Báo Cáo Bài Tập Nhóm:

DEFAULT OF CREDIT CARD CLIENTS DATA SET

DSSV:

1. Lê Hữu Nghĩa
2. Lê Minh Trí
3. Võ Hoàng Trung
4. Phạm Hoàng Viện

CANTHO UNIVERSITY

Nội dung:

I. Tổng quan tập dữ liệu.
II. Sơ lược về giải thuật MLPClassifier.
III. Huấn luyện MLPClassifier.
IV. Phân chia - Huần luyện tập dữ liệu.
V. So sánh MLP – DecisionTree.
VI. Đánh giá.

I. Tổng quan tập dữ liệu

- Nguồn:
  - https://archive.ics.uci.edu
    - Department of Information Management, Chung Hua University, Taiwan.
    - Department of Civil Engineering, Tamkang University, Taiwan.
- Dự đoán về xác suất vỡ nợ của khách hàng.
- Nội dung: Kết quả của độ chính xác dự đoán về xác suất khách hàng đáng tin cậy hoặc không đáng tin cậy.
- Khó khăn: Nhiều cột có giá trị liên tục khó khăn trong việc tính toán.

I. Tổng quan tập dữ liệu

- Số dòng dữ liệu: 30000 dòng.
- Số cột: 24 cột.
- Chi tiết thuộc tính và nhãn:
  - LIMIT_BAL : Số tiền tín dụng đã vay.
  - Gender: Giới tính.
  - Education: Trình độ giáo dục.
  - Marital status: Tình trạng hôn nhân.
  - Age: Tuổi.
  - PAY_1 - 6: Lịch sử thanh toán trong quá khứ (09/2005).
  - BILL_AMT1 - 6: Số tiền sao kê hóa đơn (09/2005).
  - PAY_AMT1 - 6: Số tiền thanh toán trước đó (09/2005).
  - Default payment next month: KH có khả năng vỡ nợ hay không.

I. Tổng quan tập dữ liệu

- Thuộc tính(Cột): 1 - 23.
- Nhãn(Cột): 24 “default payment next month”.
  - Giá trị nhãn: 1 – có, 0 – không.
- Do nhãn của tập dữ liệu có giá trị *Nhị phân*
- => Multi-layer Perceptron Classifier làm giải thuật huấn luyện mô hình đạt xác suất cao nhất.

II. Sơ lược về giải thuật MLPClassifier

- Hàm mạng: $u_b = \sum_a w_{a-b}.o_a$

- Hàm kích hoạt (*sigmoid*): $f(u) = \frac{1}{1+e^{-u/T}} \Rightarrow f'(u) = \frac{f(u)[1-f(u)]}{T}$

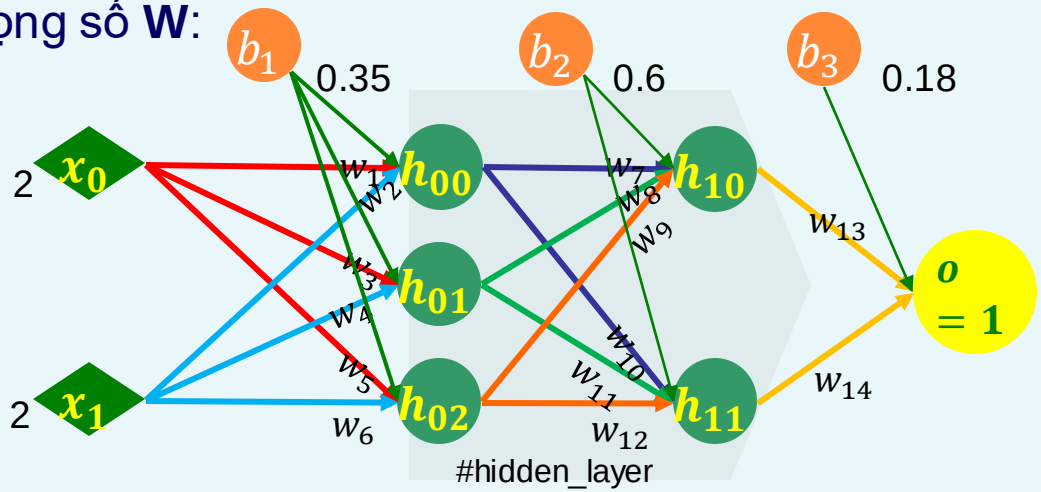
- Hàm lỗi: $E = \frac{1}{2}(y_b - o_b)^2$

