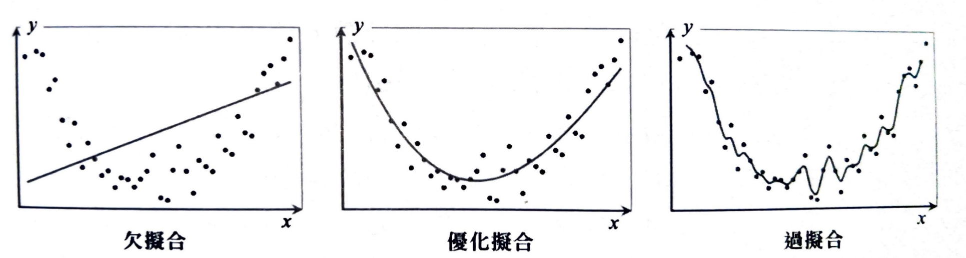
模型評估

編輯者：李紘宇

欠擬合與過擬合

有時模型太簡單，不管如何訓練，都無法很好地擬合訓練資料，稱之為「欠擬合」。有時模型太複雜，雖然很好地擬合了訓練資料，卻在測試的樣本上有很大的誤差(泛化能力差)，則稱之為「過擬合」。下面左右兩圖即是欠擬合與過擬合的例子，而我們期望的「好」模型，應是中間的最佳化擬合。但僅憑下圖無法得知，是否欠擬合或過擬和，因此之後的章節會說明判斷的方法。



解決辦法：

1. 欠擬合：

- 增加樣本特徵的數量，增加資料複雜度
  （ex：將訓練資料(人口數,新生兒數量)，擴增為(人口數,人類幸福指數,新生兒數量)）
- 提高模型複雜度
  （ex：擬合函數由二次多項式，變為六次多項式）
- 降低正則化程度
  （正則化是一種透過加入額外資訊於訓練過程中，來避免過擬合的方法）

2. 過擬合：

- 增加訓練樣本的數量
- 降低模型複雜度
- 透過正則化限制模型的複雜度

訓練集、驗證集和測試集

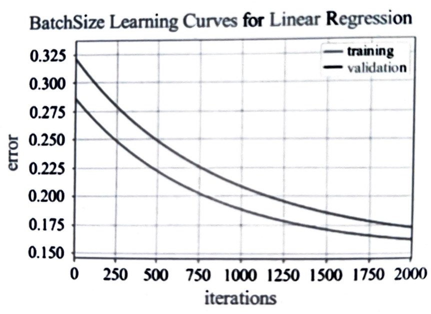
在我們訓練模型的過程中，我們一般會有三個步驟，訓練模型、調整模型和測試模型，因此我們也會將樣本資料分成三個部分，訓練集、驗證集和測試集。其功能如下：

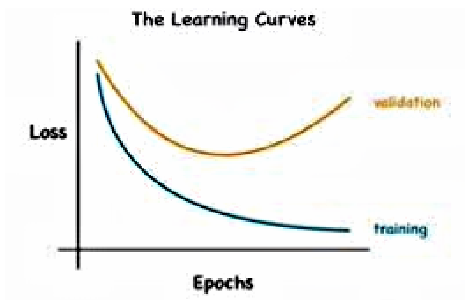
訓練集： 用來訓練模型的樣本，作為調整模型參數的依據。
驗證集： 用來評估和選擇模型，作為調整超參數的依據。
測試集： 用來測試最終模型的的預測能力。
Hint:參數是指模型自行訓練出來的變量（ex：線性回歸模型y=ax+b中的a和b）。超參數是指根據經驗事先給定的參數（ex：學習率、迭代次數）
樣本少（ex：醫學影像較難取得）時，三個樣本集通常切分成60%、20%、20%；樣本多(數十萬、數百萬)時，可以切分成90%、5%、5%。可依樣本數多寡自行決定比例。

學習曲線

學習曲線是指任何有助於判斷訓練情況的曲線，常用的有訓練曲線和驗證曲線。我們期望的好模型，訓練損失和驗證損失都應該是低的。因此透過將誤差對超參數作圖，我們可以從曲線中找到，最適合模型的超參數值。以下分別以迭代次數、訓練樣本數、模型複雜度為例。

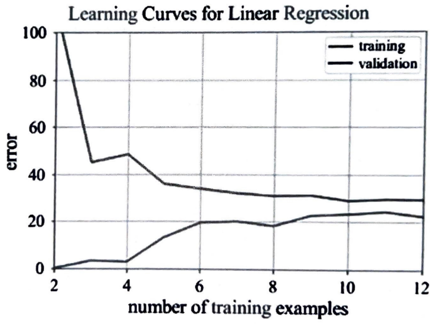
1. 損失 vs 迭代次數
   一個好的模型，訓練損失曲線和驗證損失曲線都應該越來越低（如左下圖所示）。若是出現右下圖的情況，迭代到某個值之後，驗證誤差不減反增，就代表模型開始出現過擬合，我們可以選擇提早停止迭代，讓模型停在最好的情況，此方法稱之為「**早停法**」。

