ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN MÔN HỌC

TOÁN ỨNG DỤNG VÀ THỐNG KÊ

ĐỀ TÀI: Bài toán khí hậu
Bài toán 2

Giảng viên lý thuyết: PGS.TS Nguyễn Đình Thúc
Lớp: 20TN
Thành viên thực hiện:

- 20120131 – Nguyễn Văn Lộc
- 20120536 – Võ Trọng Nghĩa
- 20120572 – Nguyễn Kiều Minh Tâm

THÀNH PHỐ HỒ CHÍ MINH, THÁNG 3-4 NĂM 2022

Mục lục

Danh sách hình vẽ

Danh sách bảng

Lời nói đầu

1 Đặt vấn đề

1.1 Xác định và hình thức hóa mục tiêu của bài toán

Bài toán: Dự báo nhiệt độ trung bình năm ở 5 trạm khí tượng theo dữ liệu của Cơ quan Quản lý Khí quyển và Đại dương Quốc gia Hoa Kỳ, bao gồm:

- Trạm Lieksa Lampela ở Phần Lan (FIE00144982).
- Trạm Jena Sternwarte ở Đức (GM000004204).
- Trạm Den Helder 1 ở Hà Lan (NLM00006235).
- Trạm Uppsala Aut ở Thụy Điển (SWE00139148).
- Trạm New York City's Central Park ở Hoa Kỳ (USW00094728).

Hình thức hóa mục tiêu của bài toán: dùng biểu đồ phân tán, tìm các hệ số hồi quy.

1.2 Dạng bài toán

Bài toán này thuộc dạng bài toán **Dự đoán** (từ dữ liệu hiện có trong hiện tại và quá khứ để đưa ra dự đoán dữ liệu trong tương lai chưa biết).

1.3 Đối tượng được chọn cho bài toán

Dữ liệu về nhiệt độ trung bình năm của 5 trạm nêu trên.
Cột TAVG của các tập tin **FIE00144982.csv**, **GM000004204.csv**, **NLM00006235.csv**, **SWE00139148.csv**, **USW00094728.csv** trong thư mục **gsoy-latest**, lấy từ dataset về Global Summary of the Year của NOAA.

1.4 Phạm vi, mức độ, quy mô của bài toán

Theo không gian: dữ liệu được xử lý trong bài toán đại diện cho 5 bang/tỉnh ở 5 quốc gia khác nhau.
Theo thời gian: dữ liệu được thu thập trong vòng 30 năm, từ năm 1991 đến năm 2020.

2 Thu thập và xử lý dữ liệu

2.1 Thu thập dữ liệu

Dữ liệu xử lý trong bài toán này được thu thập từ dữ liệu của Cơ quan Quản lý Khí quyển và Đại dương Quốc gia Hoa Kỳ (NOAA), phần dữ liệu tổng hợp theo năm (Global Summary of the Year).

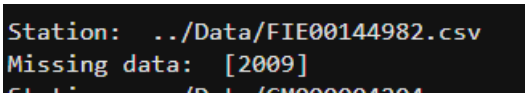
2.2 Xử lý dữ liệu

2.2.1 Trích xuất dữ liệu

Dữ liệu nhiệt độ trung bình năm từ các tập tin được trích xuất thành các tập dữ liệu **data0.csv**, **data1.csv**, **data2.csv**, **data3.csv**, **data4.csv**, theo thứ tự được nêu ra bên trên nhờ các hàm của module `pandas`.

2.2.2 Xử lý phần dữ liệu bị khuyết

Dữ liệu từ tập tin gốc đầu tiên (tập tin **FIE00144982.csv**) bị khuyết phần dữ liệu của năm 2009., do đó, nhóm chúng em đã tìm cách "điền" vào những ô bị khuyết này. Nhiệt độ trung bình năm của năm 2009 trong tập tin này được tính bằng trung bình cộng của nhiệt độ trung bình năm 2008 và nhiệt độ trung bình năm 2010.



Hình 1: Phần dữ liệu bị khuyết

3 Phân tích, đánh giá và kết luận

Sau khi được tiền xử lý, dữ liệu được trực quan hóa (visualized)thành biểu đồ phân tán nhờ hàm hỗ trợ của module `matplotlib`.