HỌC PHẦN: NGÔN NGỮ LẬP TRÌNH KHOA HỌC
PHIẾU GIAO BÀI TẬP SỐ 2
(Assignment 2)

Mục tiêu:

- **Vận dụng kiến thức/ kỹ năng tổng hợp của học phần để giải quyết các bài toán tính toán khoa học trong thực tế.**
- **Thực hiện được thiết kế, cài đặt, kiểm thử và vận hành chương trình trên máy tính**

☞ *Bạn cần hoàn thành bài tập này trong vòng 60-120 phút một cách độc lập. Nếu không thể, xin ghi rõ thời gian làm bài của bạn và ghi rõ bạn có sử dụng sự trợ giúp nào trong quá trình làm bài hay không.*

☞ *File dữ liệu: kddcup.data và kddcup.test, có link download kèm theo.*

☞ *Kỹ thuật mô phỏng: Phân lớp dữ liệu bằng phương pháp k-láng giềng gần nhất với k = 1 (Data classification by k-nearest neigbors, k =1)*

ĐỀ BÀI

Cho hai bộ dữ liệu trích 5% từ bộ KDDCUP99 gồm 2 file: kddcup.data và kddcup.test. Viết các hàm sau:

- Hàm **read**: đọc toàn bộ dữ liệu của 1 tệp lên mảng, trả về mảng đọc được.
- Hàm **distance**: tính khoảng cách Euclidean giữa hai dòng bất kỳ của x.
- Hàm **distancetoall**: tính khoảng cách Euclidean từ 1 dòng test tới tất cả các dòng của x.

Chương trình:

[1]. Đọc toàn bộ tệp kddcup.data lên mảng x_train; cắt bỏ 4 cột đầu tiên và cắt cột cuối cùng của x_train ra một mảng *label*; chuyển x_train về kiểu float.

[2]. Đọc toàn bộ file kddcup.test lên mảng x_test; cắt bỏ 4 cột đầu và cột cuối cùng của x_test rồi chuyển mảng về kiểu thực.

[3]. Lấy 1 dòng (đặt tên là *test*) bất kỳ từ mảng x_test hãy dự đoán *label* của nó bằng cách: tính khoảng cách từ nó tới tất cả các dòng của x_train. Label của *test* là label của dòng gần với nó nhất (sinh viên có thể tổ chức thành hàm **predict** với đầu vào là mảng các khoảng cách - *dis* và mảng *label*)

Hình thức nộp bài:	Trực tuyến
Định dạng file khi nộp:	Sinh viên nộp file HOVATEN.py.
Thời gian hoàn thành:	01 tuần kể từ ngày giao bài.
Hỗ trợ:	Sinh viên có thể thảo luận với giảng viên hoặc sinh viên khác, nhưng làm bài độc lập.