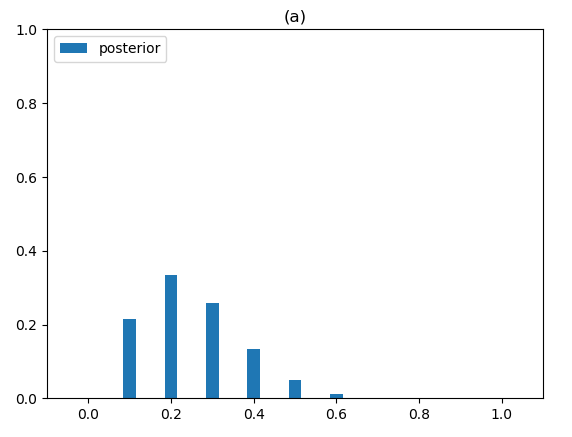
此外中間有一個步驟是用很多的p(機率)相乘，由於每段文字，字數很多，全部乘一起會變成很小很小的數不方便計算，所以我們用Log把相乘的p變成相加

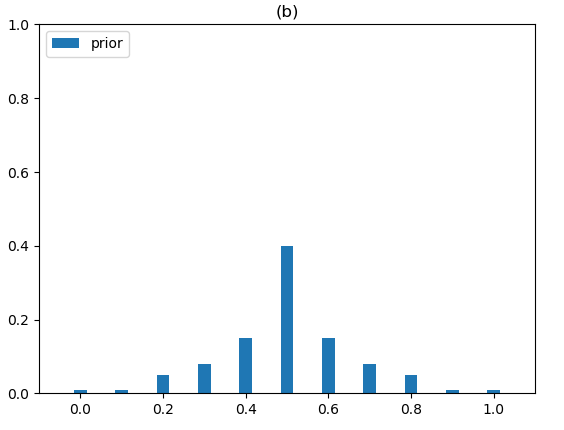
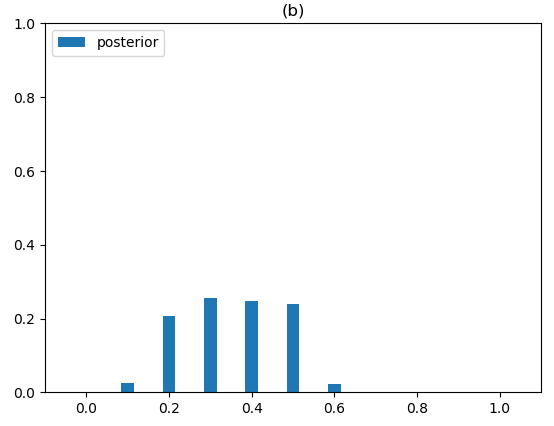
我們也不需要去計算Bayes Theorem的分母，因為不管是要計算Positive還是Negative，分母都一樣，所以只需要比分子大小即可。

3. Bayesian Estimation

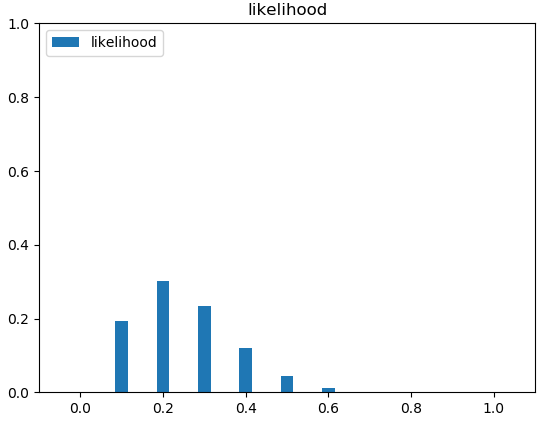
(1).

