**ML Foundation HW1**

**1. D**

因為這種問題的確有一些pattern，也不好定義怎麼樣的芒果是好的。如果有足夠資料那這種問題有可能可以使用ML。  
因為給的是圖片檔，最直接的方式是把所有pixel直接丟給模型訓練；但是我們也可用一些可能相關的concrete data給模型，像是芒果的顏色鮮豔度、大小等等的參數以某個量表顯示之後跟著丟給模型等等。

**2. E**

( A ) 這完全是機率性的做法，沒有用到資料或潛藏pattern。  
( B ) 這並沒有電腦輔助計算潛藏pattern，比較像是蒐集資料時會做的是。  
( C ) 這已經完美定義好了。已經輕鬆定義解法的問題並不是ML的範疇。  
( D ) 他並沒有學習的成分在內。比較精確地說，他沒有在Learning problem模型上講的Learning Algorithm及Hypothesis set。他並不是從hypothesis set中取一個更好的函數，而是已經先行定義好了一個函數(像是要多於幾個才會超標、超過幾percent才會算進列表等等的參數)。機器學習應該會試著優化這些參數以獲得更接近理想函數的函數。

**3. D**

基本上就如( 4 )所說，我們可以把這題的敘述換成：「不要動xnxn，但是我們把更新規則換成：  
wt+1←wt+14ynxnwt+1←wt+14ynxn  
」。要解釋的話可以說那個1212乘到xnxn上還是整個ynxnynxn上還是一樣的式子。  
如此一來，只需要重導一次公式即可。就如同教授隨信附贈的證明講義，下面只陳述大綱，並且用下括號定義一些符號：  
wt+1←wt+14yn(t)xn(t),w0=0⇒⎧⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪⎨⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪⎩wTfwt+1=wTf(wt+14yn(t)xn(t)))=wTfwt+14wTfyn(t)xn(t)≥wTfwt+14⋅minn(ynwTfxn)∥wf∥⋅ρ′∥wt+1∥2=∥wt+14yn(t)xn(t)∥2=∥wt∥2+2⋅14yn(t)wTtxn(t)+116⋅∥yn(t)xn(t)∥2≤∥wt∥2+116⋅∥yn(t)xn(t)∥2≤∥wt∥2+116⋅maxn(∥ynxn∥2)=∥wt∥2+116⋅maxn(∥xn∥2)(R′)2wt+1←wt+14yn(t)xn(t),w0=0⇒{wfTwt+1=wfT(wt+14yn(t)xn(t)))=wfTwt+14wfTyn(t)xn(t)≥wfTwt+14⋅minn(ynwfTxn)⏟‖wf‖⋅ρ′‖wt+1‖2=‖wt+14yn(t)xn(t)‖2=‖wt‖2+2⋅14yn(t)wtTxn(t)+116⋅‖yn(t)xn(t)‖2≤‖wt‖2+116⋅‖yn(t)xn(t)‖2≤‖wt‖2+116⋅maxn(‖ynxn‖2)=‖wt‖2+116⋅maxn(‖xn‖2)⏟(R′)2  
接下來我們跟投影片一樣定義：  
⎧⎪

⎪

⎪⎨⎪

⎪

⎪⎩ρ=minn(ynwTf∥wf∥x)R2=maxn(∥xn∥2){ρ=minn(ynwfT‖wf‖x)R2=maxn(‖xn‖2)  
從上面的式子，可知：  
⎧⎪

⎪⎨⎪

⎪⎩ρ′=14ρ(R′)2=116R2⇒R′=14R{ρ′=14ρ(R′)2=116R2⇒R′=14R  
而現在，我們也有：  
{wTfwt+1≥wTfwt+∥wf∥⋅ρ′∥wt+1∥2≤∥wt∥2+(R′)2{wfTwt+1≥wfTwt+‖wf‖⋅ρ′‖wt+1‖2≤‖wt‖2+(R′)2  
我們就可以完全照證明講義的講法證明出在投影片16頁中所謂的constant了。因為打字關係我把導出的式子拿掉了。基本上有點類似歸納法。得出：  
w0=0⇒{wTfwT≥∥wf∥⋅ρ′⋅T∥wT∥2≤(R′)2⋅T⇒1≥cosθT=wTfwT∥wf∥∥wT∥≥√T⋅ρ′R′w0=0⇒{wfTwT≥‖wf‖⋅ρ′⋅T‖wT‖2≤(R′)2⋅T⇒1≥cos⁡θT=wfTwT‖wf‖‖wT‖≥T⋅ρ′R′  
所以我們可以得知1≥√T⋅ρ′R′1≥T⋅ρ′R′，所以T≤(R′)2(ρ′)2T≤(R′)2(ρ′)2，因為TT是學習次數/時間，所以學習速度應該算是(ρ′)2(R′)2(ρ′)2(R′)2。對比原本的學習速度ρ2R2ρ2R2：  
(ρ)2/(R′)2ρ2/R2=116/1161=1(ρ)2/(R′)2ρ2/R2=116/1161=1  
所以學習速度不變。