- Gradient hàm lỗi đối với *tầng đầu ra*: $\frac{\partial E}{\partial o_b} = -(y_b - o_b)$

- Gradient hàm lỗi đối với *tầng bất kỳ* khác tầng đầu ra: $\frac{\partial E}{\partial w_{a-b}} = \frac{\partial E}{\partial o_b}.f'(u_b).o_b$

III. Huấn luyện MLPClassifier

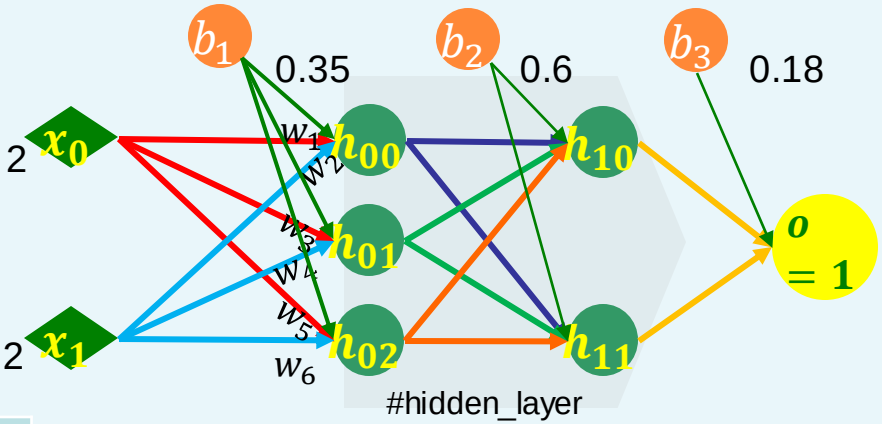
- Huấn luyện:
  - Chọn 1 bất kì với 2 trường làm input layer.
  - Tốc độ học: 0.2
  - Số lần lập: 1
  - Mô hình khởi tạo hidden_layer = 2 [3, 2].
  - Bias = [0.35, 0.6, 018]
  - Khởi tạo ngẫu nhiên trọng số **W**:



III. Huấn luyện MLPClassifier

PAY1	PAY2	Y
2	2	1

Bảng 1: Dữ liệu input_layer



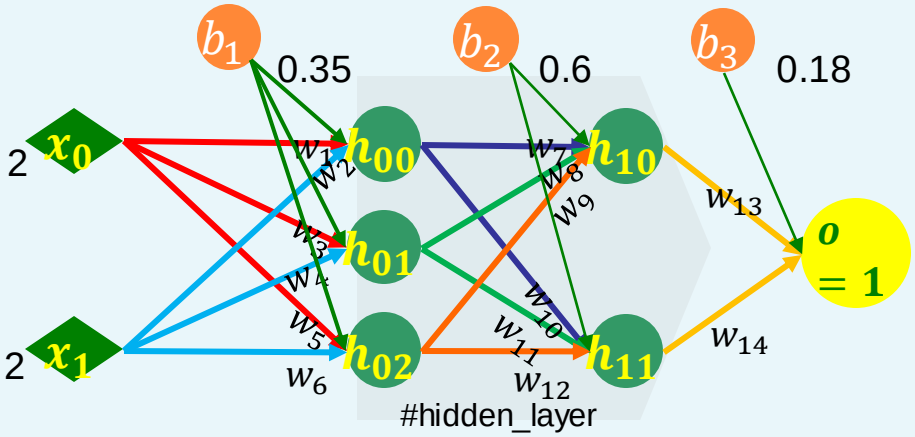
W$[x_i; h_1]$	h_{00}	h_{01}	h_{02}
x_0	$w_1 = 0.03$	$w_3 = 0.15$	$w_5 = 0.04$
x_1	$w_2 = 0.01$	$w_4 = 0.02$	$w_6 = 0.18$

Bảng 2: Vector trọng số giữa input_layer – hl_1

III. Huấn luyện MLPClassifier

W[h_1; h_2]	h_{10}	h_{11}
h_{00}	$w_7 = 0.13$	$w_{10} = 0.11$
h_{01}	$w_8 = 0.21$	$w_{11} = 0.22$
h_{02}	$w_9 = 0.3$	$w_{12} = 0.5$

