2.Overfitting and Model Choice

機器學習基本框架

機器學習的目標是從訓練資料中學習一個函數，然後用這個函數來預測測試資料。

- **訓練資料（Training Data）**：一組已知輸入 (x) 和其對應真實標籤 (y^) 的資料集。$(x1, y1), (x2, y2), \ldots, (xN, yN)$
- **測試資料（Testing Data）**：一組只有輸入 x 的資料集，需要用訓練好的模型進行預測。$xN + 1, xN + 2, \ldots, xN + M$

訓練步驟：

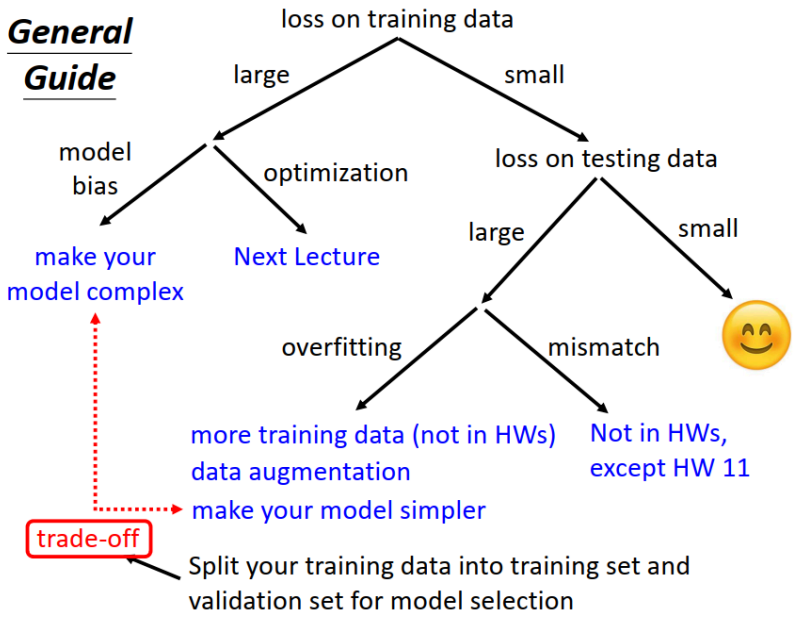
1. **定義函數集合（Model）**：選擇一個包含未知參數 θ 的函數 y=fθ(x)。
2. **定義損失函數（Loss Function）**：根據訓練資料定義一個損失函數 L(θ)，用來衡量模型的好壞。
3. **最佳化（Optimization）**：找到一組最佳參數 θ∗，使得損失函數 L(θ) 達到最小值。

   θ∗=arg minθL(θ)
4. **應用於測試資料**：使用找到的最佳函數 y=fθ∗(x) 來預測測試資料的標籤。

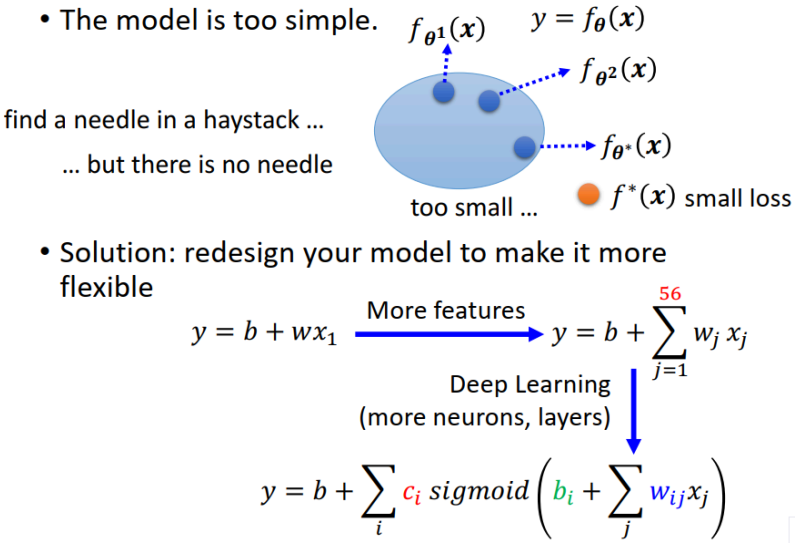
模型訓練問題診斷與應對

當模型在訓練後表現不佳時，通常可以透過觀察訓練損失和測試損失來診斷問題。

訓練損失 (Training Loss)	測試損失 (Testing Loss)	問題診斷
大	N/A	**模型偏差 (Model Bias) 或 最佳化問題 (Optimization Issue)**
小	**大**	**過擬合 (Overfitting)**
小	**小**	**成功！** 😊