**4. C**

基本上就是第三題重導一遍：  
wt+1←wt+yn(t)xn(t)∥xn(t)∥⇒⎧⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪⎨⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪

⎪⎩wTfwt+1=wTf(wt+yn(t)xn(t)∥xn(t)∥)=wTfwt+wTfyn(t)xn(t)∥xn(t)∥≥wTfwt+minn(ynwTfxn∥xn∥)∥wf∥⋅^ρ∥wt+1∥2=∥wt+yn(t)xn(t)∥xn(t)∥∥2=∥wt∥2+2⋅yn(t)wTtxn(t)∥xn(t)∥+∥yn(t)xn(t)∥xn(t)∥∥2≤∥wt∥2+∥yn(t)xn(t)∥xn(t)∥∥2=∥wt∥2+1wt+1←wt+yn(t)xn(t)‖xn(t)‖⇒{wfTwt+1=wfT(wt+yn(t)xn(t)‖xn(t)‖)=wfTwt+wfTyn(t)xn(t)‖xn(t)‖≥wfTwt+minn(ynwfTxn‖xn‖)⏟‖wf‖⋅ρ^‖wt+1‖2=‖wt+yn(t)xn(t)‖xn(t)‖‖2=‖wt‖2+2⋅yn(t)wtTxn(t)‖xn(t)‖+‖yn(t)xn(t)‖xn(t)‖‖2≤‖wt‖2+‖yn(t)xn(t)‖xn(t)‖‖2=‖wt‖2+1  
特別需要注意的是：因為|yn|=1|yn|=1且ynwTfxnynwfTxn恆正，代表：  
minn(ynwTfxn∥xn∥)=minn(|wTfxn|∥xn∥)=∥wf∥⋅^ρminn(ynwfTxn‖xn‖)=minn(|wfTxn|‖xn‖)=‖wf‖⋅ρ^  
且∥yn(t)xn(t)∥xn(t)∥∥2‖yn(t)xn(t)‖xn(t)‖‖2因為|yn|=1|yn|=1且xnxn被normalize，所以為一。  
一樣根據證明講義的證法列出：  
{wTfwT≥∥wf∥⋅^ρ⋅T∥wT∥2≤T⇒1≥cosθT=wTfwT∥wf∥∥wT∥≥√T⋅^ρ{wfTwT≥‖wf‖⋅ρ^⋅T‖wT‖2≤T⇒1≥cos⁡θT=wfTwT‖wf‖‖wT‖≥T⋅ρ^  
可以導出T≤1^ρ2T≤1ρ^2。

**5. D**

首先，我們知道⌊x+1⌋>x⌊x+1⌋>x。需要注意的是整數點位置可能有例外，不過在這裡沒有例外，所以寫嚴格大於是正確的。  
我們就拿上一題的式子：wt+1←wt+η⋅yn(t)xn(t)wt+1←wt+η⋅yn(t)xn(t)。  
所以：  
ynwTt+1xn=yn(wt+η⋅ynxn)Txn=ynwTtxn+η⋅y2n⋅xTnxn=ynwTtxn+η⋅∥xn∥2>0⇒ynwTtxn+η⋅∥xn∥2>0η>−(ynwTtxn∥xn∥2)ynwt+1Txn=yn(wt+η⋅ynxn)Txn=ynwtTxn+η⋅yn2⋅xnTxn=ynwtTxn+η⋅‖xn‖2>0⇒ynwtTxn+η⋅‖xn‖2>0η>−(ynwtTxn‖xn‖2)  
根據最剛開始的不等式，選擇η=⌊−ynwTtxn∥xn∥2+1⌋η=⌊−ynwtTxn‖xn‖2+1⌋。

