TRƯỜNG ĐẠI HỌC THỦY LỢI

**KHOA CÔNG NGHỆ THÔNG TIN**

A blue and white logo

Description automatically generated with low confidence

**BÀI TẬP LỚN**

HỌC PHẦN: HỌC MÁY

**ĐỀ TÀI: DỰ ĐOÁN THỜI TIẾT**

**Phần 1: Lý thuyết về Mạng Nơ-ron Nhân tạo (Neural Network)**

**a. Mạng nơ-ron nhân tạo là gì?**

Mạng nơ-ron nhân tạo (Artificial Neural Network - ANN) là một mô hình mô phỏng cách mà hệ thống thần kinh con người hoạt động. ANN gồm các nơ-ron kết nối với nhau thông qua các trọng số. Khi một dữ liệu được đưa vào, nó được lan truyền qua các lớp nơ-ron, qua quá trình huấn luyện, mô hình học cách phân loại và dự đoán dữ liệu mới.

**b. Các thành phần chính của một mạng nơ-ron:**

1. **Lớp đầu vào (Input Layer):**
   * Chứa các nơ-ron đầu vào, mỗi nơ-ron tương ứng với một đặc trưng (feature) từ dữ liệu. Trong bài toán dự đoán thời tiết, các đặc trưng bao gồm lượng mưa (*precipitation*), nhiệt độ cao nhất (*temp\_max*), nhiệt độ thấp nhất (*temp\_min*), và sức gió (*wind*).
2. **Lớp ẩn (Hidden Layer):**
   * Mạng nơ-ron có thể có một hoặc nhiều lớp ẩn, mỗi lớp bao gồm nhiều nơ-ron. Các nơ-ron ở lớp ẩn được kết nối với các nơ-ron ở lớp đầu vào và lớp đầu ra. Chúng xử lý thông tin và tạo ra các đặc trưng mới dựa trên trọng số liên kết giữa các nơ-ron.
3. **Lớp đầu ra (Output Layer):**
   * Dự đoán kết quả cuối cùng của bài toán dựa trên các thông tin được xử lý từ lớp ẩn. Trong bài toán này, đầu ra là một trong các trạng thái thời tiết như *sun*, *rain*, *drizzle*, *snow*, hoặc *fog*.

**c. Hàm kích hoạt (Activation Function):**

* Sigmoid: Được sử dụng chủ yếu trong các bài toán phân loại nhị phân.
* ReLU (Rectified Linear Unit): Thường được sử dụng trong các mạng nơ-ron có nhiều lớp vì khả năng khắc phục hiện tượng mất mát gradient trong quá trình huấn luyện.
* Softmax: Dùng trong các bài toán phân loại đa lớp như bài toán dự đoán thời tiết này.

**d. Hàm mất mát (Loss Function):**

* Hàm mất mát là hàm giúp đo lường sự khác biệt giữa giá trị dự đoán và giá trị thực tế.
* Đối với bài toán phân loại, đặc biệt là phân loại đa lớp, hàm mất mát thường được sử dụng là *Cross-Entropy Loss*. Hàm này giúp tối ưu hóa mạng bằng cách giảm thiểu sự khác biệt giữa nhãn dự đoán và nhãn thực tế.

**e. Lan truyền ngược (Backpropagation):**

* Đây là quá trình tối ưu hóa mạng nơ-ron, giúp mạng điều chỉnh trọng số của các liên kết giữa các nơ-ron dựa trên giá trị của hàm mất mát. Lan truyền ngược sử dụng thuật toán Gradient Descent để điều chỉnh trọng số nhằm giảm thiểu hàm mất mát sau mỗi lần huấn luyện.

**f. Điều kiện dừng của quá trình huấn luyện:**

* Quá trình huấn luyện sẽ dừng khi một trong các điều kiện sau được thỏa mãn:
  + Đã huấn luyện đủ số lần lặp (epochs) được chỉ định.
  + Đạt được độ chính xác mong muốn trên tập kiểm tra.
  + Tốc độ học (learning rate) quá nhỏ, khiến mô hình không cải thiện thêm nữa.