Bài tập số 4.

1. Trình bày và cho ví dụ các khái niệm Single Link, Compelete Link, Mean Link, Average Link, Centroid Link, Ward Link trong gom nhóm phân cấp. (2đ)
2. Với thuật toán K-Means, việc lựa chọn trung tâm nhóm ảnh hưởng đến tốc độ hội tụ của thuật toán. K-Means++ là một cải tiến của K-means trong đó có sự cải tiến việc chọn các trung tâm nhóm ban đầu. Hãy tìm hiểu và trình bày sự cải tiến này theo cách hiểu của bạn. Giải thích chi tiết tại sao nó hiệu quả hơn. (3đ)
3. Cài đặt thuật toán Kmeans: (5đ)
    a. Sử dụng bất kỳ ngôn ngữ nào. Không sử dụng thư viện liên quan đến thuật toán cần cài đặt, hay các mã nguồn có sẵn.
    b. Mục tiêu là cài đặt sao cho kết quả gom nhóm là giống với Weka nếu chỉnh các tham số giống nhau hoàn toàn.
    c. Dữ liệu đầu vào theo định dạng ARFF, giả sử tất cả các thuộc tính đều được sử dụng trong gom nhóm trừ thuộc tính cuối cùng là nhãn. Các thuộc tính có thể là rời rạc, số, và có thể bị thiếu.
    d. Các tham số có thể tùy biến như Weka. Riêng đối với 3 tham số bên dưới, ta luôn chọn (tức là chỉ cần cài đặt như sau):
        - Độ đo khoảng cách: Euclidean.
        - Điền giá trị thiếu: Không.
        - Giữ thứ tự mẫu: Có.
    e. Cài đặt đánh giá Classes to Clusters, và cho biết số mẫu gom nhóm sai.
    f. Nếu cài đặt chương trình dạng command line phải có readme.txt cho biết cách truyền tham số. Nếu cài đặt chương trình có giao diện đồ họa thì chỉ cần trình bày tương tự Weka.
    g. Thử nghiệm:
        - Tự chọn 15 tập dữ liệu ARFF có kích thước lớn hơn 100KB từ thư mục http://repository.seasr.org/Datasets/UCI/arff/
        - Các tham số chỉnh giống như mặc định của Weka, riêng số nhóm trùng với số phân lớp của dữ liệu.
        - Trình bày kết quả ở dạng bảng như sau:

Dữ liệu	Số mẫu gom nhóm sai		Thời gian gom nhóm (giây)	
	<mssv1_mssv2>	Weka	<mssv1_mssv2>	Weka
soybean.arff	…	…	…	…
…				

        - Nếu cài đúng yêu cầu thì số mẫu gom nhóm sai sẽ giống nhau cho mỗi tập dữ liệu.

Lưu ý:

Cẩn thận nộp bài và trình bày đúng tuyệt đối yêu cầu dưới đây (tên tập tin, định dạng tập tin, nội dung tập tin), hoặc bị 0 điểm.

a. Bài tập này làm theo nhóm tối đa 2 sinh viên. Thời gian nộp bài xem trên trang web.
b. Tập tin nộp bài: <mssv1_mssv2>.zip nén bên trong là duy nhất một thư mục có tên là: <mssv1_mssv2>. Thư mục này chứa:
    ① <mssv1_mssv2>_1.pdf : trả lời câu hỏi 1. (nếu có làm câu 1)
    ② <mssv1_mssv2>_2.pdf : trả lời câu hỏi 2. (nếu có làm câu 2)

    Nộp đủ như sau (nếu có làm câu 3)

    ③ <mssv1_mssv2>_src.zip: mã nguồn.
    ④ <mssv1_mssv2>_3.pdf : bảng kết quả thử nghiệm cho câu 3.f.
    ⑤ readme.txt (nếu có)

■