Entropy是從府節點到子節點減少，如果為0代表完全相同，所以有資料都在相同的類別中。

決策樹準則5分鐘前要再看一遍

所以決策樹的預測會是由水平線與垂直線所組成，

看的出來在訓練集的時候，模型沒有漏掉任何一個用戶在判斷是否購買，即使在綠色中間的一個紅點，依樣有切出來，但這就需要注意了，因為這樣的結果看起來是overfitting了，因為他將每個一個用戶都分在對的分類上，但這樣在新的資料進來的時候，沒辦法保證依樣可以分在對的分類上，

在測試集上看起來表現也還不算太差，但可以發現開始有錯誤的分類，

而評估和改善會在最後一個章節去描述

貝氏， 兩台機器生產板手，有一天把兩台機器的板手混成一堆，要找出裡面有缺陷的板手

機器1一個小時生產30支板手，機器2生產20支，在全部的產品中，不良率為1%，而不良品有一半來自機器1，一半來自機器2。而要求解的問題是從機器2生產的板手有多大的機率故障?4分

P(Mach1)=30/50=0.6

P(Mach2)=20/50=0.4

P(Defect)=0.01

P(Mach1| Defect)=0.5

P(Mach2| Defect)=0.5

要求解p(Defect|Mach2)=?

p(Defect|Mach2)= (P(Mach2| Defect)\* P(Defect))/ P(Mach2)=0.5\*0.01/0.4=0.0125

接著作者用了實際的數字帶

假設生產了1000支板手，來自機器2的就是1000\*0.4=400

而不良品為1%就是1000\*0.01=10

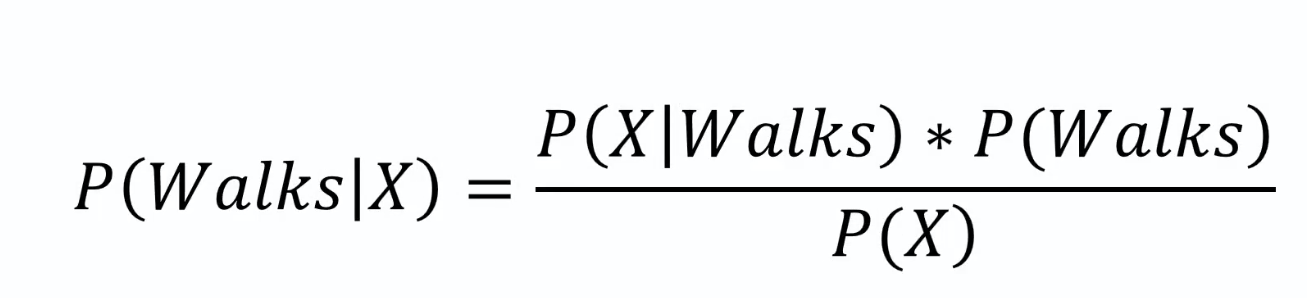
不良品中來自機器2的是50%就是10\*0.5=5

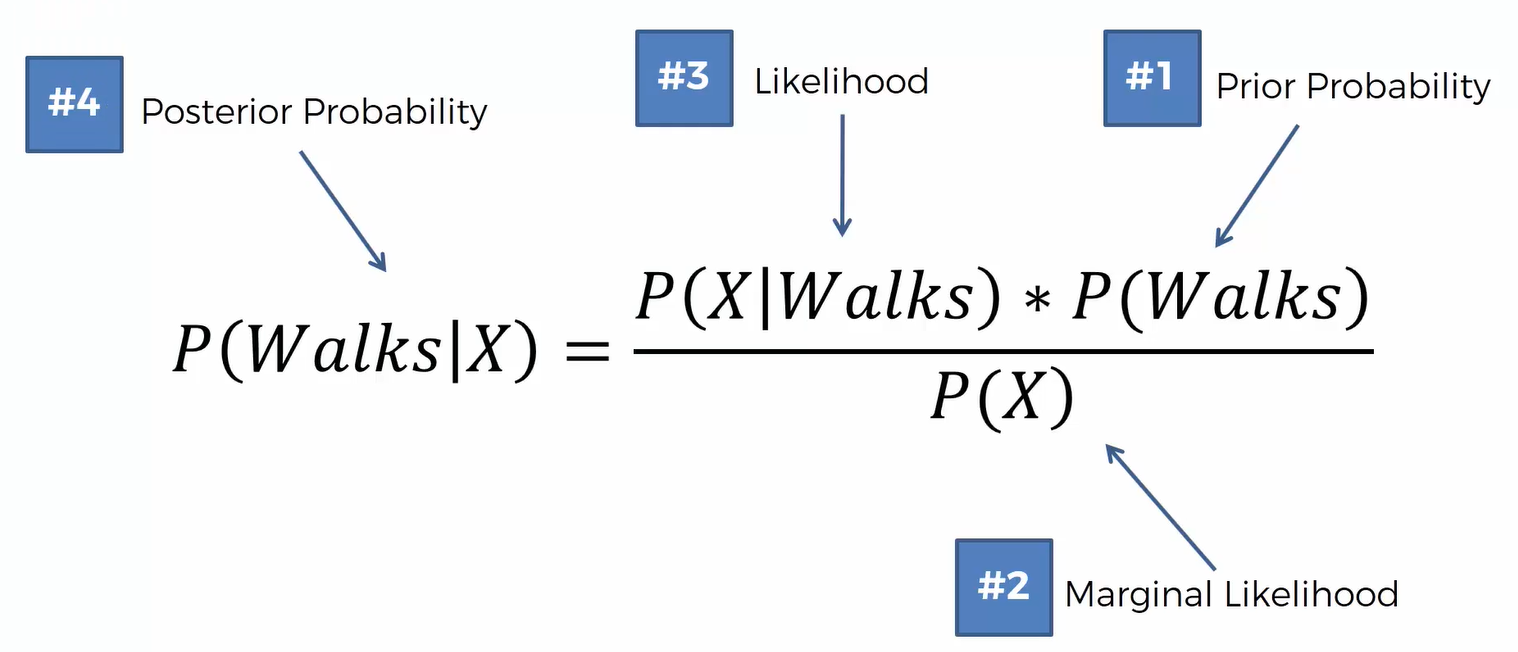
所以機器2生產的不良率為5/400=0.0125

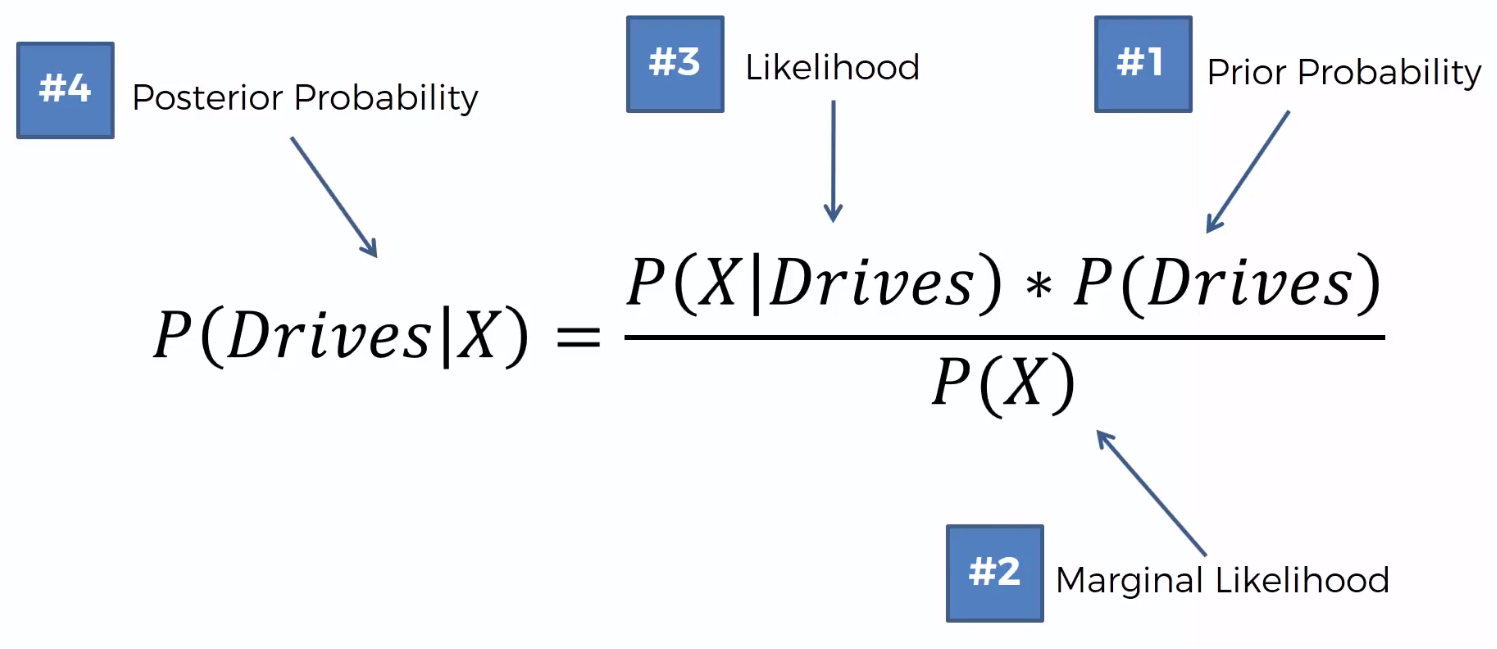
其實和上面的例子是一樣的

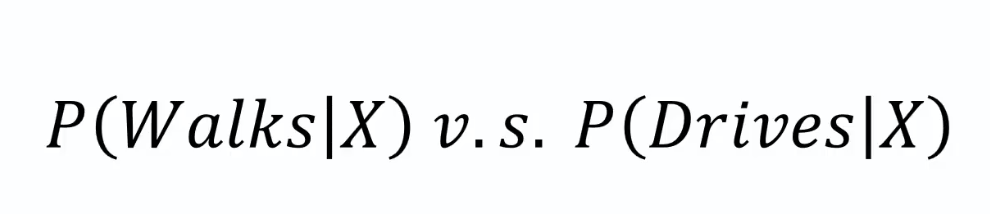
(P(Mach2| Defect)\* P(Defect))\*1000/ P(Mach2) \*1000

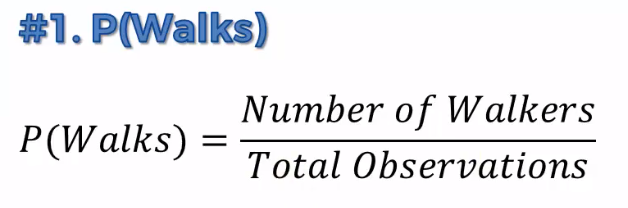
所以作者提了一個問題，為什麼不直接去計算來自機器2有缺陷的板手去除以機器2：第一，耗時，第二，會有破壞性的實驗，不可能一直去收集全部的資料。



