Đại học Khoa học Tự nhiên
Đại học Quốc gia Thành phố Hồ Chí Minh
Khoa CNTT – Môn KHDL

BÁO CÁO ĐỒ ÁN CUỐI KỲ

GVHD: Trần Trung Kiên

Người thực hiện:
Võ Phương Hòa - 1412192

I. Phát biểu bài toán

- **Vấn đề**:

  + Đọc/mua sách là nhu cầu cơ bản của con người.

  + Thị trường sách hiện tại vô cùng phong phú (thể loại, nội dung, tác giả…)

  + Người dùng muốn đọc/mua tựa sách có chất lượng tốt (nội dung, chủ đề, chất lượng giấy…)

I. Phát biểu bài toán

- Bài toán phát sinh: làm thế nào người đọc xác định được 1 cuốn sách nên đọc hay không! (tránh lãng phí thời gian và tiền bạc)

$\Rightarrow$ **Nội dung đồ án:** đánh giá tựa sách

(Phạm vi đồ án: giới hạn trong độc giả Việt Nam)

I. Phát biểu bài toán

- **Mô tả bài toán:**

Input: một tập các mẫu gồm 9 thuộc tính:

+ id	+ title	+ topic
+ nxb	+ author	+ date
+ reviews	+ rating_Count	+ price

Output: rating

=> từ kết quả thu được, đưa ra đánh giá phù hợp!

II. Giải quyết bài toán

1. Thu thập dữ liệu

- Nguồn: https://tiki.vn/
- Số mẫu dữ liệu: 29705 mẫu
- Số thuộc tính: 10 thuộc tính
- Tên các thuộc tính:

+ **id:** mã sách + **title:** tên sách + **topic:** chủ đề sách
+ **nxb:** tên nxb + **author:** tác giả + **date:** ngày xuất bản
+ **reviews:** số lượt nhận xét, phê bình + **rating_Count**
+ **price:** giá bán chính thức + **rating**

II. Giải quyết bài toán

1. Thu thập dữ liệu

- Phân chia mẫu dữ liệu:
    + Training set: 20000 mẫu
    + Validation set: 4705 mẫu
    + Test set: 5000 mẫu

II. Giải quyết bài toán

1. Tiền xử lý dữ liệu

1.1. Tiền xử lý trong giai đoạn thu thập dữ liệu

- Vấn đề:
    + Dữ liệu thiếu:
        author
        rating
        rating_Count
    + Dữ liệu có dấu

II. Giải quyết bài toán

1. Tiền xử lý dữ liệu

1.1. Tiền xử lý trong giai đoạn thu thập dữ liệu

- Cách giải quyết:

+ Chuẩn hóa dữ liệu thiếu:

author = ‘unidentify’

rating = 0

rating_Count = 0

+ Dữ liệu có dấu (unicode):

=> gọi lệnh **encode(‘utf-8’)**

II. Giải quyết bài toán

1. Tiền xử lý dữ liệu

1.1. Tiền xử lý dữ liệu tập huấn luyện

- Vấn đề:

    + Thừa dữ liệu (id và title)

    + Dữ liệu xấu (review)

    + Dữ liệu dạng chuỗi (topic, nxb, author, date)

II. Giải quyết bài toán

1. Tiền xử lý dữ liệu

1.1. Tiền xử lý dữ liệu tập huấn luyện

- Cách giải quyết:

    + Loại bỏ 3 thuộc tính id, title và review

    + Dữ liệu dạng chuỗi: thống kê số mẫu khác nhau và đánh số từ 0 -> n – 1 (n là số mẫu khác nhau).

- Chuẩn hóa dữ liệu các cột về dạng có mean = 0, std = 1.

II. Giải quyết bài toán

1. Tiền xử lý dữ liệu

1.1. Tiền xử lý dữ liệu tập huấn luyện

- Thuộc tính phân lớp **rating**: phân loại như sau

rating ϵ [0, 1) => 0

rating ϵ [1, 2) => 1

rating ϵ [2, 3) => 2

rating ϵ [3, 4) => 3

rating $\geq$ 4 => 4

II. Giải quyết bài toán

1. Tiền xử lý dữ liệu

1.1. Tiền xử lý dữ liệu tập validation và tập test

Tương tự việc tiền xử lý trên tập huấn luyện

II. Giải quyết bài toán

2. Kế hoạch sử dụng dữ liệu

- Phương pháp: ANN (Artificial Neural Network)
- Thuật toán: SGD (Stochastic Gradient Descent)
  Backpropagation Algorithm
- Activation function: Sigmoid + Softmax
- Cost function: Mean Negative Log Likelihood
- Error function: Mean Binary Error
- Regularization: Early Stopping/Weight Decay

II. Giải quyết bài toán

2. Kế hoạch sử dụng dữ liệu

Mô tả mạng neurals (ANN):

- Input layer: 6 neurals tương ứng với các thuộc tính: topic, nxb, author, date, price, rating_Count.
- Hidden layer: [50] (1 lớp ẩn có 50 neurals)
- Output layer: 5 neurals (tương ứng với thuộc tính rating).