**6. C**

我們設w0=0w0=0，因為在此題最後會證明，如果這樣會收斂，那對任意w0w0都收斂。  
( a )( b ) 會收斂。我們剛剛在第四題證明wt+1←wt+12ynxnwt+1←wt+12ynxn的時候，最後的不等式：  
1≥cosθT=wTfwT∥wf∥∥wT∥≥√T⋅ρ′R′1≥cos⁡θT=wfTwT‖wf‖‖wT‖≥T⋅ρ′R′  
與ρ′=14ρ,R′=14Rρ′=14ρ,R′=14R，基本上對所有正數的ηη都是差不多的結果，所已可以以類似的方法得到次數上限。  
( c ) 不會。事實上這就是第五題的式子。我們可知他會使ynwTt+1xn=0ynwt+1Txn=0，代表點(xn,yn)(xn,yn)在那條分界線上。如果我們把線上的點都算錯的話那一定至少有一點是錯的。  
( d )( e ) 請看第三題改造之後的式子：  
wt+1←wt+ηt⋅yn(t)xn(t)⇒wTfwt+1=wTf(wt+ηt⋅yn(t)xn(t))=wTfwt+ηt⋅wTfyn(t)xn(t)≥wTfwt+ηt⋅minn(ynwTfxn)∥wf∥⋅ρ′wt+1←wt+ηt⋅yn(t)xn(t)⇒wfTwt+1=wfT(wt+ηt⋅yn(t)xn(t))=wfTwt+ηt⋅wfTyn(t)xn(t)≥wfTwt+ηt⋅minn(ynwfTxn)⏟‖wf‖⋅ρ′  
可以從這裡導到後面的「在一定次數內收斂」的結論需要ηt>0ηt>0(注意ynwTfxn>0ynwfTxn>0)，不然沒有辦法使用「內積會遞增」的事實證明。  
( d )做到了這點，因為ηt=⎢⎢

⎢

⎢

⎢

⎢

⎢

⎢⎣>0−yn(t)wTtxn(t)∥xn(t)∥2+1⎥⎥

⎥

⎥

⎥

⎥

⎥

⎥⎦>0ηt=⌊−yn(t)wtTxn(t)⏞>0‖xn(t)‖2+1⌋⏟>0  
( e )剛好跟( d )異號，所以沒辦法接著導下去而得到會halt的結論。事實上，拿最後一大題的資料測試會發現wtwt的長度會一直變大，直到程式overflow…。  
所以會收斂的為(abd)。

如果w0≠0w0≠0，我們看wt+1←wt+yn(t)xn(t)wt+1←wt+yn(t)xn(t)的例子：  
∥w0∥2=∥w0∥2∥w1∥2≤∥w0∥2+R2...⇒∥wT∥2≤T⋅R2+∥w0∥2‖w0‖2=‖w0‖2‖w1‖2≤‖w0‖2+R2...⇒‖wT‖2≤T⋅R2+‖w0‖2  
也可用類似方法得知：wTfwT≥∥wf∥⋅ρ⋅T+wTfw0wfTwT≥‖wf‖⋅ρ⋅T+wfTw0。  
所以1≥cosθT=wTfwT∥wf∥∥wT∥≥∥wTf∥⋅ρ⋅T+wTfw0√T⋅R2+∥w0∥21≥cos⁡θT=wfTwT‖wf‖‖wT‖≥‖wfT‖⋅ρ⋅T+wfTw0T⋅R2+‖w0‖2。  
隨著TT增加，上面的成長速度(≈T≈T)大於下面的成長速度(≈√T≈T)，所以最後還是會到達1的。類似地，因為只有多一個常數項，如果使用其他用w0=0w0=0會收斂的ηtηt，也仍然會對其他任意的w0w0收斂。

**7. E**

因為他是用類似評分的系統來學習的，也沒有直接給定如「這個盤面的話要下哪步棋」的資料，所以比較像reinforcement learning。

**8. B**

(參考：[http://violin-tao.blogspot.com/2018/01/ml-lecture-21-structure-learning.html](http://violin-tao.blogspot.com/2018/01/ml-lecture-21-structure-learning.html" \t "_blank))  
他比較像是structured learning，因為我們的輸入是影片(或是圖片的sequence)，得出來的卻是人的動作(或車子的state)。兩個是很不同的object。  
因為只有少數資料有給定Y(也就是人的動作)，所以是semi-supervised。  
一批給定資料，所以是batch learning。  
雖然給定的東西還挺有意義的，但是他並未被處理過，比較像是raw feature。  
(老實說有一點選擇障礙，如果我能自選的話應該會選regression, semi-supervised, batch/online, concrete/raw都可以。)

