Lập trình nâng cao

Lê Thành Văn

24-09-2020

Khoa Hệ thống thông tin quản lý

Giới thiệu

Học phần

Học phần này bao gồm 2 phần chính :

- Làm việc với `DataFrame` (thư viện `pandas` của Python).
- Giới thiệu về machine learning và các thuật toán cơ bản.

Mục đích

- Nắm bắt được các giai đoạn và kỹ năng cơ bản khi phân tích dữ liệu
- Nắm bắt các thuật toán machine learning cơ bản

Thư viện pandas

Giới thiệu về pandas

- Là một thư viện trên Python được phát triển bởi Wes McKinney trong năm 2008.
- Là thư viện chuẩn cho việc phân tích dữ liệu khi dùng Python.

Tính năng của pandas

- Có thể xử lý tập dữ liệu khác nhau về định dạng.
- Có khả năng đọc dữ liệu từ nhiều nguồn khác nhau (csv, db/sql, excel, ...).
- Có thể xử lý nhiều phép toán cho tập dữ liệu
- Xử lý mất mát dữ liệu

Các kiểu dữ liệu trong pandas

- `Series` : kiểu dữ liệu dạng mảng 1 chiều
- `DataFrame` : kiểu dữ liệu mảng 2 chiều
- `Panel` : kiểu dữ liệu mảng 3 chiều

Cài đặt pandas

- Thường sẽ được đi kèm khi cài đặt anaconda.
- Nếu không sử dụng anaconda thì chạy `pip install pandas` trong command prompt (cmd)

Sử dụng pandas trong python

Để sử dụng `pandas` thêm dòng `import pandas as pd` ở đầu file

Machine Learning

Giới thiệu

Nói một cách đơn giản, machine learning (máy học) là những phương pháp cung cấp cho máy tính cách học hỏi mà không cần hướng dẫn chi tiết từng bước.
Machine learning là một ĩnh vực của Artificial Intelligence (AI, trí tuệ nhân tạo)

Phương pháp

Machine learning bao gồm các phương pháp sau (theo cách học) :

- Supervised Learning : học có giám sát (phân loại (classification), hồi quy (regression), ...)
- Unsupervised Learning : học không giám sát (phân nhóm (clustering), ...)
- Reinforcement Learning : học củng cố (học sâu (deep learing), mạng nơ-ron (neural networks), ...)

Hiện tại, có rất nhiều vấn đề được giải quyết bằng machine learning

- Xử lý ảnh (gắn thẻ hình ảnh, nhận diện ký tự, nhận diện khuôn mặt, ...)
- Xử lý văn bản (phát hiện spam, phân tích ngữ nghĩa, ...)
- Khai phá dữ liệu (gom nhóm, dự đoán, ...)

Quy trình cơ bản

1. Thu thập dữ liệu
2. Làm sạch dữ liệu
3. Khám phá dữ liệu
4. Chuyển đổi dữ liệu
5. Xây dựng mô hình
6. Đánh giá và điều chỉnh

Machine learning trong Python

Trong python có nhiều package về machine learning như `scikit-learn`, `TensorFlow`, ...

Tài liệu tham khảo

pandas

- Hướng dẫn về pandas
- pandas' documentation : tài liệu của pandas (phương thức, kiểu dữ liệu, ...).
- Stack Overflow về pandas : là nơi sẽ trả lời các khó khăn khi làm việc với pandas

Machine Learning

- scikit-learn's documentation : tài liệu của scikit-learn (thư viện machine learning của microsoft)
- Hướng dẫn về TensorFlow : thư viện machine learning của Google

Khóa học

- Programming with Python for Data Science : khóa học về pandas và một số thuật toán cơ bản trong machine learning.
- Khóa học machine learning của Andrew Ng

Khác

Sách :

- Understanding Machine Learning: From Theory to Algorithms

Internet

- Machine Learning cơ bản