機器學習第二章：完整筆記與模擬出題（未壓縮）

本文件包含統計學習方法的名詞解釋、名詞比較與模擬出題（中英文對照），所有內容逐字保留未壓縮

# 📘 第二章：統計學習（Statistical Learning）

## 🧠 名詞解釋（Term Explanation）

### 統計學習（Statistical Learning）

中文：研究如何從資料中學習函數 f(X) 以預測目標 Y，包含監督與非監督學習。

英文：Statistical learning studies how to learn the function f(X) from data to predict a target Y, covering both supervised and unsupervised learning.

### 監督式學習（Supervised Learning）

中文：資料中包含輸入與對應的輸出（Y），模型學習輸入與輸出之間的映射關係。

英文：Data includes inputs and corresponding outputs (Y), and the model learns the mapping between them.

### 非監督式學習（Unsupervised Learning）

中文：資料中只有輸入沒有標記，目標是探索資料內部結構（如分群）。

英文：Data contains inputs only with no output labels; the goal is to discover patterns or structures such as clustering.

### 參數方法（Parametric Method）

中文：先假設模型形式（如線性），再估計參數。計算效率高，但若模型假設錯誤會產生偏誤。

英文：Assumes a specific model form (e.g., linear) and estimates its parameters. It is efficient but may suffer from bias if the assumption is wrong.

### 非參數方法（Non-parametric Method）

中文：不假設模型形式，讓資料決定模型形狀。更彈性，但需要更多資料來估計。

英文：Does not assume a functional form; the data determines the model. It is flexible but requires more data.

### 過度擬合（Overfitting）

中文：模型過於複雜，訓練誤差低但泛化能力差。

英文：Model is too complex, resulting in low training error but poor generalization.

### 欠擬合（Underfitting）

中文：模型太簡單，無法捕捉資料趨勢，導致訓練與測試誤差皆高。

英文：Model is too simple to capture data trends, leading to high error on both training and test sets.

### 偏誤（Bias）

中文：模型預測與實際值的系統性誤差，與模型假設錯誤有關。

英文：Systematic error between model predictions and actual values, often due to incorrect model assumptions.

### 變異（Variance）

中文：模型對訓練資料的敏感度，變異高表示模型易受樣本變動影響。

英文：The model's sensitivity to training data. High variance indicates strong dependence on specific training samples.

### 訓練誤差與測試誤差（Training/Test Error）

中文：訓練誤差來自訓練資料，測試誤差用來評估模型泛化能力。

英文：Training error is calculated on the training set; test error measures model generalization.

## 📚 名詞比較（Term Comparison）

參數方法 vs 非參數方法

|  |  |  |
| --- | --- | --- |
| 項目 | 參數方法（Parametric） | 非參數方法（Non-parametric） |
| 模型假設 | 需先假設模型形式 | 無需假設模型形式 |
| 彈性 | 低 | 高 |
| 所需資料量 | 少 | 多 |
| 例子 | 線性回歸 | KNN、決策樹 |

高偏誤 vs 高變異

|  |  |  |
| --- | --- | --- |
| 項目 | 高偏誤（High Bias） | 高變異（High Variance） |
| 模型特性 | 太簡單，欠擬合 | 太複雜，過擬合 |
| 訓練誤差 | 高 | 低 |
| 測試誤差 | 高 | 高 |
| 改善方法 | 增加複雜度 | 簡化模型、正則化 |

# 📝 第二章模擬出題（Chapter 2 Mock Questions）

## 考題 1：名詞解釋 Term Explanation

中： 什麼是統計學習？其主要任務為何？

EN: What is statistical learning? What is its main objective?

答：  
統計學習是從資料中學習函數 f(X) 的方法，目的是預測目標 Y，主要任務是從輸入變數 X 預測或理解輸出變數 Y。  
Statistical learning is a method of learning a function f(X) from data to predict a target Y. Its main task is to predict or understand the output Y based on input variables X.

## 考題 2：名詞比較 Term Comparison

中： 比較參數方法與非參數方法。

EN: Compare parametric and non-parametric methods.