**9. E**

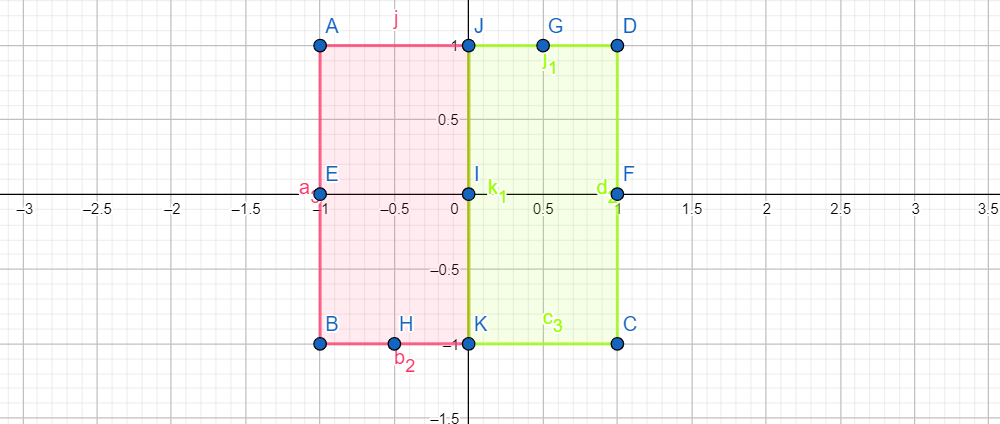
因為我們甚麼hypothesis都可能會被用到，像是：

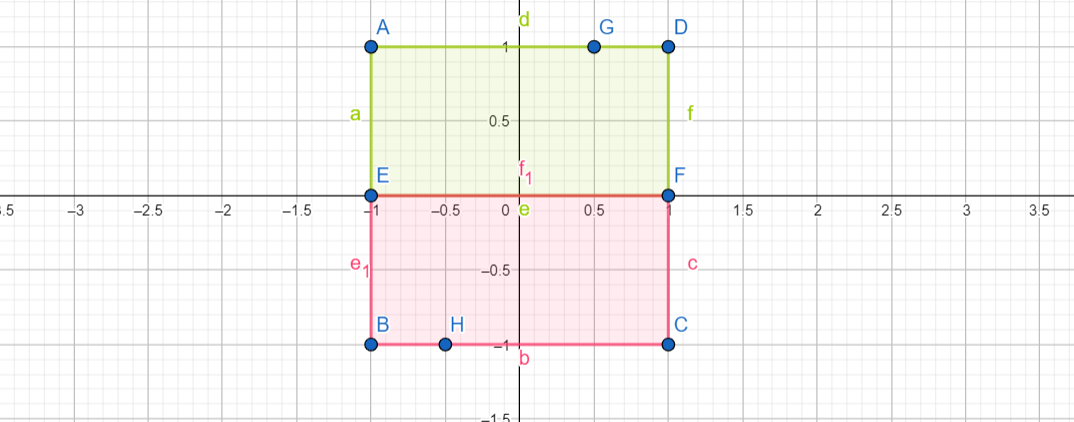
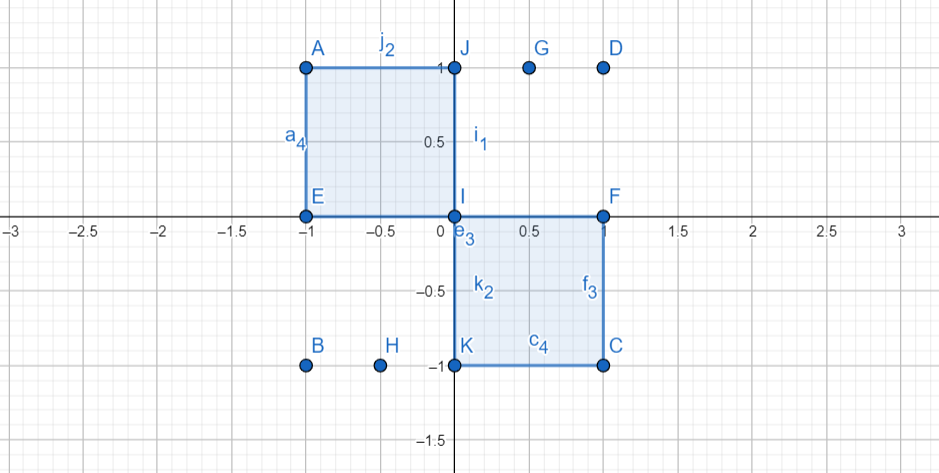
* 選前面3個，學到「只要第二格小於3就是+1，否則-1」，那Eots=1Eots=1
* 選前面3個，學到「輸出+1」，那Eots=0Eots=0