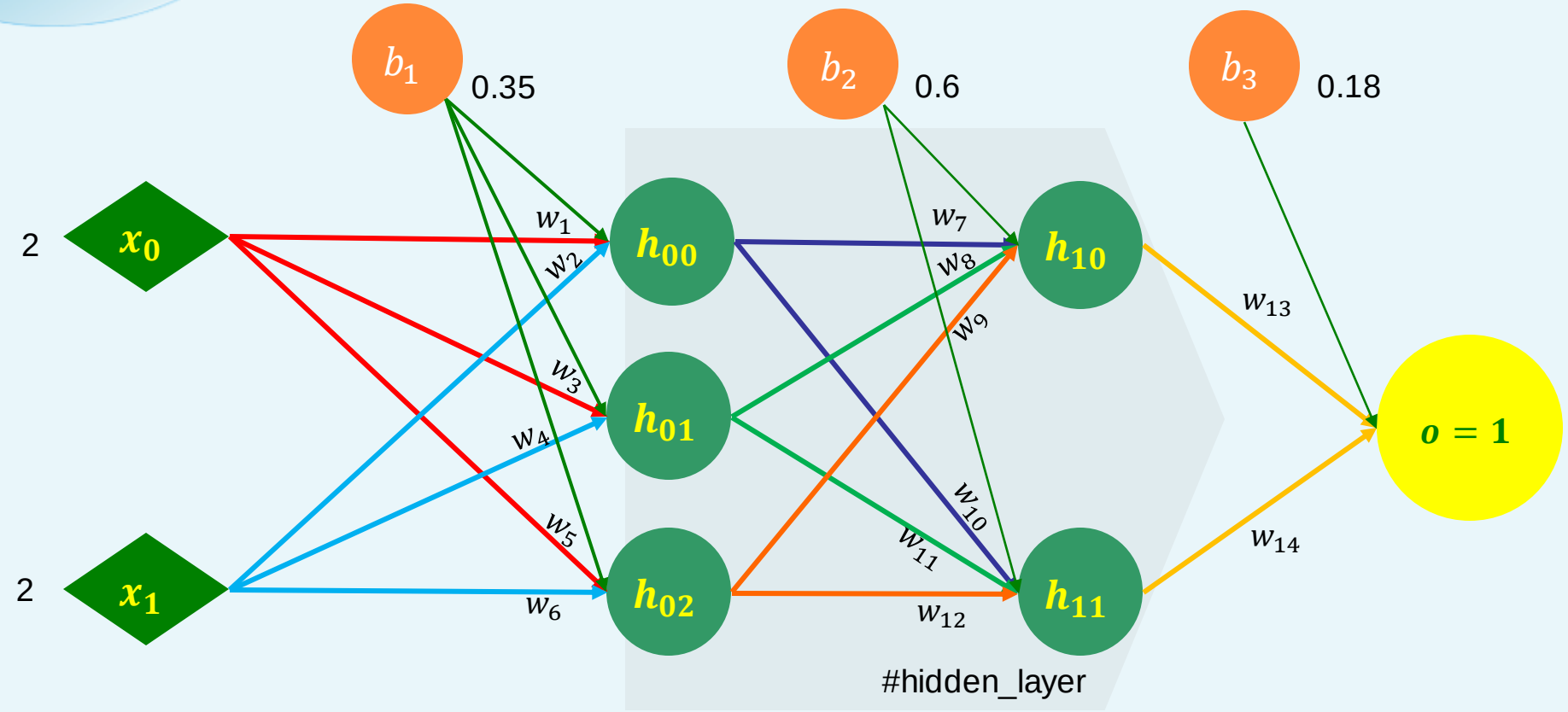
Bảng 3: Vector trọng số giữa hl_1 - hl_2



