HW2 應數碩一統計組 M072040019 梅瀚中

1. For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

   (a) The sample size n is extremely large, and the number of predictors p is small.

   (b) The number of predictors p is extremely large, and the number of observations n is small.

   (c) The relationship between the predictors and response is highly non-linear.

   (d) The variance of the error terms, i.e. $\sigma^2 = \mathrm{Var}(\epsilon)$, is extremely high.

(a)

Better, 當 n 夠大時, 靈活性大的模型比靈活性小的模型表現較好, 因為 n 夠大, 可以降低過度擬合的狀況。

(b)

Worse, 當 n 夠小時, 如果在使用靈活性大的模型, 很容易產生過度擬合的情況, 所以使用靈活性較小的模型較適合。

(c)

Better, 在高度非線性的狀況下, 靈活性較大的模型比較適合, (如果使用靈活性較小的模型容易 underfitting)

(d)

Worse, 使用靈活性大的模型在此情況容易 overfitting, 造成模型去配適不能解釋的噪音。

2. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

   (a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

   (b) We are considering launching a new product and wish to know whether it will be a *success* or a *failure*. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

   (c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

(a) regression, inference, n=500, p=3

這是一個 regression 問題, 因為 CEO 的薪水是一個連續的變數。我們感興趣的是 inference, 我們想了解自變數 (profit, number of employees, industry)是如何影響反應變數(CEO salary)的 。n=500(500firms) p=3(profit, number of employees, industry)

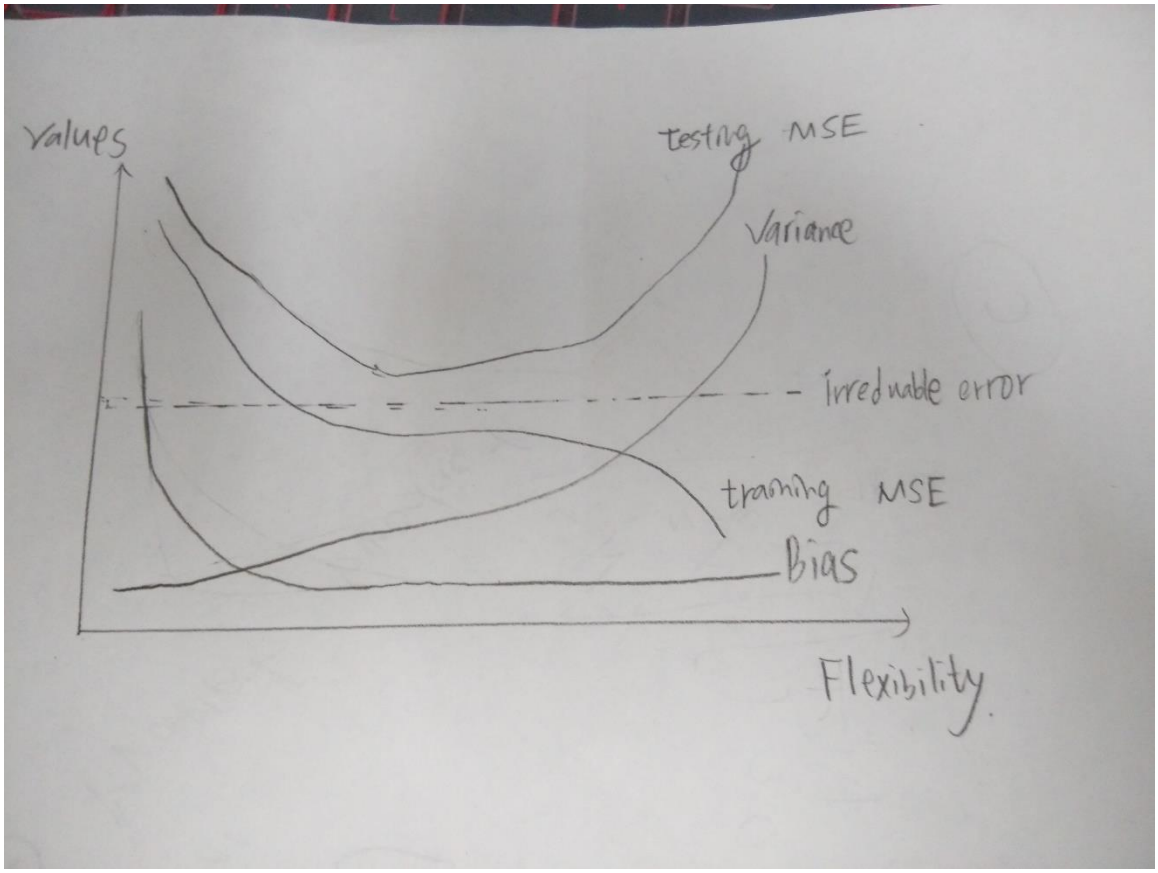
(b) classification, prediction , n=20, p=13