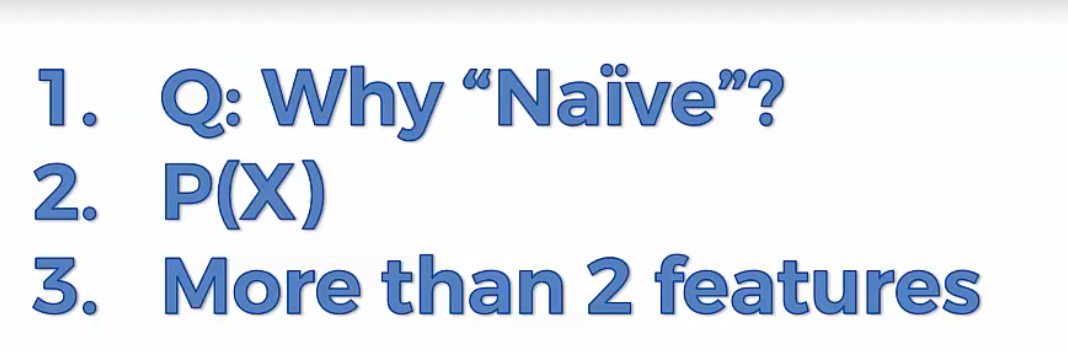






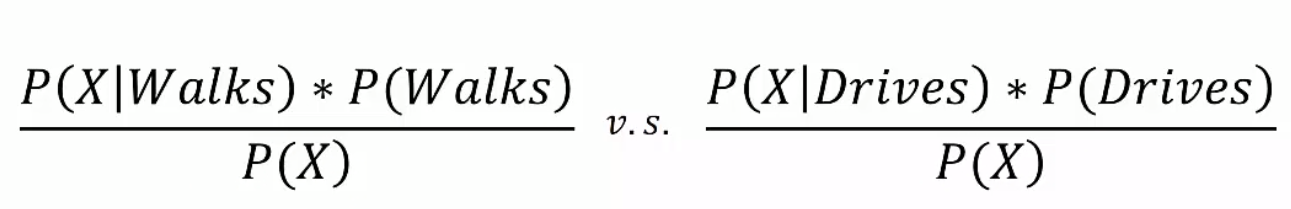
STEP1先計算

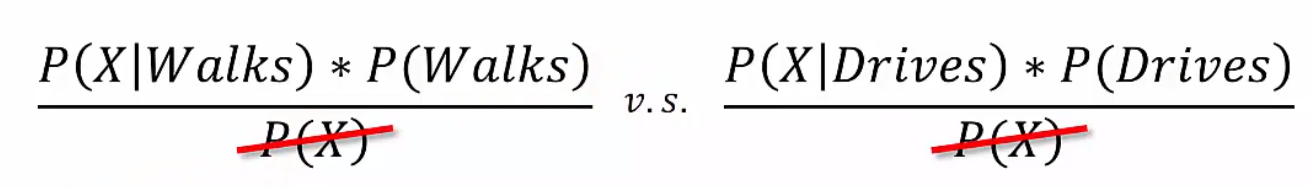
將所有的已知資料去計算P(Walks)的機率，但不將灰色的點納入計算



1. 貝氏定理要求獨立性假設，而貝氏定理是naïve bayes的基礎，但獨立性假設在一般實際的資料似乎無法滿足，所以稱之為一種單純的假設會比較貼切。作者以前面上班族的例子來說明：利用估計年薪與年齡對是走路還是開車去做分類，是假設估計年薪和年齡是獨立的，但實際上，年紀較大的人可能會因為經驗而獲得較高的年薪，是有一些相關性在的，因為不獨立，所以立論上應該不能使用貝氏定理，這就是被稱為單純貝氏的原因，因為即使變量或特徵不是獨立或是不完全獨立時，它也被應用並且有不錯的結果。

