Ví dụ: Lọc thư rác

- ▶ 1 số thực trong đoạn [1,...] thuộc kiểu capital run length average là độ dài trung bình của các chuỗi chứa toàn các kí tự hoa trong thư.
- \blacktriangleright 1 số nguyên trong đoạn [1,...] thuộc kiểu capital run length longest là đô dài của chuỗi dài nhất chứa toàn các kí tư hoa.
- \triangleright 1 số nguyên trong đoạn $[1, \ldots]$ thuộc kiểu capital run length total là tổng đô dài của các chuỗi chứa toàn các kí tư hoa, tức là tổng số kí tư hoa trong thư.

Nếu thư là thư rác thì nó được đánh dấu thuộc lớp 1, không phải thư rác thì thuộc lớp 0

MACHINE LEARNING - HUS 2021

Bài tập



- ullet Sử dụng toàn bộ 57 đặc trưng, có tham số tự do và tốc độ học $lpha=10^{-6}$.
- Sử dụng toàn bộ 57 đặc trưng, không có tham số tự do và tốc độ học $\alpha = 10^{-6}$.
- Sử dụng 55 đặc trưng (bỏ 2 đặc trưng cuối cùng) và có sử dụng tham số tự do θ_0 , tốc độ học $\alpha=10^{-6}$.
- Sử dụng 55 đặc trưng (bỏ 2 đặc trưng cuối cùng) và có sử dụng tham số tự do θ_0 , tốc độ học $\alpha=10^{-3}$.

Ví dụ: Lọc thư rác

- ► Tập dữ liệu Spambase cung cấp 4601 mẫu thư điện tử tiếng Anh dạng thư rác và không phải thư rác cho trong link:
- Tập dữ liệu này thường được dùng để đánh giá hiệu quả của các thuật toán lọc thư rác tự động.
- Một số tính chất của tập dữ liệu này:
 - Tính chất của dữ liệu: đa chiều.
 - Kích thước mẫu: 4601
 - Kiểu của đặc trưng: số nguyên và số thực.
 - Số đặc trưng: 57.

MACHINE LEARNING - HUS 2021

◂▫▸◂▱▸◂▤▸◂▤<mark>▸ ▮</mark>

Ví dụ: Lọc thư rác

Mô tả danh sách các đặc trưng

- ▶ 48 số thực trong đoạn [0, 100] thuộc kiểu word_freq_WORD là phần trăm từ trong thư là WORD, tức là 100*(số lần từ WORD xuất hiện trong thư) / tổng số từ trong thư
 - Mỗi "từ" là bất kì một chuỗi kí tự nào, có thể là từ theo nghĩa thông thường hoặc một kí hiệu (token).
 - Một số từ thuộc 48 từ được xét: make, address, all, 3d, our, money, technology, conference..
- ▶ 6 số thực trong đoạn [0,100] thuộc kiểu char_freq_WORD là phần trăm kí tự trong thư là CHAR, tức là 100*(số lần kí tự CHAR xuất hiện trong thư) / tổng số kí tự trong thư. Sáu kí tự được xét là ;, (, [, !, \$ và #.