(a)的 Prior、Posterior



(b)的Prior、Posterior

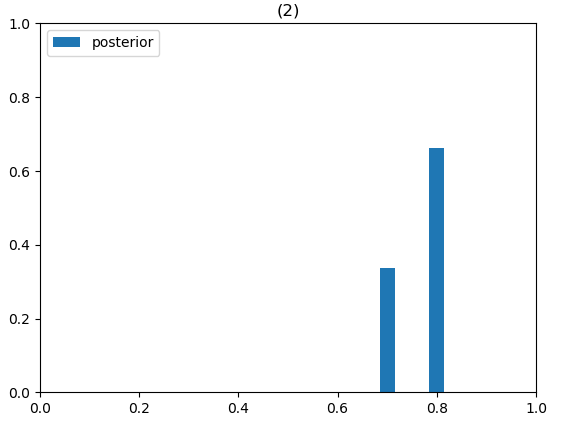
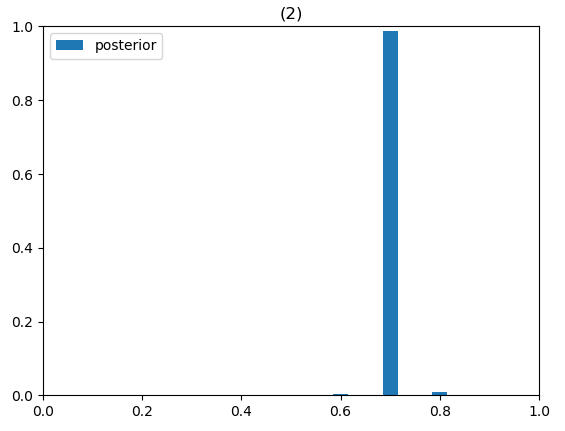
至於Likelihood都是一樣的，因為Likelihood都是看結果去估計得到的

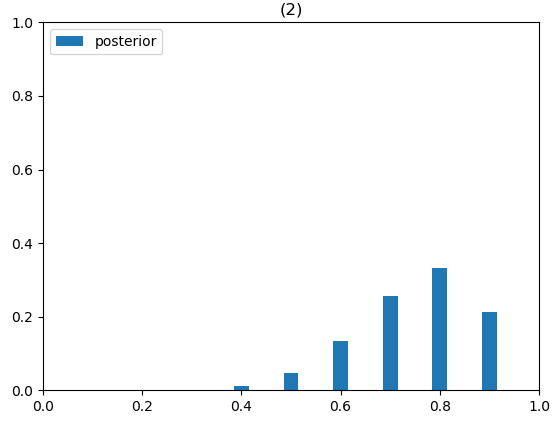
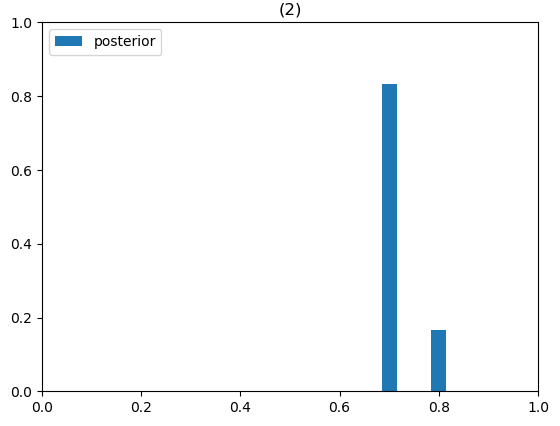


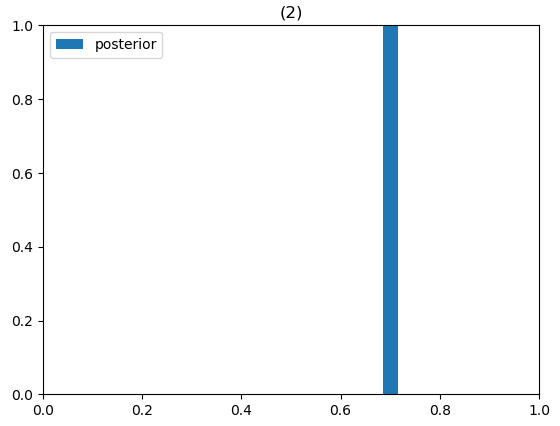
(2)

我用(a)當作prior，去擲50次硬幣，然後我是使用np.random.choice([0, 1], 10, p=[0.3, 0.7])

每次都隨機擲一次，並且把計算出來的posterior當成下次的prior，跑了50次之後MAP會收斂到一個明顯的機率，那個機率(0.7)，以下是每10個step記錄下來的結果。







這張是每個step都去算一次entropy，可以發現有慢慢變小的趨勢，很顯然entropy是用來代表一個訊息量，剛開始因為存在著很大的不確定，所以攜帶的訊息量(entropy很大)，但隨著MAP一直更新prior，也就愈來愈能確定這個模型了(穩定在p=0.7)，所以不確定性變小，entropy也就變小了。