答：  
參數方法先假設模型形式（如線性），再估計參數；非參數方法不先設模型形式，由資料決定模型形狀。參數法效率高但彈性低，非參數法彈性高但需較多資料。  
Parametric methods assume a model form and estimate parameters (e.g., linear regression), while non-parametric methods let data determine the model form. Parametric methods are efficient but less flexible; non-parametric methods are more flexible but require more data.

## 考題 3：名詞解釋 Term Explanation

中： 監督式學習與非監督式學習有何不同？

EN: What is the difference between supervised and unsupervised learning?

答：  
監督學習包含輸入與對應的輸出，用於預測；非監督學習僅有輸入，用於發掘資料結構。  
Supervised learning has both input and output for prediction; unsupervised learning has only input data and is used for discovering data structure.

## 考題 4：名詞比較 Term Comparison

中： 高偏誤與高變異的模型特性為何？如何改善？

EN: What are the characteristics of high bias vs high variance models, and how to improve them?

答：  
高偏誤模型太簡單，會欠擬合，訓練與測試誤差都高；可透過增加模型複雜度改善。高變異模型太複雜，會過擬合，訓練誤差低但測試誤差高，可透過正則化或簡化模型改善。  
High bias models are too simple, causing underfitting and high training/test error; can be improved by increasing complexity. High variance models are too complex, causing overfitting with low training but high test error; can be improved by regularization or model simplification.

## 考題 5：應用題 Application

中： 若一模型在訓練資料表現良好但在測試資料表現差，可能發生什麼問題？

EN: If a model performs well on training data but poorly on test data, what might be the issue?

答：  
這代表模型過度擬合訓練資料，缺乏泛化能力。  
This indicates overfitting—the model memorized the training data but cannot generalize to unseen data.

## 考題 6：名詞解釋 Term Explanation

中： 什麼是變異？為什麼高變異不好？

EN: What is variance? Why is high variance problematic?

答：  
變異是模型對訓練資料的敏感度，高變異代表模型受資料小變動影響大，導致不穩定預測。  
Variance refers to a model’s sensitivity to training data; high variance means the model’s predictions vary greatly with small changes in the data, leading to instability.

## 考題 7：名詞解釋 Term Explanation

中： 偏誤與變異之間存在什麼樣的取捨？

EN: What is the bias-variance tradeoff?

答：  
提升模型複雜度可降低偏誤但會增加變異，反之亦然。最佳模型需平衡兩者以達最低總誤差。  
Increasing model complexity reduces bias but increases variance, and vice versa. The goal is to find a balance that minimizes total error.

## 考題 8：應用題 Application

中： 線性回歸屬於參數或非參數方法？KNN 屬於哪類？

EN: Is linear regression parametric or non-parametric? What about KNN?

答：  
線性回歸是參數方法，KNN 是非參數方法。  
Linear regression is parametric, while KNN is non-parametric.

## 考題 9：應用題 Application

中： 欲提升模型的泛化能力，可採取哪些策略？

EN: What strategies can be used to improve model generalization?

答：  
可透過交叉驗證、簡化模型、正則化、增加資料量等方式提升泛化能力。  
Use cross-validation, simplify the model, apply regularization, or increase data volume to improve generalization.

## 考題 10：名詞解釋 Term Explanation

中： 訓練誤差與測試誤差有何不同？

EN: What is the difference between training and test error?

答：  
訓練誤差來自模型在訓練資料的預測誤差；測試誤差評估模型在新資料的預測能力。  
Training error is the prediction error on training data; test error evaluates predictive performance on unseen data.