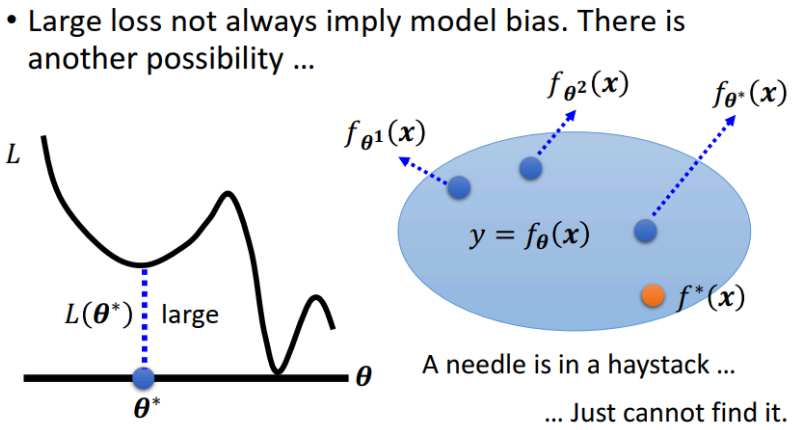


1. 模型偏差 (Model Bias)



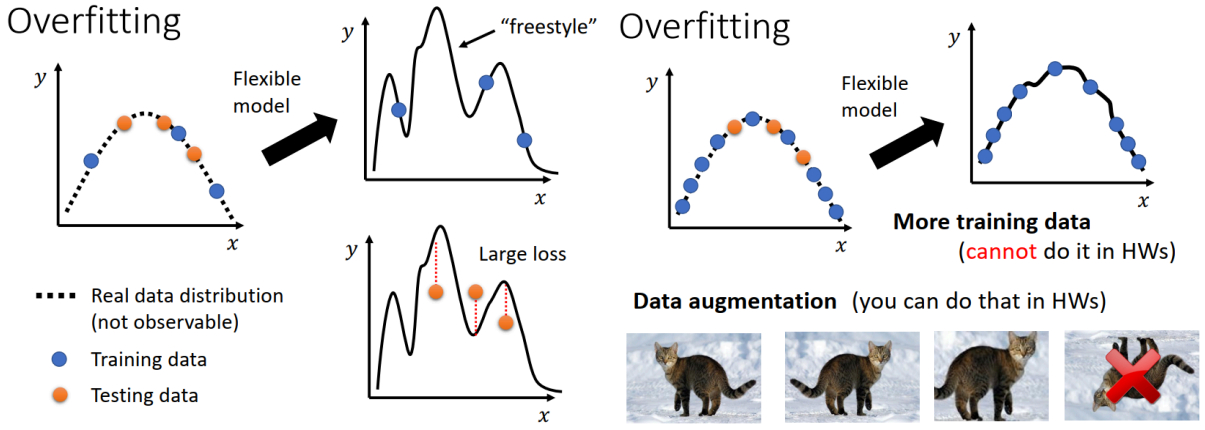
- **問題描述**：當訓練損失很大時，可能表示你選擇的模型太過簡單，無法捕捉訓練資料中的複雜模式。就像在「乾草堆裡找針，但乾草堆裡根本沒有針」。
- **解決方法**：
  - 重新設計模型，使其更具彈性（More flexible）。
  - 增加更多的特徵（Features）。
  - 增加模型的複雜度，例如在深度學習中增加神經元或層數。

2. 最佳化問題 (Optimization Issue)