**10. B**

我們丟NN次銅板，計算出頭尾出現的機率。如果較常出現的面統計出來的機率比較高我們就說我們成功找到了。  
我們定義「比較常出現的那一面」叫做SS。  
所以SS的實際出現機率是μ=12+ϵμ=12+ϵ，假設我們在樣本空間實際統計出來的SS的機率是νν，那我們知道如果ν>12ν>12的話，我們就成功了。  
這代表μμ跟νν不能差太遠，正式地說，|μ−ν|≤ϵ|μ−ν|≤ϵ。  
所以我們失敗的機率是Pr(|μ−ν|>ϵ)≤δPr(|μ−ν|>ϵ)≤δ，又從Hoeffding’s inequality，可得知：  
Pr(|μ−ν|>ϵ)≤2exp(−2ϵ2N)Pr(|μ−ν|>ϵ)≤2exp(−2ϵ2N)  
右邊那一項是上限值，我們需要上限值小於等於δδ：  
2exp(−2ϵ2N)≤δexp(−2ϵ2N)≤δ2−2ϵ2N≤logδ2N≥1−2ϵ2logδ2=12ϵ2log2δ2exp(−2ϵ2N)≤δexp(−2ϵ2N)≤δ2−2ϵ2N≤log⁡δ2N≥1−2ϵ2log⁡δ2=12ϵ2log⁡2δ  
得到N≥12ϵ2log2δN≥12ϵ2log⁡2δ。  
其實我們只要符合Pr(μ−ν>ϵ)<δPr(μ−ν>ϵ)<δ即可，也就是即使νν比μμ大超過ϵϵ也沒差，不過會需要One-sided Hoeffding Inequality。在這裡就選擇比較不那麼tight的NN。

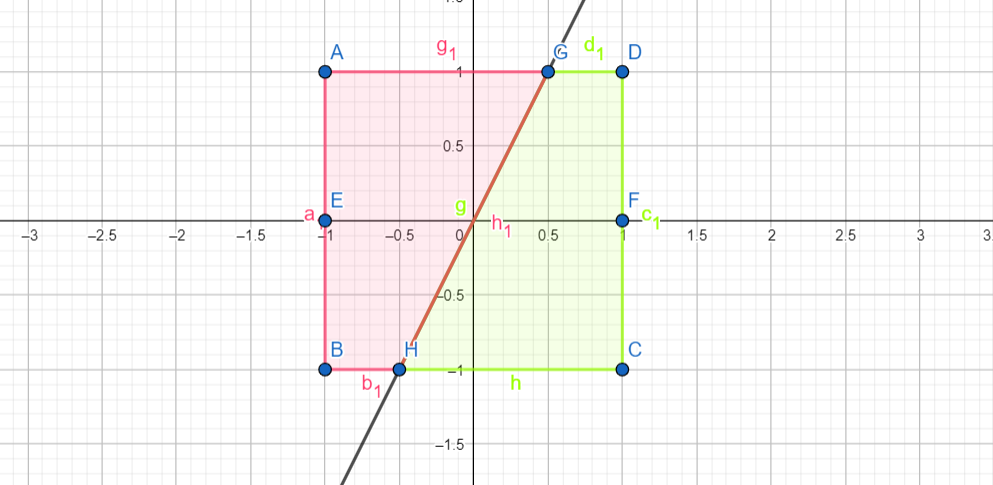
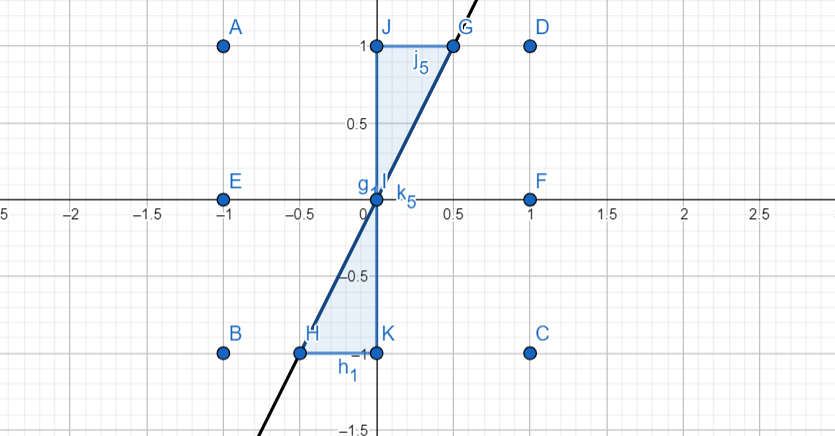
**11. C**