這是一個 classification 問題, 因為我們想把結果分為 success 和 failure。我們感興趣的是 prediction, 我們已經蒐集過以前的 20 筆資料, 藉由這 20 筆資料, 我們得到的資訊, 去預測新的一筆資料是屬於成功或失敗。n=20 (20 similar products previously launched) , p=13(price charged for the product, marketing budget, competition price, and ten other variables)

(c) regression, prediction , n=52, p=3

這是一個 regression 問題, 因為% change in the USD/Euro dollars 是一個連續型的變數。我們感興趣的是 predicting the % change in the US dollar。n=52(365/7=52.14 取整

數）,p=3(% change in US, % change in British, % change in German).
(3)(a)



(b)
training MSE : training MSE 會隨著模型靈活度越高,呈下降趨勢。但過高靈活度的模型,容易造成 overfitting 的情況,所以單看一個 training MSE 是一個很小的值,並不代表這是一個好的 model
testing MSE: testing MSE 會根據模型靈活度越高,一開始呈現下降趨勢,後隨著模型靈活度過高,造成 overfitting 狀況,也就是 training MSE 過低而 testing MSE 過高的狀況。
Bias:隨著模型靈活度越高,代表模型越去描繪 data 的狀況,造成 bias 會越低
Variance:隨著模型靈活度越高,variance 會越高。

Irreduable error:這是由 noise 造成的誤差,模型的變化對它不會有任何影響。

4. You will now think of some real-life applications for statistical learning.

   (a) Describe three real-life applications in which *classification* might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

   (b) Describe three real-life applications in which *regression* might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

   (c) Describe three real-life applications in which *cluster analysis* might be useful.

(a)

(1)成績及格的分類(>=60)或(<60)。predictor:唸書時間,睡眠時間。Prediction 問題

(2)有沒有得肺癌 。 predictor:有無抽菸,有無家人抽菸等。Prediction 問題

(3)找工作去面試有沒有上榜。Predictor:為人談吐,專業知識的多寡等。Prediction 問題

(b)

(1) 房屋的價錢 ,predictor:坪數,地段等 。inference 問題

(2) 一台車可以跑的里程數,predictor:汽缸數,燃料類型等

Inference 問題

(3)考試成績 ,predictor:唸書時間,睡眠時間,專心程度等 。Inference 問題

(c)

(1)將國家分為發達國家，發展中國家和第三世界。

(2)依薪水分為高收入,普通收入,和低收入戶。

(3)把一些相同基因表現的蛋白分為同一群。

5. What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

Flexible approach:

優點: low bias 。

缺點: (1)high variance ,(2)比較難解釋,(3)可能 overfitting。

使用時機: (1)當原本模型 underfitting 時,可以採用更 flexible 的方法來調整。(2)當資料呈現高度非線性相關時。

Less flexible approach:

優點: (1)low variance ,(2)解釋性佳。

缺點: (1)high bias ,(2)可能 underfitting

使用時機:(1)當原本模型 overfitting 時,可以採用 less flexible 的方法來調整,(2)資料呈現線性關係,(3)需要一個容易解釋的模型的時候。