

NotebookLM 研究

NotebookLM 研究摘要：多元線性回歸分析之主流方法與優化建議

主題： 多元線性回歸分析 (Multiple Linear Regression Analysis)

摘要：

多元線性回歸是統計學與機器學習中的基礎預測模型，旨在探索多個自變數（特徵）與單一應變數（目標）之間的線性關係。其核心方法為「普通最小平方法」（Ordinary Least Squares, OLS），目標是最小化預測誤差的平方和。然而，在現實世界的複雜數據中，標準的 OLS 模型往往需要進一步優化才能獲得可靠的結果。

主流進階方法：

為了解決共線性和過擬合問題，主流方法引入了「正規化」(Regularization) 技術：

1. **嶺迴歸 (Ridge Regression, L2):** 透過在損失函數中加入係數的 L2 範數懲罰項，來縮減模型係數，有效處理特徵間高度相關（多重共線性）的問題。
2. **套索迴歸 (Lasso Regression, L1):** 使用 L1 範數作為懲罰項，其特性是能將不重要的特徵係數精確地縮減至零，因此常被用作一種內嵌的特徵選擇方法。
3. **彈性網路 (ElasticNet):** 結合了 Ridge 和 Lasso 的優點，同時處理多重共線性和進行特徵選擇，在特徵數量眾多時特別有效。

模型優化建議：

當線性模型表現不佳時（如本專案 R^2 分數偏低），可從以下方向尋求改進：

- **處理非線性關係:** 現實關係很少是純線性的。可以透過「多項式迴歸」(Polynomial Regression) 增加特徵的高次項（如 x^2 ），或對特徵進行對數 (log)、平方根等數學轉換，以捕捉曲線關係。
- **特徵工程 (Feature Engineering):** 創造新的、更有預測能力的特徵。例如，根據業務理解，建立「交互作用項」（如 `alcohol * sulphates`），以捕捉特徵間的協同效應。
- **使用穩健迴歸 (Robust Regression):** 如果數據中存在大量離群值，可以改用對離群值不敏感的模型，如 `HuberRegressor` 或 `RANSAC`，以降低極端值對模型的扭曲。
- **重新定義問題:** 如評估階段所建議，若迴歸效果不彰，可嘗試將問題從「預測精確數值」轉換為「預測等級」，改用分類模型（如邏輯迴歸、決策樹）來處理。