這是ff畫出的圖，綠色區域裡的點xx之f(x)=+1f(x)=+1；紅色區域裡的點xx之f(x)=−1f(x)=−1。  


這是h2h2劃出的圖：  
  
比對區域後，會發現h2(x)≠f(x)h2(x)≠f(x)的區域為：  


可計算出[−1,+1]×[−1,+1][−1,+1]×[−1,+1]的面積為4，h2(x)≠f(x)h2(x)≠f(x)的區域面積為2，因為我們取點的方式是uniform，所以隨機取一點xx，h2(x)≠f(x)h2(x)≠f(x)的機率為1212，h2(x)=f(x)h2(x)=f(x)的機率為1212，。  
要使Ein(h2)=0Ein(h2)=0，這五點都要落在h2(x)=f(x)h2(x)=f(x)的區域內，機率(12)5=132(12)5=132。

**12. D**

首先，畫出h1h1的圖：  
  
我們知道h1≠fh1≠f的區域為：  
  
與h2≠fh2≠f的區域比較，我們知道隨機取一點，有三種可能：

1. h1=f,h2≠fh1=f,h2≠f，機率1212
2. h1≠f,h2=fh1≠f,h2=f，機率1818
3. h1=f,h2=fh1=f,h2=f，機率3838

因為我們要使Ein(h1)=Ein(h2)Ein(h1)=Ein(h2)，我們知道狀況1和狀況2的點要一樣多，否則等式不會成立。以下列出所有可能的點數目與機率：

| **狀況1** | **狀況2** | **狀況3** | **機率** |
| --- | --- | --- | --- |
| 0 | 0 | 5 | (38)5=24332768(38)5=24332768 |
| 1 | 1 | 3 | 5!3!(12)(18)(38)3=13520485!3!(12)(18)(38)3=1352048 |
| 2 | 2 | 1 | 5!2!2!(12)2(18)2(38)=4510245!2!2!(12)2(18)2(38)=451024 |

加總起來，Ein(h1)=Ein(h2)Ein(h1)=Ein(h2)的機率為24332768+1352048+451024=38433276824332768+1352048+451024=384332768。

