**Báo cáo Kỹ thuật: Các tham số ảnh hưởng đến mô hình học máy trong dự đoán thời tiết**

**1. Giới thiệu**

Trong quá trình huấn luyện và dự đoán của các mô hình học máy, có rất nhiều tham số ảnh hưởng đến hiệu quả và độ chính xác của mô hình. Dưới đây là phân tích các tham số ảnh hưởng đến ba mô hình chính:

**2. Mô hình Cây quyết định (Decision Tree - ID3 Algorithm)**

**2.1 Các tham số ảnh hưởng**

* **Max Depth (Độ sâu tối đa)**:
  + Quyết định số mức (level) tối đa mà cây có thể đạt được.
  + Nếu **max depth** quá nhỏ: cây có thể không học đủ các mối quan hệ trong dữ liệu, dẫn đến underfitting.
  + Nếu **max depth** quá lớn: cây có thể học quá nhiều chi tiết, dễ dẫn đến overfitting.
* **Criterion (Tiêu chí)**:
  + Được sử dụng để đo lường chất lượng của mỗi phép chia. Ở đây, chúng ta sử dụng **entropy** để tính toán mức độ hỗn loạn (impurity).
* **Min Samples Split (Số mẫu tối thiểu để tách)**:
  + Số lượng mẫu tối thiểu cần thiết để tách một nút bên trong.
  + Nếu giá trị này quá thấp, cây sẽ quá phức tạp; nếu quá cao, cây có thể bị đơn giản hóa quá mức.

**2.2 Cách chọn tham số phù hợp**

* Tham khảo các mô hình đã được chứng minh và thử nghiệm.
* Dùng phương pháp **cross-validation** để kiểm tra hiệu quả mô hình với nhiều tập dữ liệu khác nhau và tránh overfitting.

**3. Mô hình Mạng nơ ron nhân tạo (Neural Network - MLP Classifier)**

**3.1 Các tham số ảnh hưởng**

* **Hidden Layers (Số lớp ẩn)**:
  + Quyết định độ sâu và khả năng học các đặc trưng phức tạp của mô hình.
  + Số lớp ẩn quá ít: mạng nơ ron khó học các đặc điểm phức tạp.
  + Số lớp ẩn quá nhiều: mô hình dễ dẫn đến overfitting.
* **Neurons per Layer (Số lượng nơ ron mỗi lớp)**:
  + Nếu quá ít, mô hình không thể học được các mối quan hệ trong dữ liệu; nếu quá nhiều, mô hình sẽ quá phức tạp và dẫn đến overfitting.
* **Learning Rate (Tốc độ học)**:
  + Quyết định mức độ thay đổi trọng số sau mỗi lần cập nhật.
  + Nếu **learning rate** quá nhỏ: mô hình sẽ cần nhiều thời gian để học.
  + Nếu **learning rate** quá lớn: mô hình có thể bỏ qua điểm tối ưu và không hội tụ được.
* **Iterations (Số lần lặp)**:
  + Số lần mô hình cập nhật trọng số trong quá trình huấn luyện.
  + Iterations quá nhỏ: mô hình chưa kịp hội tụ.
  + Iterations quá lớn: dễ dẫn đến overfitting.

**3.2 Cách chọn tham số phù hợp**

* Dùng phương pháp **Grid Search** để thử nghiệm nhiều giá trị của các tham số khác nhau.
* Thử nghiệm với các tập giá trị **learning rate** từ [0.01, 0.1] và **hidden layers** khác nhau để tìm ra sự kết hợp tối ưu.

**4. Mô hình Perceptron**

**4.1 Các tham số ảnh hưởng**

* **Learning Rate (Tốc độ học)**:
  + Ảnh hưởng đến mức độ điều chỉnh trọng số của mô hình sau mỗi lần cập nhật.
  + Tương tự như Neural Network, nếu **learning rate** quá cao: mô hình có thể học quá nhanh và bỏ qua điểm tối ưu.
  + Nếu **learning rate** quá thấp: mô hình học chậm và mất nhiều thời gian để đạt được hội tụ.
* **Max Iterations (Số lần lặp tối đa)**:
  + Số vòng lặp để mô hình cập nhật trọng số cho đến khi hội tụ.
  + Iterations quá ít: mô hình có thể không học đủ.
  + Iterations quá nhiều: dẫn đến overfitting.

**4.2 Cách chọn tham số phù hợp**

* Sử dụng **cross-validation** để kiểm tra các giá trị khác nhau của **learning rate** và **iterations** để tránh overfitting và underfitting.

**5. Kết luận**

Các tham số trong mỗi mô hình đều có ảnh hưởng rất lớn đến hiệu suất và tính chính xác của mô hình. Việc lựa chọn các tham số phù hợp yêu cầu quá trình thử nghiệm và đánh giá kỹ lưỡng trên nhiều tập dữ liệu khác nhau. Grid Search và Cross-Validation là các kỹ thuật phổ biến giúp chọn ra bộ tham số tối ưu nhất cho từng mô hình.