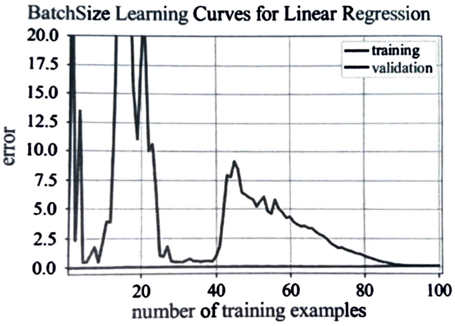




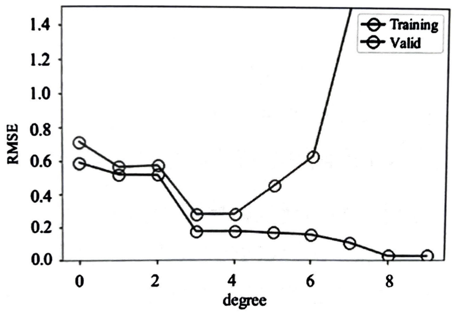
2. 損失 vs 訓練樣本數
   左下圖可以發現，當訓練樣本變多時，因為模型越來越難擬合所有樣本，因此訓練誤差會逐漸增大。並且可以看到，驗證誤差會越來越小，到最後幾乎不變，這時便可以早停，不需再增加訓練樣本。但也可能出現右下圖的情

況，樣本數40之前起伏劇烈，之後驗證誤差才開始慢慢接近訓練誤差，這時我們的訓練樣本數量應該超過40個。





3. 損失 vs 模型複雜度
   以模型複雜度作損失曲線，可以找出最適合的模型複雜度，以下圖的多項式擬合為例，可以得知三次或四次多項式是較為適合的模型。



   ps：↓ 裡面好像有模型複雜度的一些介紹，但我還沒看懂
   https://medium.com/機器學習基石系列/機器學習基石-4-vc-dimension和模型複雜度-5398ed1c8a5e

偏差與方差

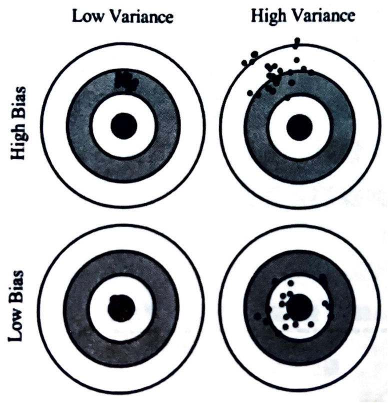
在某個問題中，引數x和因變數y應滿足關係式y=f(x)，然而在我們獲取樣本的過程中，可能會出現雜訊，導致我們獲取的樣本(x_i, y_i)偏離真實值，也就是$y \neq f(x)$。兩者之間的誤差通常認為符合高斯分佈，也就是$\epsilon = y - f(x) \sim N(0, \sigma^2)$。在模型訓練的時候，我們會用一個假設函數（ex：$\hat{f} = ax + b$）訓練，透過最小化$(y_i - f(x_i))^2$，來求得$\hat{f}$（也就是求a和b）。但不同的訓練集、不同的機器學習演算法，會求出不同的$\hat{f}$。因此在固定的x，我們可以定義誤差的平均如下

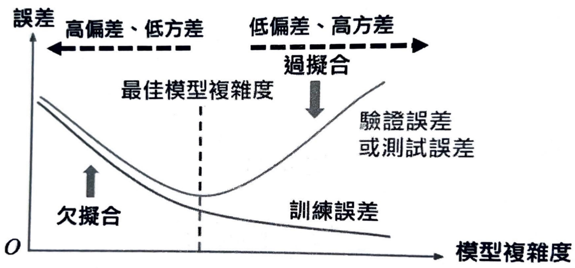
- **期望誤差** $= E[(y - \hat{f}(x))^2] = (Bias[\hat{f}(x)])^2 + Var[\hat{f}(x)] + \sigma^2$
  其中，$Bias[\hat{f}(x)] = E[\hat{f}(x)] - E[f(x)]$，稱作**偏差**。

$$Var[\hat{f}(x)] = E[\hat{f}(x)^2] - E[\hat{f}(x)]^2 = E(\hat{f}(x) - E[\hat{f}(x)])^2$$

，稱作**方差**。

偏差，可用來檢測模型預測值和真實值的差距大小，用來評估模型的準確度。方差，可用來檢測模型預測值的分散程度，用來評估模型的精確度。可以看左下圖來理解偏差與方差，其中靶心是真實值，黑點表示每次模型訓練完後的預測值。偏差、方差與模型複雜度的關係可以由右下圖看出，透過作模型複雜度曲線，我們可以在欠擬合與過擬合之間找到最佳的模型複雜度。





正則化

正則化是一種透過技術手段限制模型，以降低模型複雜度的方法。前面提到的早停法便是一種正則化方法，它透過限制模型的誤差增大，來降低模型複雜度。

另一種方法是對損失函數增加懲罰項。舉一個線性擬合問題的例子，其模型的假設函數是$f(\mathbf{x}) = \mathbf{x} \cdot \mathbf{w}$，且共有m個訓練樣本 $(\mathbf{x}^{(i)}, y^{(i)})$，我們定義添加正則項的損失函數為 $L(\mathbf{w}) = \frac{1}{2m} \sum_{i=1}^{m} \|\mathbf{x}^{(i)} \cdot \mathbf{w} - y^{(i)}\|^2 + \lambda \|\mathbf{w}\|^2$，其中第一項是我們原始的損失函數，第二項是正則項。這個正則項可以避免權重$\mathbf{w}$過大。λ是超參數，需視情況調整，用來控制懲罰項的貢獻。