TRƯỜNG ĐẠI HỌC THỦY LỢI

**KHOA CÔNG NGHỆ THÔNG TIN**

A blue and white logo

Description automatically generated with low confidence

**BÀI TẬP LỚN**

HỌC PHẦN: HỌC SÂU VÀ ỨNG DỤNG

**ĐỀ TÀI: DỰ ĐOÁN CHẤT LƯỢNG SỮA**

Giáo viên hướng dẫn: Nguyễn Thị Kim Ngân

Sinh viên/nhóm sinh viên thực hiện:

1. Trần Thanh Thảo, lớp 62TH-NB

2.Sái Thị Nga, lớp 62TH-NB

3. Nguyễn Hà Linh, lớp 63CNTT2

**Hà Nội, năm 2023**

**MỤC LỤC**

[**Phần 1: Tổng quan** 2](#_Toc135352641)

[***1.*** ***Mô tả bài toán*** 2](#_Toc135352642)

[***2. Phương pháp học sâu được sử dụng trong bài tập lớn*** 3](#_Toc135352643)

[**2.1 Phương pháp Softmax Regression** 3](#_Toc135352644)

[**2.2 Phương pháp Artificial Neural Network** 6](#_Toc135352645)

[**Phần 2: Thực nghiệm** 10](#_Toc135352646)

[***1.Mô tả tập dữ liệu*** 10](#_Toc135352647)

[***2.Đánh giá mô hình*** 11](#_Toc135352648)

[***Kết luận*** 12](#_Toc135352649)

[**Tài liệu tham khảo** 12](#_Toc135352650)

# **Phần 1: Tổng quan**

## ***Mô tả bài toán***

* Tên bài toán: Dự đoán chất lượng sữa
* Mô tả: Dựa vào các chỉ số của sữa để đánh giá chất lượng sữa
* Input: tập dữ liệu gồm các thuộc tính và đã được gán nhãn.
* Output: kết quả dự đoán chất lượng của sữa mở mức độ nào
* Tóm tắt công việc thực hiện của bài toán:
* Đọc dữ liệu từ file
* Chuẩn hóa dữ liệu
* Chia tập dữ liệu thành 2 phần: 70% dùng để huấn luyện mô hình, 30% dùng để kiểm tra sự phù hợp của mô hình
* Sử dụng các phương pháp để giải quyết bài toán:
* Phương pháp Softmax Regression
* Phương pháp Artificial Neural Network
* Đánh giá mô hình dựa trên các độ đo: accuracy, precission, recall, F1-score

## ***2. Phương pháp học sâu được sử dụng trong bài tập lớn***

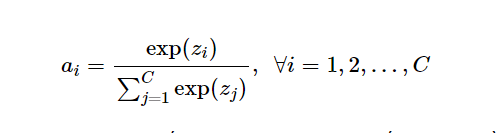
### **2.1 Phương pháp Softmax Regression**

#### 2.1.1 Giới thiệu

* Softmax regression, còn được gọi là regression softmax hoặc hồi quy softmax, là một thuật toán học máy được sử dụng trong bài toán phân loại đa lớp. Nó là một phương pháp mở rộng của hồi quy logistic (logistic regression) để xử lý bài toán phân loại có nhiều hơn hai lớp.
* Softmax regression là một trong những phương pháp phân loại phổ biến và được sử dụng rộng rãi trong nhiều lĩnh vực, bao gồm nhận dạng chữ viết tay, nhận dạng ảnh, xử lý ngôn ngữ tự nhiên và nhiều ứng dụng khác.

