# 📘 第六章：分類（Classification）

本章介紹分類任務中的基本概念與演算法，包括貝氏定理、邏輯回歸、KNN分類、模型訓練方法（GD, SGD）與效能評估指標（如Confusion Matrix、Precision、Recall、F1、AUC-ROC）等。

## 🧠 名詞解釋（Term Explanation）

* 分類任務（Classification Problem）

中文：根據已標記資料訓練模型，以預測輸入屬於哪一類別。

英文：A classification task involves training a model on labeled data to predict which class a new input belongs to.

* 貝氏定理（Bayes' Theorem）

中文：計算後驗機率的公式：P(y|x) = P(y)P(x|y) / P(x)，更新對事件的信念。

英文：Bayes' theorem computes the posterior probability P(y|x) using prior P(y), likelihood P(x|y), and evidence P(x), to update beliefs based on data.

* 邏輯回歸（Logistic Regression）

中文：用於二元分類的演算法，透過sigmoid函數輸出機率值。

英文：A binary classification algorithm that outputs probabilities using the sigmoid function.

* Sigmoid 函數

中文：將任意實數輸入轉換為 0 到 1 間的機率值。

英文：The sigmoid function maps any real-valued input to a probability between 0 and 1.

* 最大概似估計（Maximum Likelihood Estimation, MLE）

中文：估計使觀察資料最可能發生的參數值，常用於邏輯回歸。

英文：A method to estimate parameters that maximize the likelihood of the observed data; commonly used in logistic regression.

* K最近鄰分類（K-Nearest Neighbors, KNN）

中文：非參數模型，根據鄰近樣本進行分類。

英文：A non-parametric model that classifies based on the labels of the k closest training samples.

* Confusion Matrix

中文：分類問題中常用的效能評估表格，列出 TP、TN、FP、FN。

英文：A table used to evaluate classification performance by listing TP, TN, FP, FN outcomes.

* 準確率（Accuracy）

中文：正確預測數占總預測數比例。

英文：The proportion of correct predictions over total predictions.

* 精確率（Precision）

中文：正確為正例的預測占所有預測正例的比例：TP / (TP + FP)。

英文：Precision is the proportion of true positives among all predicted positives: TP / (TP + FP).

* 召回率（Recall）

中文：正確為正例的預測占所有實際正例的比例：TP / (TP + FN)。

英文：Recall is the proportion of true positives among all actual positives: TP / (TP + FN).

* F1 分數（F1 Score）

中文：Precision 與 Recall 的調和平均值：2 \* (P \* R) / (P + R)。

英文：The harmonic mean of precision and recall: 2 \* (P \* R) / (P + R).

* AUC-ROC

中文：ROC 曲線下的面積，越接近 1 表示模型越好。

英文：Area under the ROC curve, where higher values (closer to 1) indicate better classification ability.

## 📚 名詞比較（Term Comparison）

1. 📌 Logistic Regression vs. K-Nearest Neighbors

中文：邏輯回歸適用於線性邊界問題，KNN 無需訓練，適用於非線性。

英文：Logistic regression works well for linear boundaries; KNN requires no training and adapts to nonlinear patterns.

1. 📌 K-Fold vs. Stratified K-Fold vs. LOOCV vs. Time-Series CV

中文：K-Fold 平均分資料；Stratified 維持類別比例；LOOCV 留一筆資料；Time-Series 保留時間順序。

英文：K-Fold splits data randomly; Stratified K-Fold preserves class ratios; LOOCV uses one point at a time; Time-Series CV keeps temporal order.

📘 第六章：分類（Classification）

本章介紹分類任務中的基本概念與演算法，包括貝氏定理、邏輯回歸、KNN分類、模型訓練方法（GD, SGD）與效能評估指標（如Confusion Matrix、Precision、Recall、F1、AUC-ROC）等。

🧠 名詞解釋（Term Explanation）

分類任務（Classification Problem）  
根據已標記資料訓練模型，以預測輸入屬於哪一類別。

貝氏定理（Bayes' Theorem）  
計算後驗機率的公式：P(y|x) = P(y)P(x|y) / P(x)，更新對事件的信念。

邏輯回歸（Logistic Regression）  
用於二元分類的演算法，透過sigmoid函數輸出機率值。

Sigmoid函數  
將任意實數輸入轉換為0到1之間的機率值：1 / (1 + e^(-x))。

最大概似估計（MLE）  
估計使觀察資料最可能發生的參數值，常用於邏輯回歸損失函數中。

K最近鄰分類（KNN）  
非參數模型，根據K個最接近鄰居的標籤進行預測。

Confusion Matrix  
分類問題中常用的效能評估表格，列出TP、TN、FP、FN。

準確率（Accuracy）  
正確預測數占總預測數比例。

精確率（Precision）  
正確為正例的預測占所有預測正例的比例：TP / (TP + FP)。

召回率（Recall）  
正確為正例的預測占所有實際正例的比例：TP / (TP + FN)。

F1分數（F1 Score）  
Precision 與 Recall 的調和平均值：2 \* (Precision \* Recall) / (Precision + Recall)。