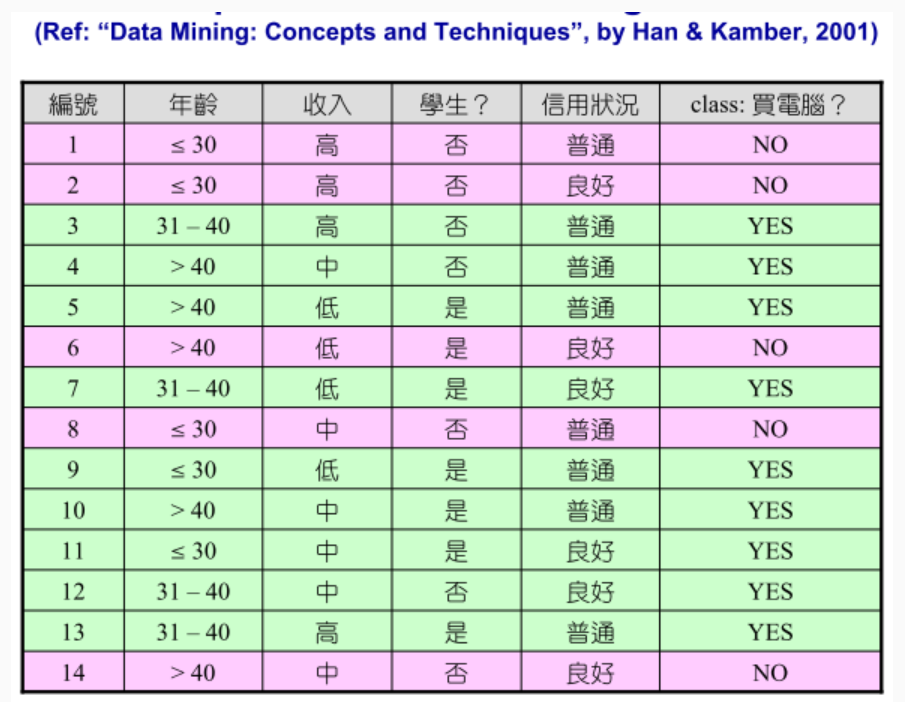
補充：Naive Bayes Classifiers (單純貝氏分類器)此分類器模型會給問題實例分配用特徵值表示的類標籤，類標籤取自有限集合。它不是訓練這種分類器的單一算法，而是一系列基於相同原理的算法：所有單純貝氏分類器都假定樣本每個特徵與其他特徵都獨立。

1. 在前面的步驟2中，p(X)是在欲分類的點為中心，畫一個圈，拿掉要預測的點後，而來自此資料集的隨機選擇點的將會與我們即將添加的數據點顯示相似的特徵，而通常在不同的分類計算上P(X)的數值都是一樣的。



所以當你在比較兩邊的大小的時候，可以做一個簡化的動作，但如果是想要算出完整的機率值，就不能做省略的步驟。





而如果超過兩個分類以上的話，因為如果是兩個分類，計算出第一個分類的機率，就可以自動得知另一個分類的機率，因為相加會等於1

做這個練習嘗試了解naïve bayes

