BÀI THỰC HÀNH SỐ 2

I. Mục đích:

Để sử dụng và hiện thực các hàm cơ bản ứng dụng trong thống kê và các classifier đơn giản với 1 đặc trưng và 2 lớp. Histogram được sử dụng để xác định biệt số (discriminant) sao cho tối thiểu misclassification.

II. Báo cáo:

Mỗi nhóm sẽ làm báo kết quả riêng, nộp kèm file source chương trình.

III. Nội dung:

1. Xây dựng một phân lớp dựa trên histogram, với tập dữ liệu (tham khảo bài toán từ Lecturer 3):
   - in_time = [(0, 27), (1, 25), (2, 16), (3, 19), (4, 26), (5, 20), (6, 19), (7, 17), (8, 10), (9, 5), (10, 4), (11, 4), (12,2)]
   - cls_late = [(5,3), (6, 5), (7, 8), (8, 15), (9, 17), (10, 18), (11, 19), (12,16), (13, 9), (14, 8), (15, 8)]

   Kiểm tra kết quả nếu rời nhà lúc 6:34, 6:35, 6:36, 6:37, 6:38.
2. Xây dựng bộ phân lớp trên tập dữ liệu 'person_data.txt' cho các trường hợp (1) đặc trưng chiều cao, (2) kết hợp chiều cao và tên.
3. Tính mean và variance của các vector đặc trưng sau:
   a. [1 2 4 6 9 10 20 7]
   b. [0 2 4 6 8 … 100]; tất cả các số chẵn từ 0 đến 100.
   c. [1 3 25 … 9801]; tất cả bình phương các số lẻ từ 1 đến 100.
   d. $\begin{bmatrix}\begin{bmatrix}2\\4\end{bmatrix} & \begin{bmatrix}3\\7\end{bmatrix} & \begin{bmatrix}4\\6\end{bmatrix} & \begin{bmatrix}5\\5\end{bmatrix} & \begin{bmatrix}2\\3\end{bmatrix}\end{bmatrix}$
4. Tính covariance matrix của các vector đặc trưng sau:
   X=[2 3 6 3 7 8] và Y=[5 7 9 6 7 8].
5. Tạo hàm mật độ của phân bố Gauss với mean là 5 và variance là 3. Plot hàm kết quả.
6. Tạo hàm mật độ của phân bố Gauss khác với mean là 2 và variance là 1.5. Plot hàm này trong cùng cửa sổ với hàm được tạo ra ở câu 3. Cho nhận xét.

7. Tạo hàm mật độ phân bố Gauss 2 chiều với mean [1 3] và variance [2 2]. Plot hàm này trên lưới [-10 10] x [-10 10] và tính khoảng cách Mahalanobis đối với các mẫu [0 0], [3 4], và [1 2].

8. Xây dựng bộ classifier sử dụng 1 đặc trưng có sẵn.
   - ***Load data:***
   - Load 2 file tương ứng cho 2 class là: *class1.txt* và *class2.txt*.
   - Cho biết số mẫu và số đặc trưng của mỗi class.
   - Tính mean, variance, và covariance của các vector đặc trưng.
   - Trích chọn 1 đặc trưng.
   - ***Xây dựng classifier ứng với 1 đặc trưng được chọn:***
   - Chia tập dữ liệu thành 2 tập con, tập huấn luyện gồm 50%.
   - Tính histogram và plot:
   - Xác định giá trị biệt số (ngưỡng). Ứng với histogram được tính, chọn biệt số sao cho giá trị misclassification là nhỏ nhất.
   - Đánh giá trên tập dữ liệu test: ứng với giá trị ngưỡng được chọn (biệt số), chúng ta cần phải đánh giá trên 1 tập dữ liệu khác (dữ liệu không được dùng trong huấn luyện); tập dữ liệu test. Dựa vào các giá trị đặc trưng tưng ứng cho các mẫu test và giá trị ngưỡng được xác định, tính phần trăm lỗi.

Tương tự, thực nghiệm với các cách phân chia tập dữ liệu khác nhau (tập dữ liệu huấn luyện là 60%, 70%, và 80%) và các bin của histogram khác nhau.

Cho biết đặc trưng có khả năng phân loại tốt nhất.

9. Thực hiện tương tự như bài 8 với 2 tập dữ liệu là cross (cross.dat) và twoclass (twoclass.dat). Chú ý dữ liệu cho các class đều gom chung 1 file và 2 đặc trưng cuối để chỉ ra class, ví dụ [1 0] tương ứng cho class1 và [0 1] tương ứng cho class2.