#### 2.1.2 Công thức của Softmax

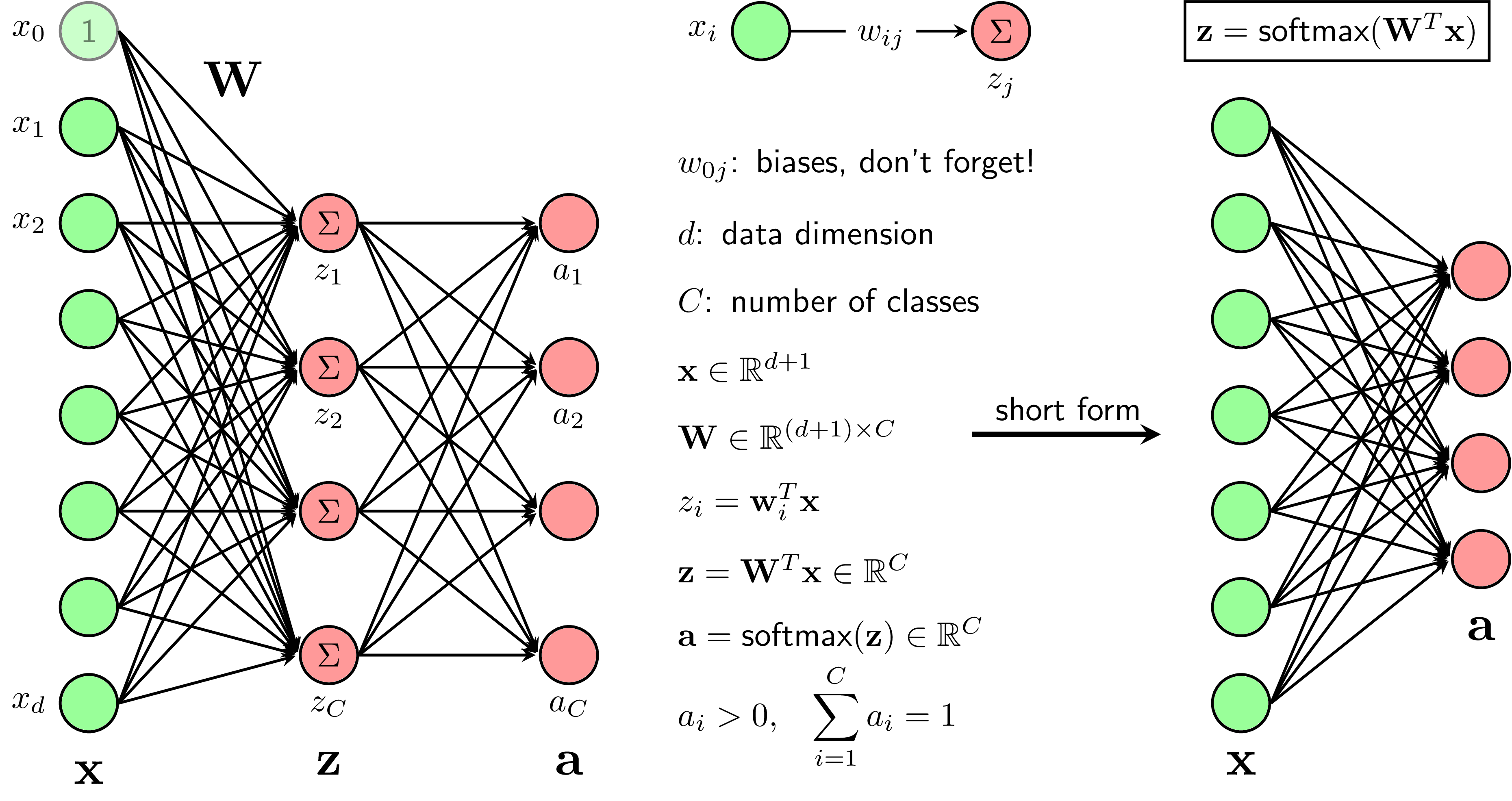
* Chúng ta cần một mô hình xác suất sao cho với mỗi input x,  thể hiện xác suất để input đó rơi vào class i. Vậy điều kiện cần là các  phải dương và tổng của chúng bằng 1. Để có thể thỏa mãn điều kiện này, chúng ta cần nhìn vào mọi giá trị  và dựa trên quan hệ giữa các  này để tính toán giá trị của . Ngoài các điều kiện  lớn hơn 0 và có tổng bằng 1, chúng ta sẽ thêm một điều kiện cũng rất tự nhiên nữa, đó là: giá trị  càng lớn thì xác suất dữ liệu rơi vào class i càng cao. Điều kiện cuối này chỉ ra rằng chúng ta cần một hàm đồng biến ở đây.
* Chú ý rằng có thể nhận giá trị cả âm và dương. Một hàm số mượt đơn giản có thể chắc chắn biến  thành một giá trị dương, và hơn nữa, đồng biến, là hàm . Điều kiện mượt để thuận lợi hơn trong việc tính đạo hàm sau này. Điều kiện cuối cùng, tổng các  bằng 1 có thể được đảm bảo nếu:

**

* Hàm số này, tính tất cả các  dựa vào tất cả các , thõa mãn tất cả các điều kiện đã xét: dương, tổng bằng 1, giữ được thứ tự của . Hàm số này được gọi là softmax function. Chú ý rằng với cách định nghĩa này, không có xác suất  nào tuyệt đối bằng 0 hoặc tuyệt đối bằng 1, mặc dù chúng có thể rất gần 0 hoặc 1 khi  rất nhỏ hoặc rất lớn khi so sánh với các ,
* Lúc này, ta có thể giả sử rằng:

