Machine Learning Foundations

Homework #1

R08922115 吳冠霖

Problem 1.

Problem 2.

我認為 semi-supervised learning 的方式非常適合用來做樂器聲音的識別。因為相同樂器的音色同質性很高，只需要做少量的 label 就可以用 clustering 的方式將各樂器區分開來，若是音色較接近的樂器也會因 label 的輔助而辨識的更好。而我覺得識別樂器有個很棒的應用是將混好的音樂中的各軌重新分開來，可以做出很多更棒的創作。

Problem 3.

由題目敘述，可以得知 testing set 的元素數量為 L 個，每個結果有+1 或-1 兩種可能，因此 target f 共有2^L種。我們可以推導出每個 hypothesis g的 OTS 期望值為「每種f出現的機率×該種f的錯誤量」。

假設該種f出現的機率為P_f，則：

$$\mathbb{E}_f\{E_{OTS}(\mathcal{A}(\mathcal{D}), f)\} = P_f \sum_{i=0}^{L} i \binom{L}{i}$$

另外，由於演算法求出的任何 hypothesis g都有對應的 target f使g答錯的數量為 0、1、2、……、L 個，而題目又有說這些f出現的機率相等，亦即上式中的P_f為定值，因此所有g的 OTS 期望值都為定值。

$$\Rightarrow \mathbb{E}_f\{E_{OTS}(\mathcal{A}(\mathcal{D}), f)\} = constant$$

Problem 4.

在 A、B、C、D 四種骰子中，只有 A 和 D 兩種含有「green 1」。因此符合條件的情況是五顆骰子為{A、D}的任意組合。假設四種骰子的數量相等，則可將抽中每種骰子的機率視為 1/4。於是抽五次的組合就可寫成下式以求出解答：

$$\sum_{i=0}^{5}\binom{5}{i}\left(\frac{1}{4}\right)^{i}\left(\frac{1}{4}\right)^{5-i}=\frac{1}{32}$$

Problem 5.

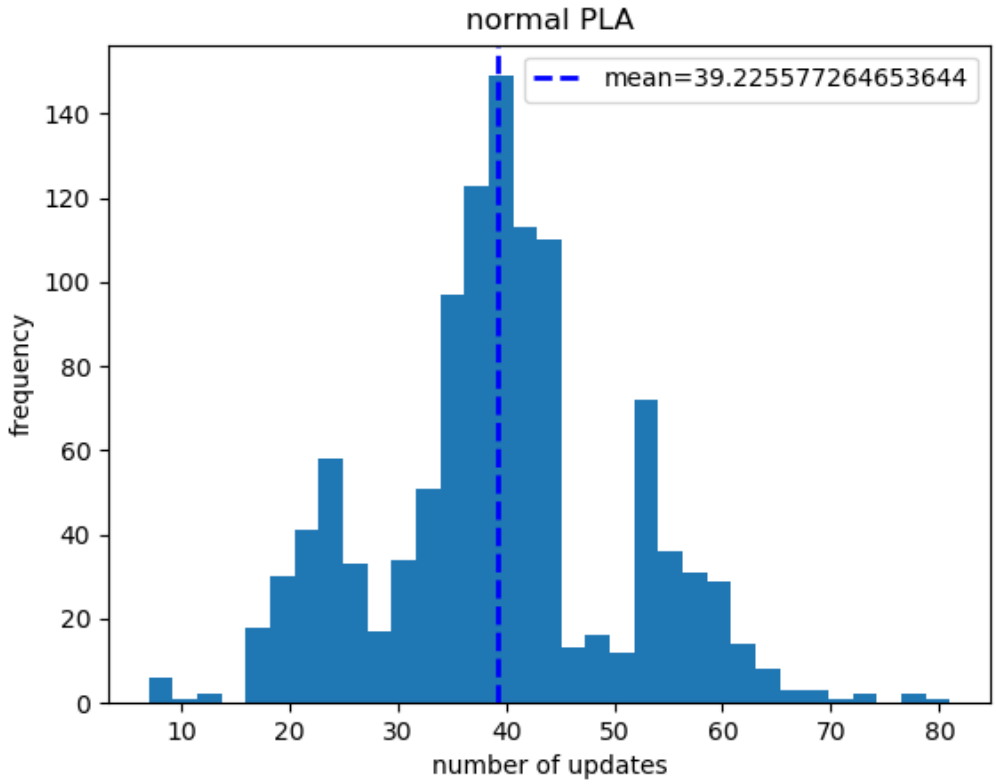
題目的要求為抽中五顆骰子中必須有某個數字在五顆骰子中的顏色都相同。觀察 A、B、C、D 的特性，可以發現{A、B}和{C、D}各為互斥集合，因此符合條件的情況為{A、C}、{A、D}、{B、C}、{B、D}四種集合的任意組合。各集合的任意組合數算法與上題相同，答案都為 1/32。然而最主要的差異在於**集合間有交集的情形發生，交集發生在抽中的五顆骰子全為相同種類的時候。**五顆骰子皆為同類型的機率可以用下式表達：