W[h_2;out]	Out
h_{10}	$w_{13} = 0.06$
h_{11}	$w_{14} = 0.1$

Bảng 4: Vector trọng số giữa hl2 – out_layer

III. Huấn luyện MLPClassifier



Mạng nơ-ron đa tầng với 1 – input_layer, 2 – hidden_layer(3,2), 1 – output_layer

III. Huấn luyện MLPClassifier

- Quá trình lan truyền tiến (Signal forward):
  - $net_{h00} = 0.03 * 2 + 0.01 * 2 + 0.35 * 1 = 0.43$
  - $out_{h00} = \frac{1}{1+ e^{-0.43}} = 0.61$
  - $net_{h01} = 0.15 * 2 + 0.02 * 2 + 0.35 * 1 = 0.69$
  - $out_{h01} = \frac{1}{1+ e^{-0.69}} = 0.66$
  - $net_{h02} = 0.04 * 2 + 0.18 * 2 + 0.35 * 1 = 0.79$
  - $out_{h02} = \frac{1}{1+ e^{-0.79}} = 0.68$
  - $net_{h10} = 0.13 * 0.43 + 0.21 * 0.69 + 0.3 * 0.79 + 0.6 * 1 = 1.04$
  - $out_{h10} = \frac{1}{1+ e^{-1.04}} = 0.73$
  - $net_{h11} = 0.11 * 0.43 + 0.22 * 0.69 + 0.50 * 0.79 + 0.6 * 1 = 1.19$
  - $out_{h11} = \frac{1}{1+ e^{-1.19}} = 0.76$

III. Huấn luyện MLPClassifier

- Quá trình lan truyền tiến (Signal forward):
  - $net_o = 1.04 * 0.06 + 1.19 * 0.1 + 0.18 * 1 = 0.36$
  - $out_o = \frac{1}{1+ e^{-0.36}} = 0.59$

III. Huấn luyện MLPClassifier

- Quá trình lan truyền ngược (Backpropation):
  - Tính toán kích thước lỗi của Output:
    - $E_{total} = \frac{1}{2}(1 - 0.59)^2 = 0.08405$
  - Ngưỡng chịu lỗi của từng trọng số và cập nhật lại trọng số:
    - $\frac{\partial E_{total}}{\partial w_{14}} = \frac{\partial E_{total}}{\partial out_o} * \frac{\partial out_o}{\partial net_o} * \frac{\partial net_o}{\partial w_{14}} = -0.07$
      - w_{14} =0.1 – 0.2*(-0,07) = 0,08
    - $\frac{\partial E_{total}}{\partial w_{13}} = \frac{\partial E_{total}}{\partial out_o} * \frac{\partial out_o}{\partial net_o} * \frac{\partial net_o}{\partial w_{13}} = -0.07$
      - w_{13}=0,06 - 0,2(-0,07) = 0,046
    - $\frac{\partial E_{total}}{\partial w_{12}} = \frac{\partial E_{total}}{\partial out_{h11}} * \frac{\partial out_{h11}}{\partial net_{h11}} * \frac{\partial net_{h11}}{\partial w_{12}} = -9.835^{-4}$
      - $w_{12} = 0{,}5 - 0{,}2 * (-9.835^{-4}) = 0{,}499$
    - $\frac{\partial E_{total}}{\partial w_{11}} = \frac{\partial E_{total}}{\partial out_{h11}} * \frac{\partial out_{h11}}{\partial net_{h11}} * \frac{\partial net_{h11}}{\partial w_{11}} = -9.546^{-4}$
      - w_{11} = 0,22 – 0,2* (-9.546^{-4}) = 0,219
    - $\frac{\partial E_{total}}{\partial w_{10}} = \frac{\partial E_{total}}{\partial out_{h11}} * \frac{\partial out_{h11}}{\partial net_{h11}} * \frac{\partial net_{h11}}{\partial w_{10}} = -8.823^{-4}$
      - w_{10} = 0,11 – 0,2*(-8.823^{-4}) = 0,109

III. Huấn luyện MLPClassifier

- Quá trình lan truyền ngược (Backpropation):
  - Ngưỡng chịu lỗi của từng trọng số:
    - $\frac{\partial E_{total}}{\partial w_9} = \frac{\partial E_{total}}{\partial out_{h10}} * \frac{\partial out_{h10}}{\partial net_{h10}} * \frac{\partial net_{h10}}{\partial w_9} = -6.75^{-4}$
      - $w_9 = 0.3 - 0.2 * (-6.75^{-4}) = 0,299$
    - $\frac{\partial E_{total}}{\partial w_8} = \frac{\partial E_{total}}{\partial out_{h10}} * \frac{\partial out_{h10}}{\partial net_{h10}} * \frac{\partial net_{h10}}{\partial w_8} = -6.5517^{-4}$
      - $w_8 = 0.21 - 0.21*(-6.5517^{-4}) = 0,209$
    - $\frac{\partial E_{total}}{\partial w_7} = \frac{\partial E_{total}}{\partial out_{h10}} * \frac{\partial out_{h10}}{\partial net_{h10}} * \frac{\partial net_{h10}}{\partial w_7} = -6.055^{-4}$
      - $w_7 = 0.13 - 0.2* (-6.055^{-4}) = 0.129$
    - $\frac{\partial E_{total}}{\partial w_6} = \frac{\partial E_{total}}{\partial out_{h02}} * \frac{\partial out_{h02}}{\partial net_{h02}} * \frac{\partial net_{h02}}{\partial w_6} = 8.4408^{-4} = \frac{\partial E_{total}}{\partial w_5}$
      - $w_6 = 0.18 - 0.2*(8.4408^{-4}) = 0.179$
      - $w_5 = 0.04 - 0.2* (8.4408^{-4}) = 0.0398$

