1. **Overfitting là gì?**

* Overfitting (quá khớp) là một hành vi không mong muốn xảy ra khi mô hình học máy đưa ra dự đoán chính xác cho dữ liệu huấn luyện nhưng lại có hiệu suất kém trên dữ liệu mới. Khi sử dụng mô hình học máy để đưa ra dự đoán, trước tiên họ huấn luyện mô hình trên một tập dữ liệu đã biết. Sau đó, dựa trên thông tin này, mô hình cố gắng dự đoán kết quả cho các tập dữ liệu mới. Một mô hình Overfit có thể đưa ra dự đoán không chính xác và có sai số cao trên các dữ liệu mới.

1. **Nguyên nhân dẫn đến Overfitting?**

* Kích thước dữ liệu quá nhỏ và không chứa đủ mẫu dữ liệu để thể hiện chính xác tất cả dữ liệu đầu vào khả thi.
* Dữ liệu đào tạo chứa một lượng lớn thông tin không liên quan, được gọi là dữ liệu nhiễu.
* Mô hình đào tạo quá lâu trên một tập dữ liệu mẫu duy nhất, dẫn đến hiện tượng ghi nhớ dữ liệu.
* Do có độ phức tạp cao, mô hình học cả phần nhiễu trong dữ liệu đào tạo.

1. **Làm thế nào để phát hiện Overfitting?**

* Phát hiện overfitting là một bước quan trọng để đảm bảo mô hình học máy có thể tổng quát hóa tốt trên dữ liệu mới, thay vì chỉ hoạt động tốt trên dữ liệu huấn luyện. Có rất nhiều phương pháp để phát hiện Overfitting như: So sánh độ chính xác (accurary) hoặc sai số trên tập huấn luyện và tập dữ liệu kiểm tra (tập dữ liệu mới), Quan sát đồ thị learning curve (đường học),… Nhưng một trong những phương pháp mạnh mẽ nhất được sử dụng trong thực tế để phát hiện hiện tượng Overfitting, đặc biệt khi dữ liệu hạn chế và bạn muốn tận dụng tối đa toàn bộ dữ liệu là phương pháp **K-Fold Cross-Validation** (Xác thực chéo K lần).
* Cách **K-Fold Cross-Validation** (Xác thực chéo K lần) hoạt động.
  + **Chia dữ liệu:** Chia tập dữ liệu đào tạo thành K tập con hoặc tập mẫu có kích thước bằng nhau, được gọi là nhóm.
  + **Huấn luyện và kiểm tra mô hình:** Giữ một tập con làm dữ liệu xác thực (tập A) và đào tạo mô hình máy học trên K-1 tập con còn lại. Mô hình sau đó được đánh giá trên tập A này, và kết quả được lưu lại.
  + **Lặp lại quá trình:** Quá trình này được lặp lại K lần, mỗi lần sử dụng một tập dữ liệu khác nhau làm tập kiểm tra và K-1 tập con khác nhau làm tập huấn luyện.
  + **Tính toán kết quả trung bình:** Sau K lần lặp, kết quả đánh giá (thường là accuracy, F1-score, hoặc sai số) được trung bình hóa để có được một đánh giá cuối cùng về hiệu suất của mô hình.

1. **Các phương pháp giải quyết Overfitting ở câu 1.**

* **Logistic Regression**:
  + **Regularization (Chuẩn hóa)**: Logistic Regression hỗ trợ các chuẩn hóa như L2 (Ridge) hoặc L1 (Lasso), giúp giảm độ phức tạp của mô hình bằng cách ràng buộc các hệ số của đặc trưng về gần 0. Sử dụng tham số penalty và C (điều chỉnh mức độ regularization) để điều chỉnh mức phạt trong Logistic Regression.
* **Random Forest (cho cả Classification và Regression)**:
  + **Giảm độ sâu của cây**: Giới hạn độ sâu tối đa của mỗi cây bằng cách điều chỉnh tham số max\_depth để tránh cây phát triển quá phức tạp.
  + **Giảm số lượng đặc trưng khi phân chia**: Giới hạn số lượng đặc trưng sử dụng tại mỗi nút bằng tham số max\_features.
  + **Số lượng cây**: Tăng số lượng cây (n\_estimators) có thể giúp mô hình học tổng quát hơn.
* **SVM (Support Vector Machine)**:
  + **Regularization (C parameter)**: Tham số C điều chỉnh mức độ penalize đối với các sai lệch. Nếu C quá lớn, mô hình có xu hướng overfit, trong khi giá trị C nhỏ sẽ giúp mô hình đơn giản hơn.
  + **Kernel Trick**: Chọn kernel phù hợp với dữ liệu (ví dụ: linear, rbf, hoặc poly).
* **SVR và Linear Regression (cho Regression)**:
  + **Regularization cho SVR**: Tham số C trong SVR điều chỉnh độ mềm dẻo của biên, giúp giảm overfitting.
  + **Regularization cho Linear Regression**: Ridge (L2) và Lasso (L1) Regression có thể được sử dụng để thêm regularization cho Linear Regression.

1. **Từ kết quả ở phần code, ta có thể rút ra một số nhận xét sau:**

* **Classification Accuracy**
  + **Trước khi áp dụng regularization:** Các mô hình như Random Forest và SVM có độ chính xác cao hơn so với Logistic Regression. Đặc biệt, Random Forest có độ chính xác cao nhất là 0.89.
  + **Sau khi áp dụng regularization:** Độ chính xác của các mô hình đều giảm, đặc biệt là với Logistic Regression và SVM. Random Forest với độ sâu giới hạn cũng giảm độ chính xác xuống còn 0.74.
  + **Nhận xét:** Việc áp dụng regularization trong các mô hình phân loại đã làm giảm overfitting, nhưng cũng dẫn đến việc giảm độ chính xác trên tập kiểm tra. Điều này có thể là do regularization làm giảm tính phức tạp của mô hình, giúp mô hình tổng quát tốt hơn trên dữ liệu mới nhưng giảm khả năng dự đoán chính xác với dữ liệu huấn luyện.
* **Mean Squared Error (MSE) trong Regression**
  + **Trước khi áp dụng regularization:** Random Forest có MSE thấp nhất (0.12), cho thấy khả năng dự đoán khá tốt. Các mô hình Linear Regression và SVR có MSE cao hơn, khoảng 0.45 - 0.47.
  + **Sau khi áp dụng regularization:** MSE của các mô hình như Ridge Regression và SVR chỉ giảm nhẹ. Với Random Forest Regressor khi giới hạn độ sâu, MSE tăng lên 0.30, nhưng vẫn duy trì ở mức khá thấp so với các mô hình khác. Lasso Regression có MSE cao nhất (0.63), cho thấy nó không phù hợp tốt với dữ liệu này sau khi áp dụng regularization.
  + **Nhận xét:** Regularization giúp giảm khả năng overfitting trong các mô hình hồi quy, tuy nhiên mức giảm MSE không quá lớn. Các mô hình đơn giản hơn như Ridge Regression và Lasso Regression có khả năng tổng quát tốt hơn, nhưng không đạt được hiệu suất cao như Random Forest.

Tham khảo từ:

<https://bachtuan91.wordpress.com/2018/12/25/overfitting-and-regularized-doi-voi-hoi-quy-tuyen-tinh-va-hoi-quy-logistic/>

<https://aws.amazon.com/vi/what-is/overfitting/>

<https://cuonglv1109.blogspot.com/2018/11/overfitting-la-gi.html>