**

* Trong đó, được hiểu là xác suất để một điểm dữ liệu **x** rơi vào class thứ i nếu biết tham số mô hình (ma trận trọng số) là **W**
* Hình vẽ dưới đây thể hiện mạng Softmax Regression dưới dạng neural network:



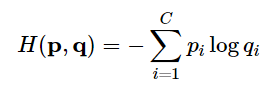
Ở phần bên phải, hàm tuyến tính Σ và hàm softmax (activation function) được tách riêng ra để phục vụ cho mục đích minh họa. Dạng short form ở bên phải là dạng hay được sử dụng trong các Neural Networks, lớp **a** được ngầm hiểu là bao gồm cả lớp **z**.

#### 2.1.3 Cross Entropy

Cross entropy giữa hai phân phối**p** và **q** được định nghĩa là:

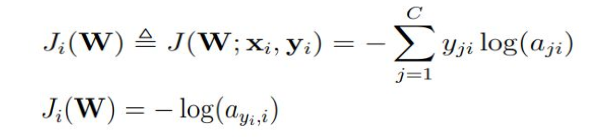


Với **p** và **q** là rời rạc (như **y** và**a** trong bài toán của chúng ta), công thức này được viết dưới dạng:



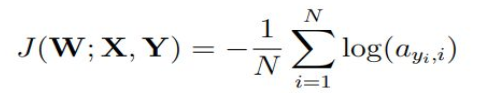
#### 2.1.4 Hàm mất mát cho Softmax Regression

- Với Softmax Regression, trong trường hợp có C lớp dữ liệu, mất mát giữa đầu ra dự đoán và đầu ra thực sự của một điểm dữ liệu xi với label (one-hot)  được tính bằng:



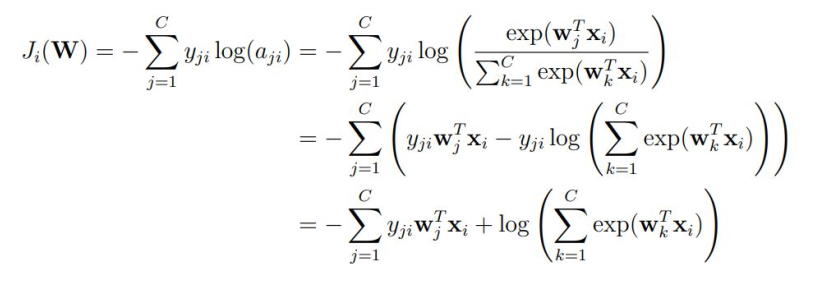
Với yji và aji lần lượt là là phần tử thứ j của vector (xác suất) yi và ai.

- Kết hợp tất cả các cặp dữ liệu xi,yi,i=1,2,…,N, chúng ta sẽ có hàm mất mát cho Softmax Regression như sau:

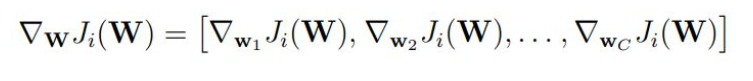


#### 2.1.5 Tối ưu hàm mất mát

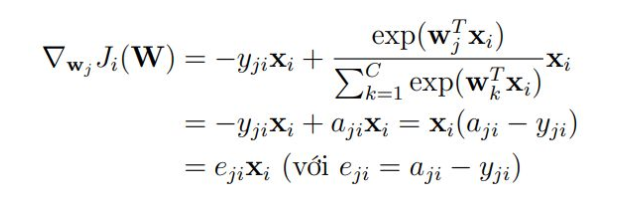
- Với chỉ một cặp dữ liệu (xi,yi) ta có:



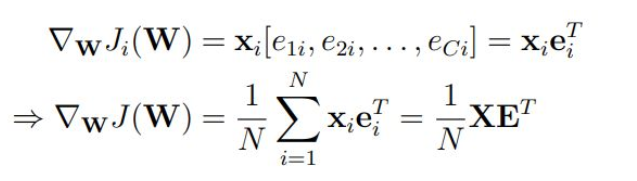
- Tiếp theo ta sử dụng công thức tính đạo hàm:



* Trong đó, gradient theo từng cột wj có thể tính được dựa theo công thức trên:



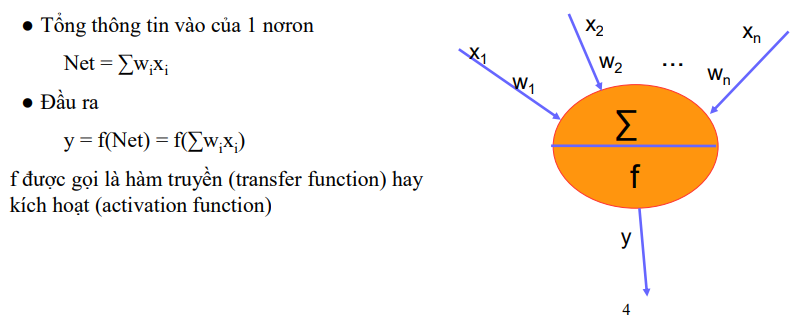
* Giá trị eji=aji−yji, E= A-Y có thể được coi là sai khác giữa đầu ra dự đoán và đầu ra thực sự. Ta có:



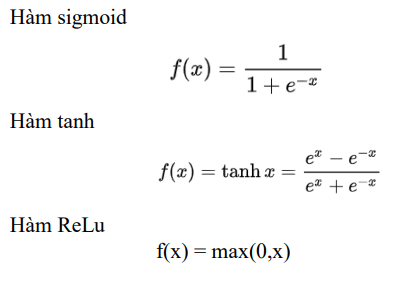
### **2.2 Phương pháp Artificial Neural Network**

#### 2.2.1 Giới thiệu

* Neural Network là một mô hình toán học phức tạp được đưa ra để xử lý thông tin, giải quyết các vấn đề phổ biến trong lĩnh vực học máy, trí tuệ nhân tạo AI, deep learning và tìm kiếm tất cả các mối quan hệ cơ bản trong một tập hợp các dữ liệu. Tất cả điều này được Neural Network thực hiện thông qua việc bắt chước các hoạt động của của tế bào thần kinh và khớp thần kinh từ bộ não của con người.
* Hiểu đơn giản, Neural Network tương tự như bộ não của con người, nó sẽ kết nối các nút đơn giản và các nút này sẽ tạo thành một mạng lưới. Do đó, nó được xem là hệ thống của các tế bào thần kinh nhân tạo. Trong Neural Network, một loạt các thuật toán sẽ được sử dụng với mục đích xác định và nhận ra tất cả các mối quan hệ có trong các tập dữ liệu.



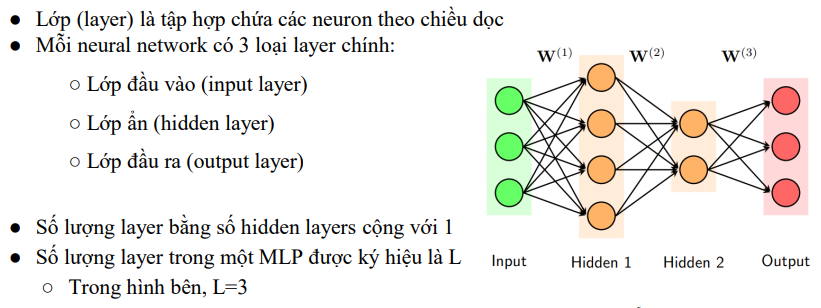
* ***Một số hàm kích hoạt phổ biến***



#### 2.2.2 Cấu trúc mạng Nơron

Neural Network là sự kết hợp của các tầng perceptron (perceptron đa tầng). Và mỗi mạng Neural Network thường bao gồm ba thành phần chính như sau:

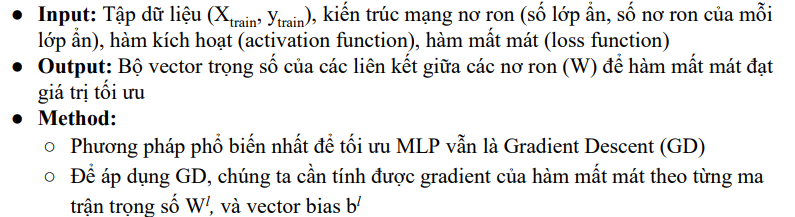
* **Tầng vào (input layer):**Tầng vào nằm ở phía bên trái của Neural Network. Tầng này đại diện cho các đầu vào của mạng nơron nhân tạo.
* **Tầng ra (output layer):**Tầng ra nằm ở phía bên phải của Neural Network. Nó đại diện cho những đầu ra của mạng nơron nhân tạo.
* **Tầng ẩn (hidden layer):** Đây là tầng nằm xen giữa hai tầng là tầng vào và tầng ra thể hiện cho quá trình suy luận logic của Neural Network. Nó đại diện cho nút trung gian phân chia không gian đầu vào các vùng mềm (vùng ranh giới). Tầng ẩn sẽ nhận một tập hợp các đầu vào có trọng số và thông qua một chức năng kích hoạt để tạo ra kết quả đầu ra.