AUC-ROC  
ROC曲線下的面積，代表模型分類能力。值越接近1越好，0.5為隨機猜測。

📚 名詞比較（Term Comparison）

Logistic Regression vs. K-Nearest Neighbors

- Logistic Regression:  
 • 適用於線性可分的二元分類  
 • 使用 Sigmoid 函數輸出機率  
 • 有模型參數，可用 GD/SGD 訓練  
- KNN:  
 • 適用於非線性資料  
 • 無模型訓練階段，預測需計算距離  
 • 可用於分類或回歸

K-Fold vs. Stratified K-Fold vs. LOOCV vs. Time Series CV

* - K-Fold:  
   • 隨機等分資料  
   • 適用於一般任務  
  - Stratified K-Fold:  
   • 保留分類比例，適合分類任務  
  - LOOCV:  
   • 每次只留一筆資料做驗證，偏差低但計算成本高  
  - Time Series CV:  
   • 適用於時間序列資料，保留順序避免資料洩漏

## 📝 模擬出題與答案（題目與答案寫在一起）

Q1. (a)(b) 題組：貝氏定理與後驗機率

(a) What is Bayes' theorem and how is it used in classification?  
Answer: Bayes' theorem updates the probability of a class label based on evidence. It is used in Naive Bayes classification to compute posterior probability.  
（a）什麼是貝氏定理？它在分類問題中如何使用？  
答：貝氏定理透過觀察資料更新類別機率，Naive Bayes 模型即根據此計算後驗機率。  
  
(b) What is the difference between prior and posterior probability?  
Answer: Prior is the belief before seeing data; posterior is the updated belief after observing data.  
（b）請比較先驗機率與後驗機率的差別。  
答：先驗機率為觀察前的假設，後驗機率則是考慮新資料後更新的結果。

Q2. (a)(b) 題組：邏輯回歸與損失函數

(a) Why do we use the sigmoid function in logistic regression?  
Answer: It maps real-valued input into a 0-1 probability range, suitable for binary classification.  
（a）為何在邏輯回歸中使用 Sigmoid 函數？  
答：Sigmoid 函數將任意數轉換為 0 到 1 的機率值，適合用於二元分類。  
  
(b) What loss function is commonly used in logistic regression?  
Answer: The negative log-likelihood, also known as cross-entropy loss, is commonly used.  
（b）邏輯回歸常使用哪種損失函數？  
答：常用負對數概似函數（交叉熵）來衡量預測誤差。

Q3. (a)(b) 題組：KNN原理與參數選擇

(a) What does the parameter k represent in KNN?  
Answer: k represents the number of nearest neighbors used to determine the class of a new sample.  
（a）KNN 中的 k 代表什麼？  
答：k 代表用來決定新樣本所屬類別的鄰近點數量。  
  
(b) What happens if k is too small or too large?  
Answer: Too small k may cause overfitting, too large k may smooth out decision boundaries too much.  
（b）若 k 太小或太大會發生什麼事？  
答：k 太小易過擬合，太大會讓模型失去區別能力。

Q4. (a)(b) 題組：分類效能評估指標

(a) Explain precision, recall, and F1-score.  
Answer: Precision measures accuracy of positive predictions. Recall measures how many actual positives are captured. F1 combines both.  
（a）請解釋 precision、recall 和 F1-score。  
答：Precision 是預測為正確的比例；Recall 是找出所有實際正例的能力；F1 是兩者的調和平均。  
  
(b) Why might accuracy be misleading in imbalanced datasets?  
Answer: Accuracy can be high even if the model predicts only the majority class. Precision/Recall are more informative.  
（b）為何在類別不平衡的情況下 Accuracy 可能誤導？  
答：即使模型只猜多數類別也可能有高 Accuracy，因此 Precision 與 Recall 更具參考價值。

Q5. (a)(b) 題組：Sigmoid 函數與分類應用（圖：sigmoid\_curve.png）

(a) What does the sigmoid curve represent in logistic regression?  
Answer: It shows how input features are transformed into probabilities between 0 and 1 for classification tasks.  
（a）Sigmoid 曲線在邏輯回歸中代表什麼意義？  
答：它將模型輸出轉換為介於 0 與 1 間的機率值，幫助進行二元分類。  
  
(b) Why is the output of sigmoid bounded and differentiable important?  
Answer: The bounded output helps interpret as probabilities; differentiability is needed for gradient descent optimization.  
（b）為什麼 sigmoid 的輸出需具備界限性與可微分性？  
答：界限性讓結果可解釋為機率；可微分性則便於使用梯度下降進行優化。

圖 Q5：Sigmoid Function in Logistic Regression

