Bài 01 \_ Bài tập thêm

Giả sử tồn tại một bảng có số dòng là vô tận do dữ liệu liên tục được thêm vào. Để

đọc hết toàn bộ dữ liệu sẽ mất rất nhiều thời gian nên yêu cầu đặt ra là chỉ được đọc

tất cả một lần duy nhất.

a) Sinh viên hãy thiết kế một mô hình để áp dụng có hiệu quả thuật toán Naïve Bayes

trên dữ liệu này.

* Khi có số lượng dữ liệu lớn thì chúng ta nên sử dụng một số bộ phân loại chạy song song với nhau **(Parallelize Probability Calculations** ) để tiết kiệm thời gian. Chúng ta có thể lưu trữ một khoảng dữ liệu nhất định, thay thế những dữ liệu cũ nhất bằng dữ liệu mà ta mới vừa thu thập được.

b) Người ta muốn theo dõi, so sánh sự thay đổi của mô hình phân lớp theo thời gian

(ví dụ: mô hình phân lớp của tuần trước so với hiện tại...). Sinh viên hãy gợi ý

phương pháp thực hiện điều này.

* Để khám phá sự phát triển của lược đồ phân loại, ta có thể sử dụng cho một số bộ phân loại song song với nhau.
* Ta có thể giữ một bộ phân loại dựa trên:
* Toàn bộ lịch sử của dữ liệu.
* Dữ liệu tuần trước.
* Chỉ dữ liệu ngày hôm trước.
* Ta có thể duy trì số lượng riêng trong bảy ngày trước đó cho bộ phân loại hàng tuần.
* Sau đó, ta có thể loại bỏ số đếm của ngày cũ nhất và thay thế chúng bằng số đếm của ngày trước đó vào cuối mỗi ngày.
* Ta có thể duy trì các số đếm riêng biệt cho mỗi giờ cho bộ phân loại hàng ngày và tại mỗi giờ, có thể thay thế các số đếm cũ nhất bằng các số đếm cho giờ trước đó.