**13. B**

最首先，我們看hihi及hi+dhi+d，i∈[1,d],i∈Ni∈[1,d],i∈N：  
因為∀x∈D,hi(x)=−hi+d(x)∀x∈D,hi(x)=−hi+d(x)，所以原本hihi預測對的，hi+dhi+d反而都會預測錯；反之亦成立。或說[hi(xk)=yk]+[hi+d(xk)=yk]=1[hi(xk)=yk]+[hi+d(xk)=yk]=1。  
我們也可得知：  
{Eout(hi)+Eout(hi+d)=1Ein(hi)+Ein(hi+d)=1{Eout(hi)+Eout(hi+d)=1Ein(hi)+Ein(hi+d)=1  
所以Eout(hi)=1−Eout(hi+d),Ein(hi)=1−Ein(hi+d)Eout(hi)=1−Eout(hi+d),Ein(hi)=1−Ein(hi+d)，所以：  
|Eout(hi)−Ein(hi)|=|(1−Eout(hi+d))−(1−Ein(hi+d))|=|−Eout(hi+d)+Ein(hi+d)|=|Eout(hi+d)−Ein(hi+d)||Eout(hi)−Ein(hi)|=|(1−Eout(hi+d))−(1−Ein(hi+d))|=|−Eout(hi+d)+Ein(hi+d)|=|Eout(hi+d)−Ein(hi+d)|  
所以hihi與hi+dhi+d算出來的誤差是一樣的，所以只要hihi的誤差不超過ϵϵ，那hi+dhi+d的誤差也不會超過。所以我們判斷[BAD D for HBAD D for H]的時候，考慮i∈[1,d]i∈[1,d]的hihi即可。  
所以：  
Pr([BAD D for H])=Pr(2d⋃i=1[BAD D for H)=Pr(d⋃i=1[BAD D for hi)≤d∑i=1Pr([BAD D for hi)≤d∑i=12⋅exp(−2ϵ2N)=2d⋅exp(−2ϵ2N)=C⋅2⋅exp(−2ϵ2N)Pr([BAD D for H])=Pr(⋃i=12d[BAD D for H)=Pr(⋃i=1d[BAD D for hi)≤∑i=1dPr([BAD D for hi)≤∑i=1d2⋅exp(−2ϵ2N)=2d⋅exp(−2ϵ2N)=C⋅2⋅exp(−2ϵ2N)  
所以C=dC=d。  
因為上述討論的hihi與hi+dhi+d間的關係，可以不用考慮後面dd個hypothesis，所以CC可以從原本的2d2d壓成dd。

**14. D**

先列出每種hypothesis的機率：

|  | **1** | **2** | **3** | **4** | **5** | **6** |
| --- | --- | --- | --- | --- | --- | --- |
| A | 橘 | 綠 | 橘 | 綠 | 橘 | 綠 |
| B | 橘 | 綠 | 綠 | 綠 | 橘 | 橘 |
| C | 橘 | 橘 | 橘 | 橘 | 橘 | 綠 |
| D | 橘 | 綠 | 綠 | 橘 | 綠 | 橘 |
| 橘機率 | 11 | 1414 | 1212 | 1212 | 3434 | 1212 |
| 綠機率 | 00 | 3434 | 1212 | 1212 | 1414 | 1212 |

拿到五個綠3的機率為(12)5(12)5。  
( a )00  
( b )(14)5(14)5  
( c )(34)5(34)5  
( d )(12)5(12)5  
( e )(14)5(14)5  
所以是D。

**15. C**

拿出來的骰子只有ABCD四種且機率uniform，也可以知道假設五顆骰子中只有BC兩種並且兩種同時出現的話(例如BCBBC或BBBBC等組合，不包含BBBBB與CCCCC)，就不可能可以有pure green的數字出現。  
統計一下，如果五顆骰子中出現的種類組合只有出現下列組合中任意種：  
A, B, C, D, AB, AC, AD, BD, ABD  
那就會有一個數字pure green。  
視每次的骰子不同的話，總共有45=102445=1024種可能性。把所有上述組合數的列出來：

| **組合** | **數量** |  | **組合** | **數量** |
| --- | --- | --- | --- | --- |
| A | 15=115=1 |  | AC | C2025−C2115+C2205=30C0225−C1215+C2205=30 |
| B | 15=115=1 |  | AD | C2025−C2115+C2205=30C0225−C1215+C2205=30 |
| C | 15=115=1 |  | BD | C2025−C2115+C2205=30C0225−C1215+C2205=30 |
| D | 15=115=1 |  | ABD | C3035−C3125+C3215+C3305=150C0335−C1325+C2315+C3305=150 |
| AB | C2025−C2115+C2205=30C0225−C1215+C2205=30 |  |  |  |

注意ABD代表的是「只有出現ABD且同時要出現」。需要排除例如A沒出現等等的狀況。  
所以機率為1⋅4+30⋅4+1501024=27410241⋅4+30⋅4+1501024=2741024。

**16. B**

做出來就是11。

**17. B**

做出來是-7。

**18. C**

做出來是15。

**19. D**

做出來是17。

**20. D**

做出來是17。  
事實上的確應該跟19題一樣，因為ηη如果是正的常數應該不會影響學習速率

發表於 [**HackMD**](https://hackmd.io/)

 70

[讚賞](https://hackmd.io/@Kaiserouo/SJjJDPnHD) [收藏](https://hackmd.io/@Kaiserouo/SJjJDPnHD) [訂閱](https://hackmd.io/@Kaiserouo/SJjJDPnHD)