III. Huấn luyện MLPClassifier

- Quá trình lan truyền ngược (Backpropation):
  - Ngưỡng chịu lỗi của từng trọng số:
    - $\frac{\partial E_{total}}{\partial w_4} = \frac{\partial E_{total}}{\partial out_{h01}} * \frac{\partial out_{h01}}{\partial net_{h01}} * \frac{\partial net_{h01}}{\partial w_4} = 1.580^{-4} = \frac{\partial E_{total}}{\partial w_3}$
      - w_4 = 0.02 – 0.2*(1.580^{-4}) = 0.01969
      - w_3 = 0.15 – 0.2*(1.580^{-4}) = 0.149
    - $\frac{\partial E_{total}}{\partial w_2} = \frac{\partial E_{total}}{\partial out_{h00}} * \frac{\partial out_{h00}}{\partial net_{h00}} * \frac{\partial net_{h00}}{\partial w_2} = -1.359^{-4} = \frac{\partial E_{total}}{\partial w_1}$
      - w_2 = 0.01 – 0.2* (-1.359^{-4}) = 0.01
      - $w_1 = 0.03 - 0.2 * (-1.359^{-4})$ = 0.03

IV. Phân chia - Huần luyện tập dữ liệu

- Phân chia tập dữ liệu theo nghi thức Hold-Out:
  - Tập train: 80% - 24000 (rows)
  - Tập test: 20% - 6000 (rows)
  - Chọn random_state:
    - Với giá trị = 0:
      - Accuracy is: 78.38333333333334
    - Với giá trị = 5:
      - Accuracy is: 77.68333333333334
    - Với giá trị = 10:
      - Accuracy is: 78.05

IV. Phân chia - Huần luyện tập dữ liệu

- Huần luyện với các tham số khác nhau(MLPClassifier):
  - activation = “tanh”, solver=“adam”, random_state=“0”, $\eta = 0.01$:
    - Accuracy: 78.38333333333334
  - activation = “tanh”, solver=“sgd”, random_state=“0”, $\eta = 0.04$:
    - Accuracy: 78.38333333333334
  - activation = “logistic”, solver=“adam”, random_state=“5”, $\eta = 0.1$:
    - Accuracy: 78.38333333333334
  - activation = “logistic”, solver=“sgd”, random_state=“100”, $\eta = 0.2$:
    - Accuracy: 78.38333333333334

V. So sánh MLP - DecisionTree

- So sánh độ chính xác tổng thể sau 10 lần lặp giữa MLPClassifier - DecisionTreeClassifier:

Số lần lặp	MLPClassifier (%)	DecisionTreeClassifier(%)
1	78.38	82.95
2	78.05	82.45
3	77.71	82.36
4	78.63	81.86
5	78.05	82.75
6	78.88	82.58
7	77.06	81.75
8	78.53	81.89
9	78.60	82.71
10	78.01	82.08
Trung bình	78.38	82.95

VI. Đánh giá

- Nhận xét:
  - MLPClassifier: Theo lý thuyết thì khi nhãn có giá trị là nhị phân thì giải đây là giải thuật thích hợp. Nhưng do giá trị các cột data của tập dữ liệu có nhiều giá trị lớn và liên tục nên xác xuất ra vẫn chưa được tối ưu.
  - DecisionTreeClassifier: So với MLP thì giải thuật lại tối ưu hơn. Nếu có sử dụng trong ứng dụng thực tế thì nên chọn giải thuật này.

Cảm ơn Cô và mọi người đã lắng nghe!