Statistical Learning and Data mining

Homework 2

M052040003 鍾冠毅

1.a. 較佳。以迴歸分析為例，當樣本數多時，一階模型可能造成更多的變異，若使用二階獨立變數，例如二次方項或是交叉項，將使模型的預測更加有彈性，進而降低變異。

1.b. 較差。與上題相反，過多的變數將導致模型 overfit，因此可以使用個總選模方法，或探討共線性問題，以減少不必要之變數。

1.c. 較佳。若執意使用較低彈性的線性迴歸，將造成更大的變異，可依照資料需求，使用廣義線性迴歸，或是增加變數，以獲得較低變異的模型。

1.d. 較差。變異越高，可能因為變數過多，導致 overfit，因此應該降低其彈性。

2.a. Regression problem. Reference.
探討迴歸分析中，哪一個獨立變數對於模型有較顯著的影響。
n=500, top 500 firms in the U.S.
p: profit, number of employees, industry

2.b. Classification problem. Predict.
分析各項數據，以預測最後產品分類為成功或是失敗。可用 random forest 作為分類方法，並找到最佳的決策樹。
n=20, 20similar products that were previously launched.
p: price charged for the product, marketing budget, competition price, and ten other variables

2.c. Regression problem. Predict.
以每周的股票市場改變百分比預測下周的美津改變比率，可以使用時間序列之方法預測。
n=52, 52 weeks in 2012.
p: the % change in the US market, the British market and the German market.

4.a.i. 預測學生數學段考成績是否會及格。預測。
Predictor : 每周讀書時間、是否玩線上遊戲、是否有補習
Response : 數學成績是否及格

4.a.ii. 探討顧客回饋單中，何種因素使顧客有意願再次光臨。推論。
Predictor : 服務品質、價位、餐飲品質、店內裝潢等五等級滿意度。
Response : 是否有意願再次光臨用餐。

4.a.iii. 預測顧客是否會購買某樣化妝品 Y。預測。
Predictor : 性別、是否已婚、是否有小孩、年收入、治裝費
Response: 是否購買化妝品 Y

4.b.i. 預測癌症病患存活率。預測
Predictor : 期數、是否有其他癌症、慢性病與否、化療次數
Response: 死亡時間、未死亡時間

4.b.ii. 探討什麼原因最能影響夏季室內溫度。推論。
Predictor : 開窗與否、是否在地上潑水、電風扇強度、太陽直射與否
Response: 室內溫度

4.b.iii. 了解什麼變數與該地區電力負載量較無關係。推論。
Predictor : 氣溫、濕度、平均風速、工業區/文教區/商業區、居民收入
Response: 電力負載量

4.c.i. 欲以籃球球員各項數值，將數值表現相近的選手分為同一群。

4.c.ii. 文字探勘中找出同一段落常同時出現的字詞。

4.c.iii. 在地圖上點出 SARS 患者住家座標，探討各病例是否屬於同一個感染區

5. 彈性較大的方法在非線性的資料中非常適用，可以降低 bias；缺點則是過於彈性的模型，可能導致 overfit，使得 test error/ MSE 過高。
在非線性的資料中，線性模型過於僵化，因此可以加入更多的變數，如交叉項或是高次項，以增加模型彈性，得到較低 bias 的預測模型。
過多的變數可能導致獨立變數中有共線性存在，若要推測何種變數影響模型最多，則可能因為共線性之問題而受到影響，因此解決共線性問題，選取關鍵的變數，降低模型的彈性，方能找到其解。

6. parameter approach 給定模型的假設，例如各種方法或是分布；non-parameter approach 則無。有母數方法較為簡單，因為已經給定了模型的型式，則簡化了使用其他方法的可能。但也因為給予既定的預測模型型式，而得到 bias 較高，彈性較差的模型。