# 📝 第二章：模擬考題與答案（題目與答案寫在一起）

1. Q1. (a)(b) 題組：統計學習概念

(a) What is statistical learning and why do we need it?  
Answer: Statistical learning is a branch of statistics and ML that focuses on developing models to make predictions or decisions based on data. We need it to select appropriate models, interpret predictions, and handle issues like noise or outliers.  
（a）什麼是統計學習？為何我們需要它？  
答：統計學習是統計與機器學習的子領域，用來建構能根據資料進行預測或判斷的模型。學習統計學習有助於模型選擇、結果解釋與處理資料中的雜訊與異常值。  
  
(b) Explain the difference between statistical learning and machine learning.  
Answer: Statistical learning focuses more on interpretability and is rooted in mathematical theory, while machine learning is broader and prioritizes prediction accuracy, including neural networks and deep learning.  
（b）請解釋統計學習與機器學習的差異。  
答：統計學習強調模型可解釋性，並建立於數理理論基礎上；機器學習則更強調預測準確度，範疇更廣，涵蓋深度學習等方法。

1. Q2. (a)(b) 題組：回歸函數與誤差來源

(a) What is the regression function f(x)? Why is it useful?  
Answer: It is the expected value of Y given X, f(x) = E(Y|X). It helps make predictions by capturing the relationship between input and output.  
（a）什麼是回歸函數 f(x)？它有什麼用處？  
答：f(x) 表示在給定 X 的情況下 Y 的期望值，即 f(x) = E(Y|X)。它可用來描述輸入與輸出變數的關係並進行預測。  
  
(b) Explain the difference between reducible and irreducible errors.  
Answer: Reducible error is due to inaccurate model estimation and can be reduced by better modeling. Irreducible error is caused by inherent data noise and cannot be eliminated.  
（b）請解釋可減少與不可減少的誤差差異。  
答：可減少誤差來自模型估計不準確，可透過改進模型來降低；不可減少誤差則來自資料內在雜訊，無法完全消除。

1. Q3. (a)(b) 題組：維度詛咒與模型假設

(a) What is the curse of dimensionality? How does it affect learning?  
Answer: It refers to the sparsity of high-dimensional data, making it hard to find neighbors and generalize well. It reduces the effectiveness of non-parametric methods.  
（a）什麼是維度詛咒？對學習有何影響？  
答：維度詛咒是指高維資料中樣本變得稀疏，使得尋找鄰近樣本與泛化效果變差，特別對非參數方法影響較大。  
  
(b) Compare parametric and non-parametric methods in the context of high-dimensional data.  
Answer: Parametric methods assume a form and are more efficient in high dimensions; non-parametric methods need more data and suffer from sparsity issues.  
（b）在高維資料下比較參數與非參數方法。  
答：參數方法假設模型形式，適用於高維資料時較有效率；非參數方法依賴鄰近資料，對維度詛咒較敏感。

1. Q4. (a)(b) 題組：模型誤差與泛化

(a) Why might a model perform well on training data but poorly on test data?  
Answer: The model is likely overfitted to the training data, capturing noise instead of general patterns, leading to poor generalization.  
（a）為什麼模型在訓練資料上表現很好，卻在測試資料上表現差？  
答：這通常是因為模型過度擬合訓練資料，記住了雜訊而非真正的模式，導致泛化能力不足。  
  
(b) Name two methods to improve model generalization.  
Answer: (1) Use cross-validation; (2) Apply regularization.  
（b）舉出兩種提升模型泛化能力的方法。  
答：(1) 使用交叉驗證；(2) 使用正則化技術。

1. Q5. (a)(b) 題組：模型彈性與偏誤變異分析（圖：bias\_variance\_tradeoff.png）

(a) According to the figure 'Bias-Variance Tradeoff', explain why test error decreases and then increases as model flexibility grows. What does the vertical dashed line represent?  
（a）根據圖中的「偏誤-變異權衡」曲線，解釋為什麼測試誤差在模型彈性增加時先下降後上升？圖中的虛線代表什麼？  
  
Answer:  
As model flexibility increases, the model fits training data better, reducing bias. However, too much flexibility leads to overfitting and increased variance, causing test error to rise. The dashed line represents the optimal complexity point that minimizes test error.  
隨著模型彈性提升，偏誤降低，但過度彈性會導致過擬合，使變異升高、測試誤差增加。虛線表示測試誤差最小的最佳模型彈性位置。

圖 Q5：Bias-Variance Tradeoff

