申論題 1&2

1.請上 Kaggle, 在 Competitions 或 Dataset 中找一組競賽或資料:

我選的是數字辨識的競賽，這個競賽的目的是希望能透機器學習的方式，使電腦能夠辨識手寫的數字，

Q1. 你選的這組資料為何重要?

Ans1:這組資料裡面提供了很多手寫數字的影像，也因為是手寫的數字，字體的複雜程度是比較高的，相較於電腦打字時的版本，更能使機器學習到不同字體的數字，不會因為字寫得比較醜而不能被機器辨識。之所以會選擇這個競賽是因為其應用非常的廣泛，雖然這個競賽只關注在辨識數字，但是這樣的概念還可以延伸到不同的問字，希望未來還能辨識各國語言的文字。而這樣的應用實際上是非常廣泛的，包括車牌辨識、貨物的批號辨識、言論管制、文章內容分類等等。

Q2.資料從何而來 (tips: 譬如提供者是誰、以什麼方式蒐集)?

Ans2:這個資料出自於 MNIST (Modified National Institute of Standards and Technology) dataset，裡面的資料是由美國的高中生以及 United States Census Bureau 的員工所提供的手寫數字影像。

Q3. 蒐集而來的資料型態為何?

Ans3:此競賽提供的資料有三個，分別是 training data, testing data，以及上傳檔案的格式。三種都是.csv file，其中 training data 裡面有 785 個 column，第一個 column 是影像所對應到的數字，所以第一個 column 裡面每一列的資料就是 0~9 的數字。其餘 784 個 column 則是影像 pixel 的編號(image size 為 28X28，總共 784 個 pixels)，由於影像是灰階影像，所以每一列的資料就是每一個 pixel 編號對應到的 pixel value，值都落在 0~255 之間，數值越大就代表顏色愈黑。而 testing data 則是只有 pixel value，並沒有數字 label。

Q4. 這組資料想解決的問題如何評估?
Ans4:想要解決的問題就是希望能夠準確的預測出影像所對應到的正確的數字，所以評估方式就是看整體的 accuracy，也就是正確的比率：

$$accuracy = \frac{correct}{correct + wrong} \times 100\ \%$$

2.想像你經營一個自由載客車隊，你希望能透過數據分析以提升業績，請你思考並描述你如何規劃整體的分析/解決方案：

Q1. 核心問題為何(tips：如何定義 「提升業績 & 你的假設」)?
Ans1:首先要提出一個提升業績的方法之前，必須先瞭解有哪些因素會影響整體的業績，如果我們能找到一個方程式，input 是這些影響因素，output 是過去的業績，也就是說如果我們可以成功預測業績，那麼我們就可以從方程式的參數來得知 input 對於 output 的影響。那麼可能影響整體業績的因素為：

1. 駕駛數量(x)
2. 駕駛與員工上班時間(y)
3. 每個員工平均一天載客的次數(z)
4. 平均載一次客人可以得到的收入(A)
5. 平均載一次客人車子所行駛的距離(B)
6. 平均載一次客人所花的時間(C)
7. 電話接客的次數(D)

列出這些因素之後就可以提出一個 model，也就是這些影響因子跟過去記載的每月業績之間的關聯性，公式如下：

$$每月業績 = f(x, y, z, A, B, C, D)$$

$f(x, y, z, A, B, C, D)$為一個方程式，會受到這七個因子影響。所以我們可以將過去每個月的業績，以及每個月所收集到的這些資料，進行 fitting，找到一組參數，假設每一個因子對於業績的影響都沒有很複

雜，都不需要加入二次式甚至更高次的函式，則我們可以將公視寫成入下列表示：

每月業績 = $f(x, y, z, A, B, C, D)$ = ax+by+cz+pA+qB+rC+oD
若這些資料和方程式可以 fit 的很好，則有機會可以成功預測每個月的業績，也可以透過觀察這些參數的值得知哪一個因子對於整體的業績影響較大，比如說假設經過 fitting 得到 a = 2, q = -1 p = 10，就可以解釋說員工數(x)以及平均載一次客人所得到的收入(A)的影響是正向的，如果員工數越多，收入越多，那麼業績也就越大；然而相反的，如果每一次載客的距離(B)都很遠，對於業績來說可能就會有負面的影響，可能就要考慮減少載客的距離。總結針對這個問題的回答，如果將資料 fit 過去的業績，找到一個資料和業績之間的關係，就可以知道每個資料對業績的影響是正面的還是負面的，也就可以進一步規劃提升業績的方法。

Q2. 資料從何而來 (tips：哪些資料可能會對你想問的問題產生影響 & 資料如何蒐集)
Ans2:上述的七種資料都可以透過網路即時的回傳資料給總部，由總部相關單位進行紀錄，並且分析。收集方式也不太困難，例如員工數由人事資料即可得知、上班時間以打卡的方式做紀錄、載客距離與時間則可以透過導航的紀錄得知。

Q3. 蒐集而來的資料型態為何?
Ans3: data type 都是數字，可以是 int，也可以是 float。並且格式都是 csv 的檔案格式，方便我們進行 model fitting。

Q4. 你要回答的問題，其如何評估 (tips：你的假設如何驗證)?
Ans4:從 Q1 得知，我們要先找到一個可以成功預測業績的 function，所以首先我們要先來衡量這個被找到的 function 是不是真的可以預測。怎麼做呢？在實作上我們通常都會定義一個 cost function，可以是 RMSE，也可以是 CrossEntropyLoss，後者比較常用在分類的問題，所以這一次我們將使用 RMSE，也就是預測的業績跟 ground truth(實際的每月業績)之間的平均離差平方和。只要找到一個 function，可以使這一個離差評方盒的直來到最小，就可以得到最佳的

function。找到最佳解之後就可以觀察這些 input 的係數分別是多少，得知 input 對於 output 的影響。