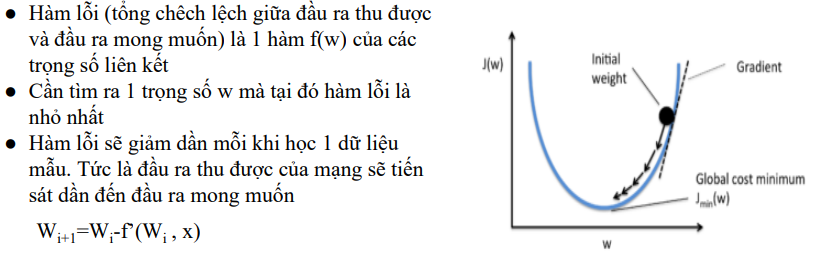
Với mạng Neural Network, mỗi nút mạng sẽ được coi là sigmoid nơron nhưng hàm kích hoạt của chúng lại hoàn toàn khác nhau. Thực tế, người ta sẽ thường sử dụng hàm kích hoạt cùng loại để việc tính toán trở nên thuận lợi hơn. Số lượng nút mạng tại mỗi tầng có thể khác nhau tùy thuộc vào cách giải quyết của bạn. Tuy nhiên:

* Các tầng ẩn sẽ được để với số lượng nơron khác nhau.
* Những nơron nằm ở tầng thường sẽ liên kết với nhau để tạo thành một mạng kết nối đầy đủ. Điều này sẽ giúp người dùng có thể dựa vào tầng và số lượng nơron để tính toán kích cỡ mạng.

#### 2.2.3 Thuật toán



#### 2.2.4 Phương pháp giảm Gradient



#### **2.2.5 Ứng dụng mạng Nơron**

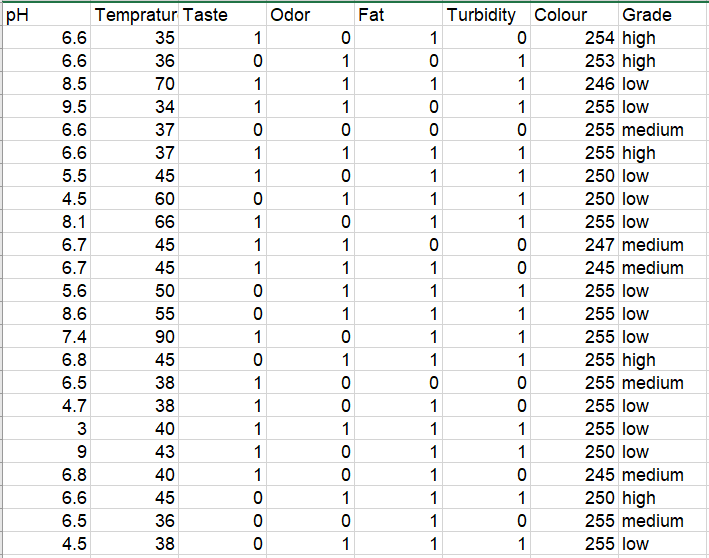
Với số lượng các ứng dụng được triển khai ngày càng lớn, Neural Network, máy học và trí tuệ nhân tạo được ứng dụng trong nhiều công việc khác nhau như:

