TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM

ĐỀ THI KẾT THÚC HỌC PHẦN
Học kỳ 1 – Năm học 2023-2024

(do phòng KT-ĐBCL ghi)
ĐA - CK2324-1
CSC14119

Tên học phần:	**Nhập môn Khoa học Dữ liệu**	**Mã HP:**	**CSC14119**
Thời gian làm bài:	**90 phút**	Ngày thi:	29/12/2023

Ghi chú: *Sinh viên được phép sử dụng tài liệu khi làm bài (không thiết bị thu phát sóng).*

Câu 1 (*1 điểm*).

Làm thế nào bạn sẽ tiếp cận việc phân tích dữ liệu từ một bộ dữ liệu lớn và phức tạp? Ngoài ra, những môn học nào trong lộ trình của Khoa học Dữ liệu sẽ cung cấp kiến thức và công cụ cần thiết để thực hiện phân tích này?

Câu 2 (*2.5 điểm*).

Khi tham gia lớp học của Giáo sư Sprout, Harry Potter được yêu cầu xem xét sự phát triển của cây nhân sâm để chữa lành cho những người bị con rắn của chúa tể Voldermort hóa đá. Danh sách sau là chiều cao (mm) của 30 mẫu cây khi đào lên trong số 1000 cây trong vườn:

180, 206, 200, 206, 202, 205, 205, 203, 200, 202, 203, 201, 205, 201, 199, 199, 203, 204, 208, 203, 204, 203, 201, 200, 223, 202, 204, 205, 204, 202

a) Tính toán các giá trị thống kê về tính trung tâm và tính phân bố.
b) Xác định tứ phân vị thứ nhất (Q1) và tứ phân vị thứ 3 (Q3) của dữ liệu.
c) Vẽ lược đồ histogram và dựa trên đó để xác định các điểm ngoại lai (outlier) đi kèm với lý giải tóm tắt.
d) Trực quan hóa dữ liệu bằng box-plot sau khi đã loại bỏ các điểm ngoại lai.
e) Một cây cần đạt chiều cao từ 204 mm trở lên mới có thể làm thuốc được, nếu bạn là Harry Potter và GS Sprout hỏi bạn liệu có thu hoạch được chưa vì thời gian khá gấp rút rồi. Bạn cần lý giải cho nhận định của mình.

Câu 3 (*3 điểm*).

Tỷ lệ phần trăm theo năm của nữ giới làm công ăn lương được đưa ra trong bảng sau:

Năm	Tỉ lệ phần trăm	Năm	Tỉ lệ phần trăm
1979	61.2	**1985**	61.8
1980	60.7	**1986**	62.0
1981	61.3	**1987**	62.7
1982	61.3	**1990**	62.8
1983	61.8	**1992**	62.9
1984	61.7		

a) Xem "năm" là biến độc lập và "tỉ lệ phần trăm" là biến phụ thuộc, hãy vẽ biểu đồ phân tán (scatter plot) của dữ liệu.

b) Qua biểu đồ, liệu có tồn tại mối quan hệ giữa hai biến này không? Giải thích.

c) Giả sử dữ liệu tuân theo đường hồi quy tuyến tính có dạng: $y = \beta_1 x + \beta_2$.
Hàm lỗi sử dụng MSE (mean-squared error).
Hãy sử dụng phương pháp đạo hàm để tìm đường hồi quy của tập dữ liệu trên. Trực quan sau khi đã thực hiện.

d) Dựa trên đường hồi quy vừa tìm được, xác định dự đoán cho năm 2050. Giá trị dự đoán này có hợp lý không? Tại sao?

e) Tính hệ số tương quan của hai biến (correlation coefficient). Biết rằng biểu thức hệ số tương quan có dạng:

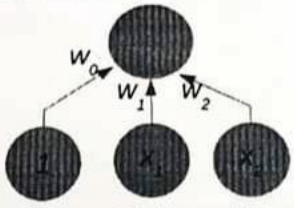
$$r = \frac{\sum xy - \frac{1}{n}\sum x \sum y}{\sqrt{(\sum x^2 - \frac{1}{n}(\sum x)^2)(\sum y^2 - \frac{1}{n}(\sum y)^2)}}$$

trong đó n là kích thước của tập dữ liệu, x và y lần lượt là giá trị các biến của các mẫu.
Nhận xét hệ số tương quan đã tính ra.

f) Có bất kỳ điểm dữ liệu nào là ngoại lệ (outlier) không? Giải thích.

Câu 4 (*3.5 điểm*).

Cho một mạng nơ-ron tuyến tính 1 lớp có cấu tạo như sau:



Hàm kích hoạt (activation function) có dạng: $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$

Trọng số được khởi tạo ban đầu lần lượt là $w_0 = 0, w_1 = 1, w_2 = 0.5$.

Các mẫu dữ liệu huấn luyện:

$$\mathbf{X} = \begin{bmatrix} 1 & 2 \\ -2 & 5 \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} 1 \\ 6 \\ 1 \end{bmatrix}$$

trong đó mỗi dòng trong **X** thể hiện một điểm dữ liệu và mỗi phần tử trong **y** thể hiện giá trị đầu ra mong muốn.

a) Sử dụng mạng nơ-ron trên để xác định cho mẫu dữ liệu mới $\mathbf{x} = [2\ 3]^T$.

b) Tính độ lỗi của mạng trên tập **X** theo công thức:

$$E = \frac{1}{2}\sum(\mathbf{y} - f(\mathbf{x}))^2$$

c) Thực hiện 1 lần cập nhật độ giảm độ dốc (gradient descent) sử dụng tập dữ liệu **X** và hàm lỗi E từ câu trên với hệ số học $\eta = 0.1$.