II. Giải quyết bài toán

2. Kế hoạch sử dụng dữ liệu

Huấn luyện dữ liệu:

B1: Khởi tạo bộ trọng số W.

B2: Xáo trộn thứ tự các mẫu huấn luyện.

B3: Phân tập dữ liệu N mẫu thành m tập con (mỗi tập có size là M, với m*M = N).

B4: Với mỗi tập con, xét từng mẫu (x, y), trong đó:

$x = [x_0, x_1, \ldots, x_5]$ $(x_i \in [0, 1])$

$y \in \{0, 1, 2, 3, 4\}$

(chuyển y về dạng one_hot_Y)

II. Giải quyết bài toán

2. Kế hoạch sử dụng dữ liệu

Huấn luyện dữ liệu:

B4: Với mỗi mẫu (x, y):

- Forward-propagation:

  Dùng hàm sigmoid để tính output của các tầng ẩn

  $a_i = 1/(1+e^{-z})$ với $z = w_i^T x$

  Dùng hàm sofmax để tính output của tầng cuối cùng (tầng output).

$$a_i = \frac{\exp(z_i)}{\sum_{j=1}^{C} \exp(z_j)}, \quad \forall i = 1, 2, \ldots, C$$

II. Giải quyết bài toán

2. Kế hoạch sử dụng dữ liệu

Huấn luyện dữ liệu:

B4: Với mỗi mẫu (x, y):

- Back-propagation:

  Tính độ lỗi error từ tầng output -> tầng ẩn đầu tiên.

$$\delta_i^{(l)} = \frac{\partial e}{\partial s_i^{(l)}} = \theta'\left(s_i^{(l)}\right) \sum_{k=1}^{d^{(l+1)}} W_{ik}^{(l+1)} \delta_k^{(l+1)}$$

- Tính gradient

$$\nabla_W e(h(x), y): \frac{\partial e}{\partial W_{ji}^{(l)}} = a_j^{(l-1)} \delta_i^{(l)}$$

II. Giải quyết bài toán

2. Kế hoạch sử dụng dữ liệu

Huấn luyện dữ liệu:

B5: Cập nhật trọng số

$$W \leftarrow W - \alpha * \frac{1}{B} \sum_{(x,y) \in mb} \nabla_W e(h(x), y)$$

B6: Thực hiện lại B3 cho đến khi huấn luyện xong tất cả các tập con.

II. Giải quyết bài toán

2. Kế hoạch sử dụng dữ liệu

Sử dụng regularization trong huấn luyện:

- Dùng weight decay: tại mỗi bước cập nhật trọng số W, bổ sung: $\mathbf{W \leftarrow W - 2\eta\lambda W}$.

- Dùng early stopping: theo dõi và cập nhật độ lỗi tốt nhất (nhỏ nhất) MBE trên tập validation.

III. Kết quả thực nghiệm

1. Không sử dụng regularization

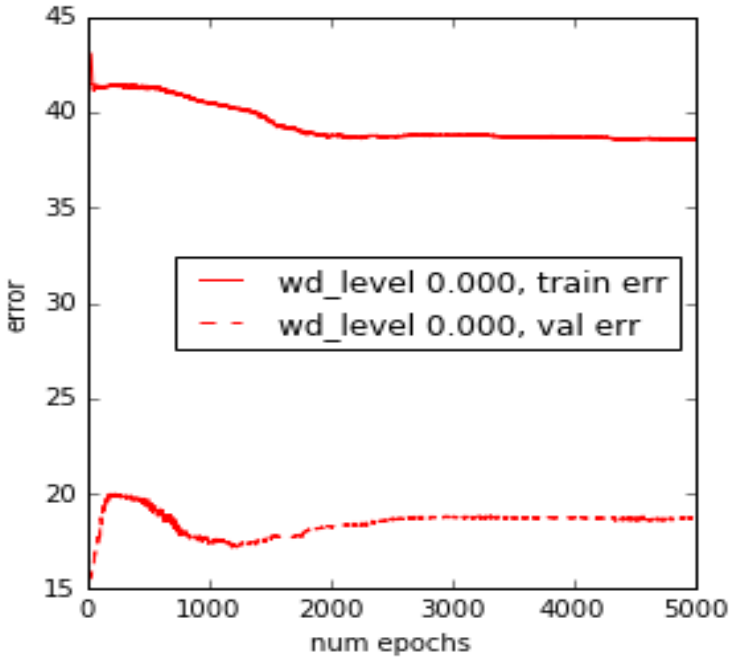
hidden_layer_sizes = [50] mb_size = 40

learning_rate = 0.0005 epoch = 5000

⇒ Kết quả huấn luyện:

train_err = 38.575%

val_err = 18.661%



III. Kết quả thực nghiệm

2. Sử dụng regularization

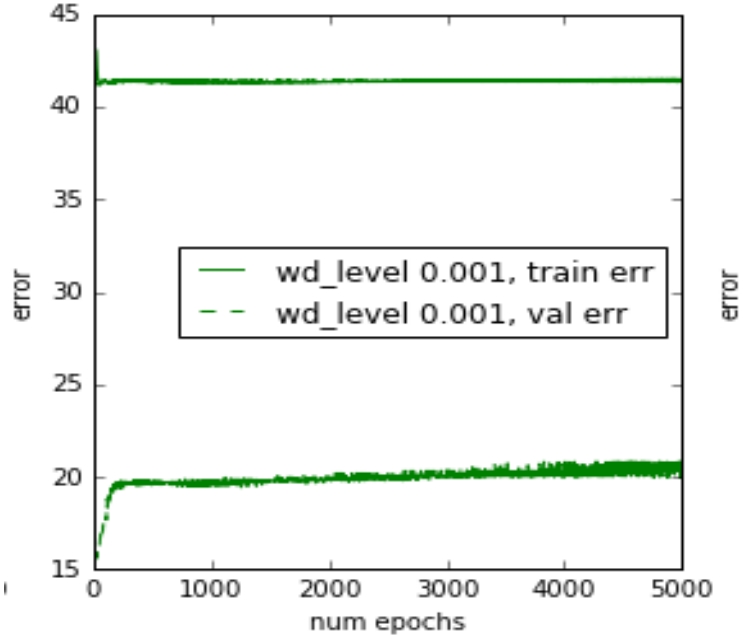
hidden_layer_sizes = [50] mb_size = 40

learning_rate = 0.0005 epoch = 5000

- Weight decay:
  - Với wd_level = 0.001:

    train_err = 41.45%

    val_err = 20.191%



III. Kết quả thực nghiệm

2. Sử dụng regularization

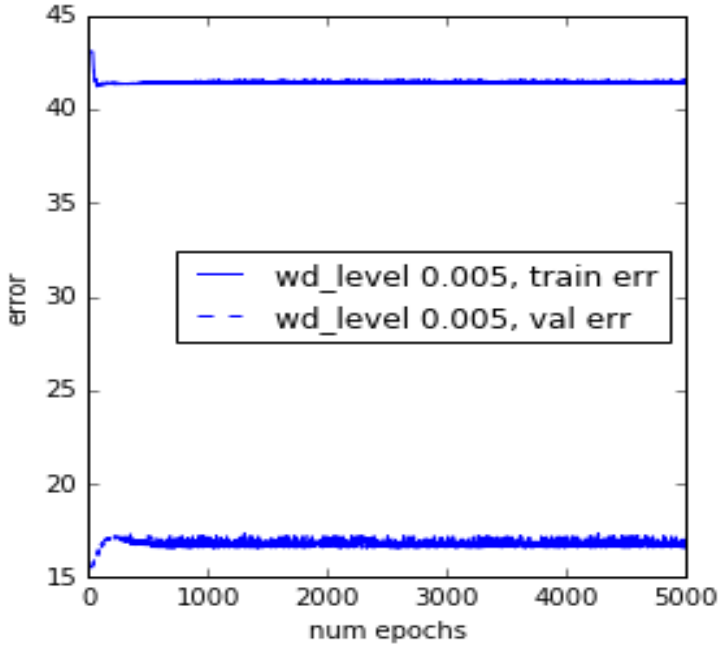
hidden_layer_sizes = [50] mb_size = 40

learning_rate = 0.0005 epoch = 5000

- Weight decay:
  - Với wd_level = 0.005

    train_err = 41.43%

    val_err = 16.642%



III. Kết quả thực nghiệm

2. Sử dụng regularization

hidden_layer_sizes = [50] mb_size = 40

learning_rate = 0.1 max_epoch = 1000000

- Early stopping:
  - max_patience = 3000:

    train_err = 18.465%

    val_err = 7.97%

    test_err = 11.32%

III. Kết quả thực nghiệm

2. Đánh giá kết quả

- So sánh các kết quả đạt được, việc sử dụng regularization cho kết quả khả quan hơn là non-regularization.
- Trong 2 regularization thử nghiệm, mô hình huấn luyện dùng early stopping cho kết quả tốt hơn weight decay.

Cảm ơn thầy đã theo dõi!