* Nhận dạng chữ viết tay: Mạng nơron nhân tạo Neural Network được sử dụng với mục đích chuyển đổi các ký tự viết tay thành các ký tự kỹ thuật số để máy tính dễ dàng nhận ra được các ký tự đó.
* Dự đoán các giao dịch chứng khoán: Việc theo dõi sàn giao dịch chứng khoán thường khó khăn và khó hiểu do có nhiều yếu tố ảnh hưởng đến thị trường này. Neural Network ra đời có thể kiểm tra được các yếu tố nói trên và dự đoán giá chứng khoán hàng ngày. Điều này mang đến nhiều lợi ích cho các nhà môi giới chứng khoán.
* Vấn đề đi lại của các nhân viên bán hàng: Neural Network giúp giải quyết và tìm ra con đường tối ưu cho các nhân viên bán hàng trong việc đi lại giữa các thành phố thuộc cùng một khu vực cụ thể đồng thời mang đến doanh thu cao và giảm thiểu chi phí.
* Nén hình ảnh: Ý tưởng đằng sau của Neural Network nén dữ liệu là lưu trữ, mã hoá và tái tạo các hình ảnh mang tính thực tế. Bằng cách sử dụng Neural Network nén hình ảnh, bạn có thể dễ dàng tối ưu hoá kích thước dữ liệu của mình. Điều này sẽ giúp bộ nhớ của bạn tiết kiệm được dung lượng đáng để.
* Ngoài ra, Neural Network cũng được sử dụng nhiều trong nhiều ứng dụng và công nghệ khác nhau như thị giác máy tính, trò chơi điện tử, dịch tự động, lọc mạng xã hội, nhận dạng giọng nói và chẩn đoán y tế. Điều đáng ngạc nhiên hơn là Neural Network cũng được sử dụng nhiều trong các hoạt động truyền thống và sáng tạo như nghệ thuật hay hội hoạ.

# **Phần 2: Thực nghiệm**

# ***1.Mô tả tập dữ liệu***

* Dữ liệu có 1060 vector gồm các thuộc tính và nhãn



* Mô tả ma trận dữ liệu (X):
* Gồm 1060 dòng và 7 cột
* Mỗi cột tương ứng với một thuộc tính lần lượt là: pH, Temprature, Taste, Odor, Fat, Turbidity, Colour

|  |  |
| --- | --- |
| **Tên thuộc tính** | **Ý nghĩa** |
| pH | Chỉ số pH |
| Temprature | Nhiệt độ |
| Taste | Mùi vị |
| Odor | Mùi hương của sữa |
| Fat | Chất béo |
| Turbidity | Độ đục |
| Colour | Màu sắc |

* Mô tả nhãn lớp (Y): gồm 117 dòng và 1 cột tương ứng là Grade

|  |  |
| --- | --- |
| **Tên nhãn** | **Ý nghĩa** |
| Grade | Chất lượng sữa:   * Low: Xấu * Medium: Trung bình * High: Tốt |

## ***Đánh giá mô hình***

Dùng tập test data để đánh giá chất lượng của mô hình được lựa chọn theo các độ đo:

* Phương pháp Artificial Neural Network

+ Độ đo của 5 lần lặp:

Text

Description automatically generated

+ Độ đo trung bình:

Text

Description automatically generated

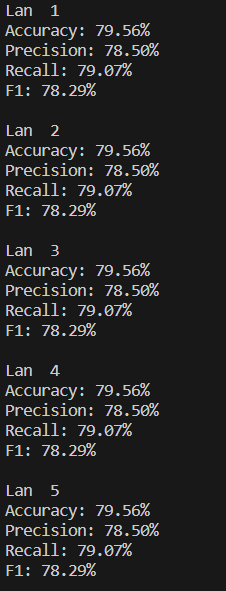
+ Độ đo tốt nhất:

Text

Description automatically generated

* Phương pháp Softmax Regression

+ Độ đo của 5 lần lặp:



+ Độ đo trung bình:

Text

Description automatically generated

+Độ đo tốt nhất:

Text

Description automatically generated

## 

## 

## ***Kết luận***

Qua bài làm về đề tài: Ứng dụng thuật toán Softmax Regression và Artificial Neural Network để dự đoán chất lượng sữa dựa vào các chỉ số, nhóm chúng em hi vọng mô hình này có thể phát triển hơn nữa. Đồng thời thông qua bài làm giúp nhóm em có những kiến thức về môn học máy cũng như áp dụng kiến thức đó vào phân tích các bài toán thực tế.

* Tìm ra tỷ lệ dự đoán đúng.
* Cho ra kết quả dự đoán dữ liệu về chất lượng sữa thông qua các chỉ số

Mặc dù đã có nhiều cố gắng trong suốt quá trình làm bài tập, nhưng kiến thức còn hạn chế, cũng như chưa có kinh nghiệm trong thực tế nên không thể tránh khỏi những sai sót. Vì vậy, nhóm em mong thầy, cô góp ý để có thể chỉnh sửa, hoàn thiện bài làm tốt hơn.

## **Tài liệu tham khảo**

1. Machine learning cơ bản: <https://machinelearningcoban.com/>
2. Kaggle: https://www.kaggle.com/datasets/cpluzshrijayan/milkquality
3. Slide bài giảng