BÀI TẬP 4
THỐNG KÊ MÁY TÍNH VÀ ỨNG DỤNG

Câu 1. (3.5 điểm) Dữ liệu của 2 đại lượng X, Y được cho trong bảng sau

X	1.11	0.00	0.47	0.23	0.14	0.29	0.53	0.61	0.83	0.65	1.05	0.31
Y	2.38	1.03	1.00	0.90	0.93	0.90	1.06	1.16	1.57	1.22	2.18	0.91

X	1.35	0.04	1.03	0.64	0.86	0.22	0.30	1.23	1.49	0.48	1.07	1.35
Y	3.32	0.99	2.12	1.21	1.65	0.90	0.91	2.82	3.98	1.01	2.25	3.32

Phần I. Biết rằng $X \sim \mathcal{U}(0, \theta)$, ta có thể dùng các ước lượng sau cho θ

$$T_1 = 2\bar{X}, \quad T_2 = 2\hat{m}, \quad T_3 = 2\sqrt{3}S, \quad T_4 = \max\{X_1, X_2, \dots, X_n\}$$

với $\bar{X}, \hat{m}, S, \max$ lần lượt là trung bình, trung vị, độ lệch chuẩn và giá trị lớn nhất của mẫu.

a) Tính các giá trị ước lượng T_1, T_2, T_3, T_4 cho θ từ mẫu dữ liệu đã cho.
b) Dùng kĩ thuật bootstrapping, so sánh sai số chuẩn của các ước lượng trên.
c) Giả sử ta có thêm thông tin là $\theta = 1 + e, e \sim \text{Exponential}(1)$ và $\theta \leq 2$. Dùng kĩ thuật suy diễn Bayes để ước lượng θ. So sánh sai số của ước lượng này với các ước lượng trên.

Phần II.

d) Tính hệ số tương quan mẫu giữa X và Y.
e) Kiểm định giả thuyết "X và Y có tương quan" bằng kiểm định hệ số tương quan trong `scipy` (https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html).
f) Dùng kĩ thuật lấy mẫu lại hoán vị, kiểm định giả thuyết "X và Y có tương quan" và so sánh kết quả với Câu (e).

Phần III. Xét mô hình hồi qui tuyến tính

$$Y = a + bX + cX^2 + \varepsilon.$$

với a, b, c là các hệ số và ε là lỗi.

g) Ước lượng các hệ số hồi qui a, b, c.
h) Dùng kĩ thuật bootstrapping, xây dựng khoảng tin cậy 95% cho a, b, c.
i) Giả sử ta dùng mô hình hồi qui trên để dự đoán giá trị cho Y là y_0 tại $x_0 = 0.5$. Dùng kĩ thuật bootstrapping, ước lượng sai số dự đoán và xây dựng khoảng tin cậy 95% cho y_0.

Câu 2. (3 điểm) Tìm hiểu bộ dữ liệu Seoul Bike Sharing Demand tại: https://archive.ics.uci.edu/dataset/560/seoul+bike+sharing+demand. Download tập tin dữ liệu tại: https://archive.ics.uci.edu/static/public/560/seoul+bike+sharing+demand.zip, lấy ra tập tin `SeoulBikeData.csv`, nạp dữ liệu và bỏ biến (cột) Date. Với bộ dữ liệu này, ta cần phân tích dự đoán số lượng xe được thuê (Rented Bike Count).

a) Nếu chỉ dùng 2 biến khác để dự đoán số lượng xe được thuê thì nên chọn 2 biến nào?
b) Thực hiện kỹ thuật hồi qui để dự đoán số lượng xe được thuê từ 2 biến đã chọn ở Câu (a) dùng 3 mô hình hồi qui khác nhau từ thư viện scikit-learn.
c) Tương tự Câu (b) nhưng dùng kĩ thuật phân tích Bayes với mô hình phù hợp. So sánh kết quả với các mô hình truyền thống ở Câu (b).

Câu 3. (3.5 điểm) Tìm hiểu bộ dữ liệu Breast Cancer Wisconsin (Original) tại: https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original. Download tập tin dữ liệu tại: https://archive.ics.uci.edu/static/public/15/breast+cancer+wisconsin+original.zip, lấy ra tập tin `breast-cancer-wisconsin.data`, nạp dữ liệu và bỏ biến Sample code number. Trong bộ dữ liệu này, ta cần phân tích dự đoán ung thư (class) theo các biến khác.

a) Thực hiện việc phân lớp dùng 3 mô hình phân lớp khác nhau từ thư viện scikit-learn.
b) Tương tự Câu (a) nhưng dùng kĩ thuật phân tích Bayes với mô hình phù hợp. So sánh kết quả với các mô hình truyền thống ở Câu (a).
c) Dùng kĩ thuật kiểm tra chéo, chọn ra mô hình "tốt nhất" giải thích ung thư (class) theo các đặc trưng (các biến còn lại).
d) Ta thấy rằng tỉ lệ mẫu ung thư (class là 4 - malignant) thấp hơn đáng kể so với mẫu lành tính (2 - benign), vấn đề này được gọi là mất cân bằng lớp (class imbalance). Tìm hiểu vấn đề mất cân bằng lớp và thư viện imbalanced-learn (https://imbalanced-learn.org/stable/index.html) là thư viện hỗ trợ các kĩ thuật giải quyết vấn đề mất cân bằng lớp. Chọn 1 thuật toán tăng mẫu (over-sampling) và 1 thuật toán giảm mẫu (under-sampling) từ thư viện imbalanced-learn để thực hiện lại Câu (a) và so sánh các kết quả.

Lưu ý: Trình bày bài làm (lời giải, công thức Toán, mã Python, kết quả, …) trong tập tin Jupyter Notebook.

--- HẾT ---