1 Tìm hiểu về machine learning

Machine Learning (ML) là một ngành con của khoa học máy tính nhằm giúp máy tính có khả năng hiểu được dữ liệu đầu vào mà không cần qua việc lập trình một cách tường tận.

Phân loại	Khái quát	Bài toán	Thuật toán cụ thể
Supervised learning	Có sẵn một số cặp đầu vào – ra, mong muốn dự đoán đầu ra với đầu vào mới.	Classification, Regression	Linear regression, K-Nearest Neighbor
Unsupervised learning	Sử dụng dữ liệu chưa gán nhãn và đầu ra không cụ thể, thường sử dụng để khai thác thêm thông tin từ dữ liệu.	Clustering, Dimensional reduction(TBA)	K-means clustering
Semi-supervised learning	TBA	TBA	TBA
Reinforcement learning	TBA	TBA	TBA

1.1 Linear Regression

Least square method, xấp xỉ hàm cần tính bằng cách tối thiểu bình phương khoảng cách giữa các điểm vào – ra (x_i, y_i) đã biết.

Đầu vào:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{\Phi} = \begin{pmatrix} \varphi_1 = \varphi_1(x_1) \\ \varphi_2 = \varphi_2(x_2) \\ \vdots \\ \varphi_n = \varphi_n(x_n) \end{pmatrix}$$

Đầu ra: $\mathbf{w} = (w_i)^T$

Hàm cần tối thiểu hóa:

$$L(\mathbf{w}) = \sum_{i=1}^{n} (w_i \varphi_i - y_i)^2$$

Tối thiểu hóa bằng đạo hàm riêng phần $\partial(L)/\partial w_i$ thu được nghiệm:

$$\mathbf{w} = \left(\Phi^T \Phi\right)^{-1} \Phi^T \mathbf{y}$$

1.2 K-nearest neighbors

Thuật toán học lười, tìm K những điểm dữ liệu có khoảng cách gần nhất và đoán đầu ra dựa trên những điểm đó.

1.3 K-means clustering

Phân các điểm dữ liệu vào các cụm cho $\|x_i - m_j\|$ min, với x_i là điểm dữ liệu, m_i là tâm mỗi cụm.

Cần tìm $\mathbf{Y} = (y_{ij})$ với $y_{ij} = \begin{cases} 1 & \text{nếu } x_i \in \text{cụm } j \\ 0 & \text{nếu } x_i \notin \text{cụm } j \end{cases}$ sao cho $\sum_{j=1}^{\text{số cụm}} y_{ij} \|x_i - m_k\|^2$ đạt min

Nếu biết trước tâm mỗi cụm " Gán mỗi điểm dữ liệu vào cụm có khoảng cách từ điểm đó đến tâm ngắn nhất

Tìm tâm mỗi cụm Giải đạo hàm riêng phần theo m_j của hàm mục tiêu được:

$$m_j = \frac{\sum_{i=1}^{n} x_i y_{ij}}{\sum_{i=1}^{n} y_{ij}} = \frac{\text{Tổng giá trị các điểm trong cụm } j}{\text{Tổng số điểm trong cụm } j}$$

Vậy m_j là điểm trung bình cộng của các điểm dữ liệu trong cùng một cụm.

1. Chọn ngẫu nhiên một tâm cho mỗi cụm
2. Gán các điểm vào cụm, dựa trên khoảng cách ngắn nhất tới tâm
3. Cập nhật tâm mới bằng cách lấy trung bình các điểm trong một cụm
4. Nếu tâm mới == tâm cũ dừng, không chuyển qua bước 2

2 Tìm hiểu về NLP

NLP là một lĩnh vực giao thoa giữa khoa học máy tính, trí tuệ nhân tạo và ngôn ngữ học.

2.1 Một số ứng dụng

- Nhận diện giọng nói
- Giám sát mạng xã hội
- Bot
- Dịch thuật máy
- Kiểm tra chính tả
- Công cụ tìm kiếm
- Trích rút thông tin
- Quảng cáo

2.2 Các quá trình cơ bản

Quá trình	Khái quát	Note
Tokenization	Chia đoạn văn bản thành những thành phần nhỏ hơn có nghĩa	Có thể có nhiều cách chia, đơn giản nhất là chia mỗi câu hoặc từ
Lọc stopwords	Loại bỏ những từ không mang lại nhiều ý nghĩa trong câu	Tùy từng ngôn ngữ và use case
Stemming	Đưa các từ về một loại duy nhất	Vì tiếng Việt không danh từ hóa như tiếng Anh nên bước này có thể sẽ vô nghĩa với tiếng Việt?
Lemmatization	Đưa các từ về cùng một thì	Tiếng Việt không chia thì như tiếng Anh
Tagging	Nhận diện từ loại của mỗi từ.	-
Entity Recognition	Nhận diện thực thể xuất hiện trong văn bản	-
Chunking	Xếp các token thành một cụm có nghĩa	Token thu được ở bước tokenization và qua xử lý ở các quá trình trên

2.3 Word embedding