$$\left(\frac{1}{4}\right)^{5}\times 4=\frac{1}{256}$$

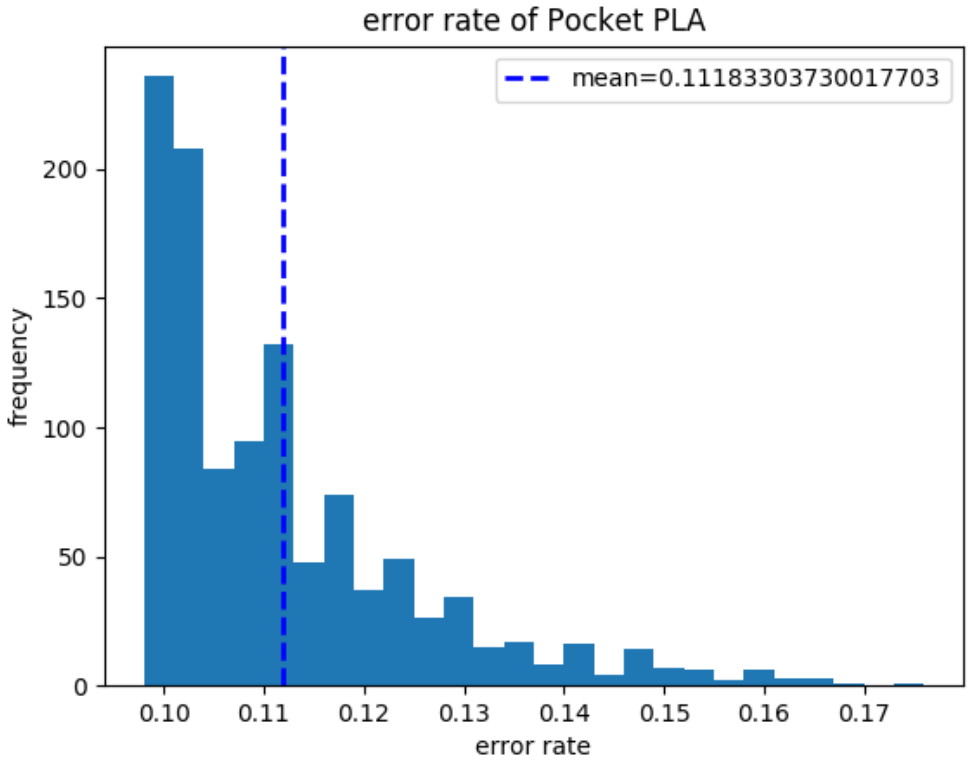
因此最終機率則為：

$$\frac{1}{32}\times 4-\frac{1}{256}=\frac{31}{256}$$

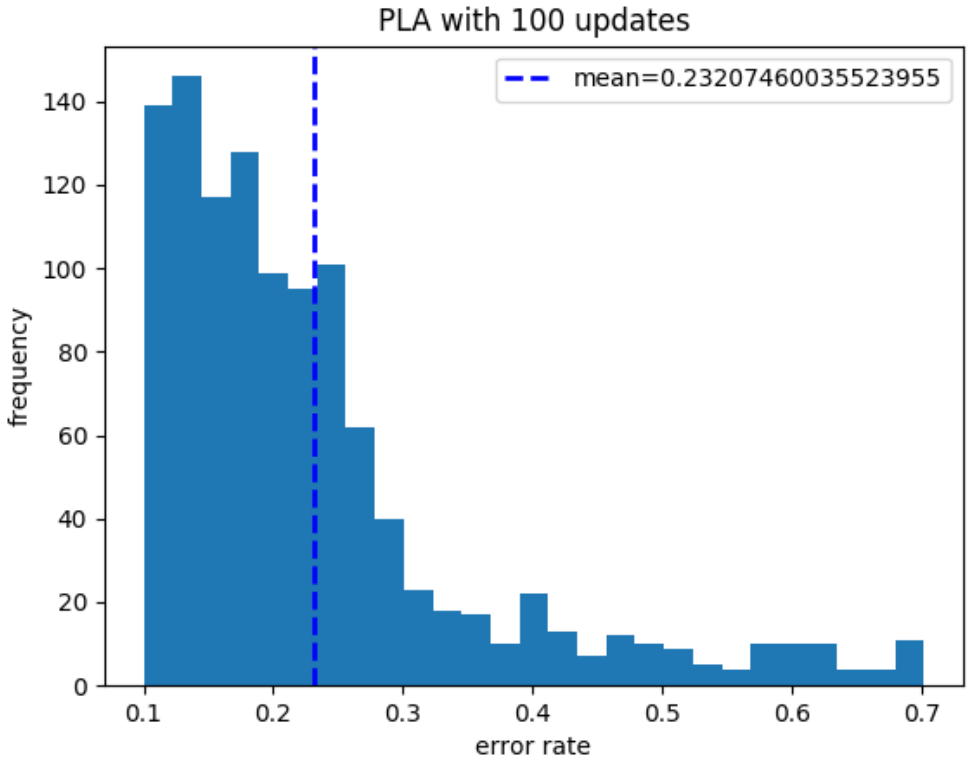
Problem 6.



Problem 7.



Problem 8.



可以觀察到兩種方法的機率分佈外觀很類似，像是伸縮平移後的結果。而在結果的表現上，有使用 pocket 的結果非常穩定，不會有超過 0.2 的機率出現，而不使用 pocket 則常常會產生比什麼事都不做還糟糕的結果（若使用 w 的初始值 0 直接測試，還有正確率 0.4 左右的表現）。

關鍵在於 pocket **每次更新時會檢查資料整體的表現**，每次更新都保證會讓結果更好。而若只考慮單一筆資料其實是類似 greedy 的作法，無法顧全大局，若是為了偏差較遠的值進行校正，很可能會導致整體的結果一落千丈。不過，由於 pocket 演算法每次都要檢查所有資料的緣故，執行時間遠遠大於後者，若要實際拿來使用必須要考慮到成果與執行時間的效益，並沒有絕對較好的選擇。

Problem 9(bonus).

我認為這個做法不會有作用，甚至可能造成反效果。首先，PLA 演算法的時間複雜度是取決於 data 的數量與其分佈，而不是數值本身。而且將每個 x 除以 10 可能會導致每個點距離更接近，讓他們的關係更加模糊且難以區分，也許透過其他數值方法將資料點彼此的距離增加會是個有效果的辦法，但題目敘述的不是。