TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM

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

MÃ LƯU TRỮ
(do phòng KT-ĐBCL ghi)

Câu 4 (*2 điểm*).

Cho bảng thống kê các khách hàng ở từng quốc gia mua cây thông cho mùa giáng sinh năm 2021:

Customer Id	Country	Age	Salary	Purchased
Id_1	France	44	72000	No
Id_2	Spain	27	48000	Yes
Id_3	Germany	30	54000	No
Id_4	Spain		61000	No
Id_5	Germany	40		Yes
Id_6	France	35	58000	Yes
Id_7	Spain		52000	No
Id_8	France	48	79000	
Id_9	Germany	50		No
Id_10	France	37	67000	Yes
Id_11	Spain	49	623000	
Id_12		38	51000	Yes
Id_13	Germany		55000	No
Id_14	France	36	68000	No
Id_15	France	39		
Id_16	Spain	41	62000	Yes

France 6
Spain 5
Ger 4
Yes 6
No 7

a) Các ô không điền giá trị nghĩa là bị thiếu. Bạn hãy đề xuất cách giải quyết vấn đề này. Biết rằng bỏ mẫu hay đánh dấu các ô bị thiếu không phải là giải pháp khả thi.

b) Hãy sử dụng phương pháp chia giỏ theo độ sâu để loại bỏ các dữ liệu nhiễu trong cột tuổi (age) từ dữ liệu ở câu (a).

c) Hãy biến đổi dữ liệu để làm cho các mô hình học máy phân tích về sau có thể chạy nhanh và chính xác. Trình bày ít nhất 3 cách biến đổi, lý giải ngắn gọn và thực hiện.

Câu 5 (*2 điểm*).

Cho 10 mẫu dữ liệu như trong bảng sau:

	1	2	3	4	5	6	7	8	9	10
y	1.45	1.93	0.81	0.61	1.55	0.95	0.45	1.14	0.74	0.98
x_1	0.58	0.86	0.29	0.2	0.56	0.28	0.08	0.41	0.22	0.35
x_2	0.71	0.13	0.79	0.2	0.56	0.92	0.01	0.6	0.7	0.73

Trong đó x_1, x_2 là biến độc lập và y là biến phụ thuộc. Giả sử mô hình tuân theo hồi quy tuyến tính có dạng $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. Hàm mất mát là MSE.

a) Nếu chúng ta bắt đầu với $\beta_0 = 1, \beta_1 = 2$, và $\beta_2 = 3$ thì giá trị của hàm mất mát là bao nhiêu?

b) Tính giá trị của $\beta_0, \beta_1, \beta_2$ sau một lần lặp của gradient descent với tốc độ học $\eta = 1$.

-- *Hết* --

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM

ĐỀ THI KẾT THÚC HỌC PHẦN

Học kỳ 1 – Năm học 2023-2024

MÃ LƯU TRỮ
(do phòng KT-ĐBCL ghi)
CK2324_1
CSC14119

Tên học phần: Nhập môn Khoa học Dữ liệu **Mã HP:** CSC14119

Thời gian làm bài: **90 phút** Ngày thi: 11/01/2024

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

Câu 1 (*1.5 điểm*).

Mô tả những gì cần làm để thu thập dữ liệu (mẫu) nhằm khám phá những câu hỏi sau đây, giải thích ngắn gọn các bước thực hiện dựa trên các kiến thức liên quan đến KHDL.

a) Tỉ lệ người tăng trên 5 kg trong sáu tháng cuối năm 2021 do lệnh giới nghiêm trong đại dịch Covid ở Việt Nam?
b) Mức học sinh lớp 1 biết đọc sau học kì 1 khác nhau giữa năm 2020 và 2021?
c) Tỉ lệ sinh viên thích và không thích môn Nhập môn KHDL ở trường KHTN qua thời gian?

Câu 2 (*3 điểm*).

Một bài báo trên tạp chí Sound and Vibration đã công bố kết quả nghiên cứu về mối tương quan giữa mức tăng huyết áp và tiếng ồn (đơn vị dB) khi tiếp xúc. Dữ liệu sau đây được trình bày trong bài báo.

Mức tăng huyết áp	1	0	1	2	5	1	4	6	2	3
Mức âm thanh	60	63	65	70	70	70	80	90	80	80

Mức tăng huyết áp	5	4	6	8	4	5	7	9	7	6
Mức âm thanh	85	89	90	90	90	90	94	100	100	100

a) Biến nào là biến độc lập, biến nào là biến phụ thuộc? Tại sao?
b) Vẽ lược đồ phân tán (scatter plot) của hai biến trên. Dựa trên lược đồ để đánh giá xem dữ liệu có tuân theo mô hình tuyến tính không? Giải thích.
c) Giả sử mô hình tuân theo hồi quy tuyến tính có dạng $f(x) = \beta_1 x + \beta_0$
Hàm lỗi tuân theo MSE (mean-square error). Hãy chọn ngẫu nhiên một cặp giá trị (β_0, β_1) để tính độ lỗi.
d) Sử dụng phương pháp đạo hàm để khớp mô hình hồi quy tuyến tính với dữ liệu trên.
e) Dự đoán mức tăng huyết áp khi tiếng ồn đạt giá trị 85 dB. Theo bạn, kết quả có hợp lý không? Tại sao?

Câu 3 (*1.5 điểm*).

a) Gradient descent là gì? Tại sao phương pháp gradient descent và đạo hàm đều có thể dùng để khớp mô hình nhưng gradient descent lại được dùng phổ biến hơn? Có trường hợp nào phương pháp gradient descent không thể áp dụng? Ví dụ.
b) Nếu hàm tối ưu hóa là lồi nghiêm ngặt (strictly convex) thì gradient descent có luôn hội tụ không?
c) Phân biệt sự khác nhau giữa gradient descent, stochastic gradient descent và batch gradient descent? Điểm lợi và bất lợi của mỗi phương pháp.