學號：B07902040 系級：資工二 姓名：吳承軒

1. (2%) 請比較實作的 generative model 及 logistic regression 的準確率，何者較佳？請解釋為何有這種情況？

   logistic regression 的準確率較佳，因為 generative model 對 probability distribution 做了假設，而 logistic regression 屬於 Discriminative model 的一種，是沒有做預先假設的。generative model 預做假設的優勢在於，當資料量較小時，預設的 distribution 能在較小的資料中得到較好的結果；當雜音較多時，預設的 distribution 也能避免這些雜音影響 model 太多。而我們的資料量夠大，資料的雜音也少，相對來說就是對 Discriminative model 有優勢。

2. (2%) 請實作 logistic regression 的正規化 (regularization)，並討論其對於你的模型準確率的影響。接著嘗試對正規項使用不同的權重 (lambda)，並討論其影響。(有官 regularization 請參考 https://goo.gl/SSWGhf p.35)

   正規化會懲罰過大的 W 項，使得 W 各項值的大小被控制，避免出現 overfitting 的現象。

   而我的 model 本身複雜度並不高，overfitting 的現象在迭代次數未超出合理範圍太多時並不明顯，也因此運用正規化之後，對於正確率僅有小幅增進。

   隨著 lambda 變大，weight 改動的速度會變慢，需要更多次迭代才能達到底部；當 lambda 過小時，接近於未用正規化前的結果，由於上段原因，到底部的 error rate 會小幅上升。

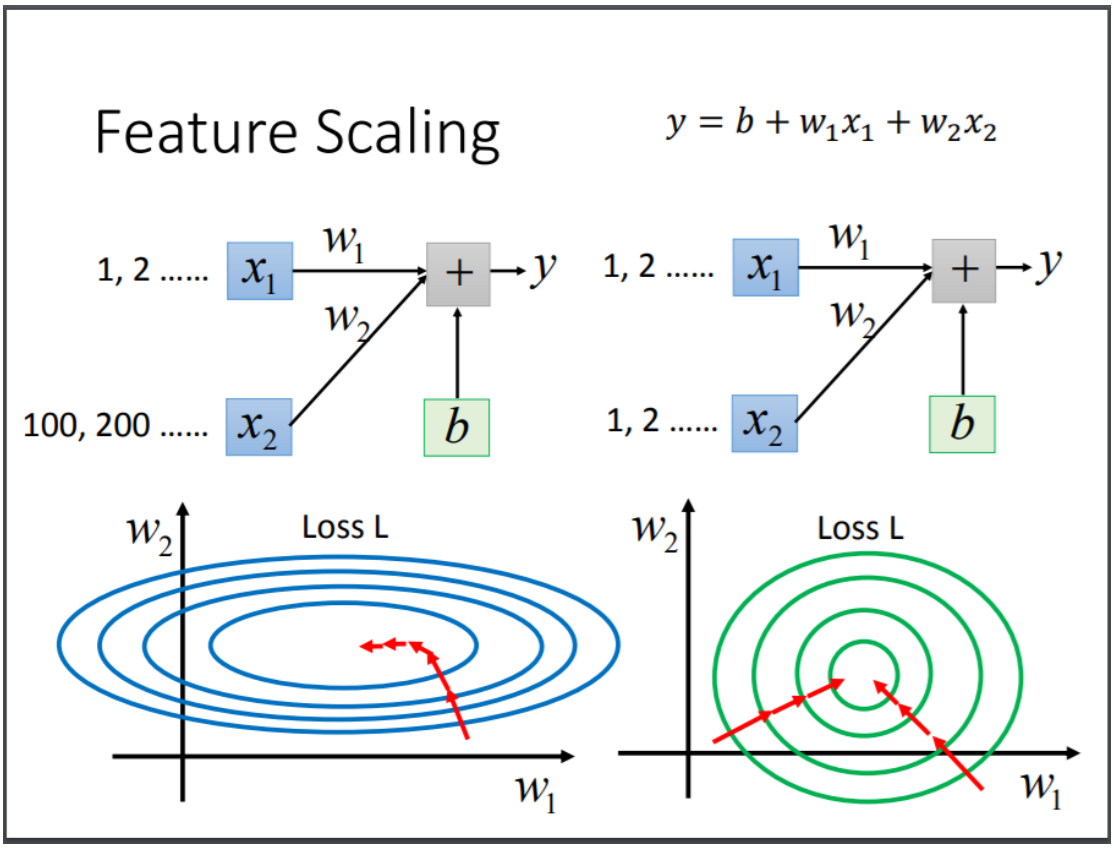
3. (1%) 請說明你實作的 best model，其訓練方式和準確率為何？

   將 not in universe、not in universe 和 do not know 的 feature 刪除，使用正規化和特徵標準化改善，將 train 的過程包裝成函式，train 多次，每次 train 之前先 shuffle，再切分原資料為 train data 和 validation data，回傳產生的 weight 在 validation 上的正確率，最後選擇正確率最高那次的 weight 作為結果。

4. (1%) 請實作輸入特徵標準化 (feature normalization)，並比較是否應用此技巧，會對於你的模型有何影響。

應用之後有幾個優點:

1.透過迭代降低 error rate 的速度更快，將原先的橢圓形變成了圓形:



(圖片來源)http://speech.ee.ntu.edu.tw/~tlkagk/courses/ML_2016/Lecture/Gradient%20Descent%20(v2).pdf)

2.避免了原資料中，平均值較大的數字對結果影響較大的問題。

3.將 weight 可視化，標準化之後，直接看 weight 各項的絕對值即可知道各個 feature 對結果的影響力。