- **問題描述**：即使模型足夠複雜，訓練損失仍然很大，這可能是因為最佳化演算法無法找到損失函數的最小值。就像「乾草堆裡有針，但你就是找不到它」。
- **解決方法**：
  - 使用更強大的最佳化演算法（這將在後續課程中討論）。
  - 可以從較淺或較簡單的網路開始，因為它們更容易最佳化。

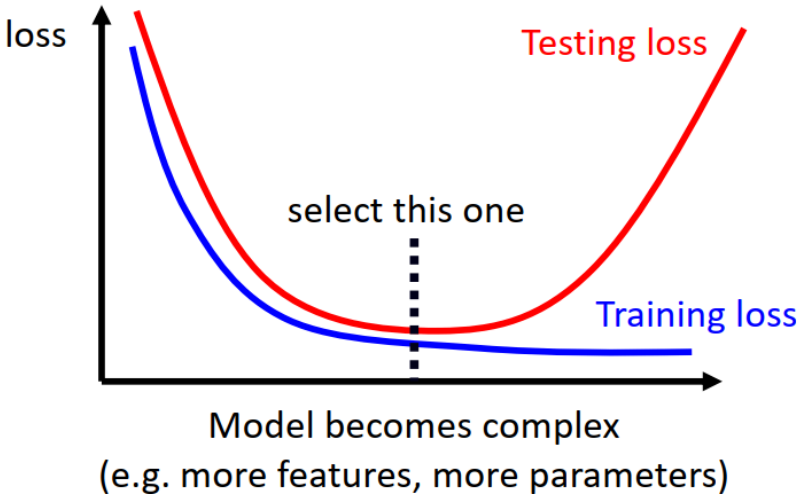
3. 過擬合 (Overfitting)



- **問題描述**：訓練損失很小，但測試損失很大。這意味著模型過度學習了訓練資料中的雜訊或特有模式，而無法泛化到新的資料上。一個極端的例子是模型只記憶訓練資料的輸入和輸出，對沒見過的資料則隨機預測。
- **解決方法**：
  - **減少模型複雜度**：使用更少的特徵或參數。
  - **提早停止（Early Stopping）**：在訓練過程中，當模型在驗證集上的表現開始變差時就停止訓練。
  - **正規化（Regularization）**：在損失函數中加入一個懲罰項，來限制模型參數的大小，以避免參數值過大。
  - **Dropout**：在訓練過程中隨機地「關閉」一部分神經元，迫使模型不依賴於任何特定的神經元。

- **增加訓練資料**：獲取更多的訓練資料，這可以幫助模型學習更通用的模式。
- **資料增強（Data Augmentation）**：透過對現有資料進行變換（如圖片的旋轉、翻轉、裁剪等）來產生新的訓練資料。

模型偏差 vs. 過擬合的權衡 (Bias-Complexity Trade-off)



- 模型複雜度與損失之間存在一個權衡關係。
- 隨著模型複雜度的增加，訓練損失會逐漸降低。
- 測試損失則先下降，達到最低點後再上升。最低點是模型在泛化能力上表現最佳的平衡點。
- 找到這個平衡點的過程就是模型選擇。

模型選擇 (Model Selection)

- **問題**：在有公共（Public）和私有（Private）測試集的競賽中，僅憑公共測試集來選擇模型可能導致過擬合，因為模型可能碰巧在公共集上表現好，但在私有集上表現差。
- **解決方法**：
  - **不要用公共測試集來選擇模型。**
  - **將訓練資料劃分為訓練集（Training Set）和驗證集（Validation Set）。**
  - **交叉驗證（Cross Validation）**：這是一種更穩健的模型選擇方法。

N-折交叉驗證 (N-fold Cross Validation)

1. 將訓練資料分成 N 個子集。

2. 進行 N 次訓練，每次使用 N-1 個子集作為訓練集，剩下的 1 個子集作為驗證集。
3. 計算每次訓練在驗證集上的平均表現，選擇平均表現最好的模型。

資料分布不匹配 (Mismatch)

- **問題描述**：訓練資料和測試資料來自不同的分佈。
- **結果**：單純增加訓練資料量並不會有幫助。
- **解決方法**：需要深入理解資料是如何生成的。