# Lập trình nâng cao

Lê Thành Văn

06-10-2020

Khoa Hệ thống thông tin quản lý

# Giới thiệu

# Mục đích

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

# Nội dung

Môn học 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.

# Thời khóa biểu

Chương trình học gồm 15 buổi (3 tiết một buổi):

- Thứ 3 (tiết 7): 01/12/2020 – 19/01/2021.
- Thứ 6 (tiết 7): 04/12/2020 – 22/01/2021.

# Điểm số

Điểm quá trình (50%):

- Điểm chuyên cần (40%).
- Điểm bài tập (60%).

Điểm cuối kỳ (50%) (thi thực hành).

# Tài liệu tham khảo

# pandas

- Hướng dẫn pandas.
- pandas' documentation: tài liệu của pandas (phương thức, kiểu dữ liệu, ...).
- Stack Overflow về pandas: hỏi đáp về 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.
- Machine Learning for Hackers: Case Studies and Algorithms to Get You Started

Internet:

- Machine Learning cơ bản

# Thư viện pandas

# Giới thiệu

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

# Tính năng

- 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

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

# Cài đặt

Cài đặt mới: `pip install pandas`.

Cập nhật: `pip install pandas --upgrade`.

# Sử dụng

Người ta thường import pandas như sau:

```
import pandas as pd
```

# Machine Learning

# Giới thiệu

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 lĩnh vực của Artificial Intelligence (AI, trí tuệ nhân tạo)

# Phương pháp

Machine learning bao gồm các phương pháp sau:

- 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), ...)

# Vấn đề giải quyết được

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

Thông thường, người ta sử dụng quy trình sau để giải quyết vấn đề bằng machine learing:

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

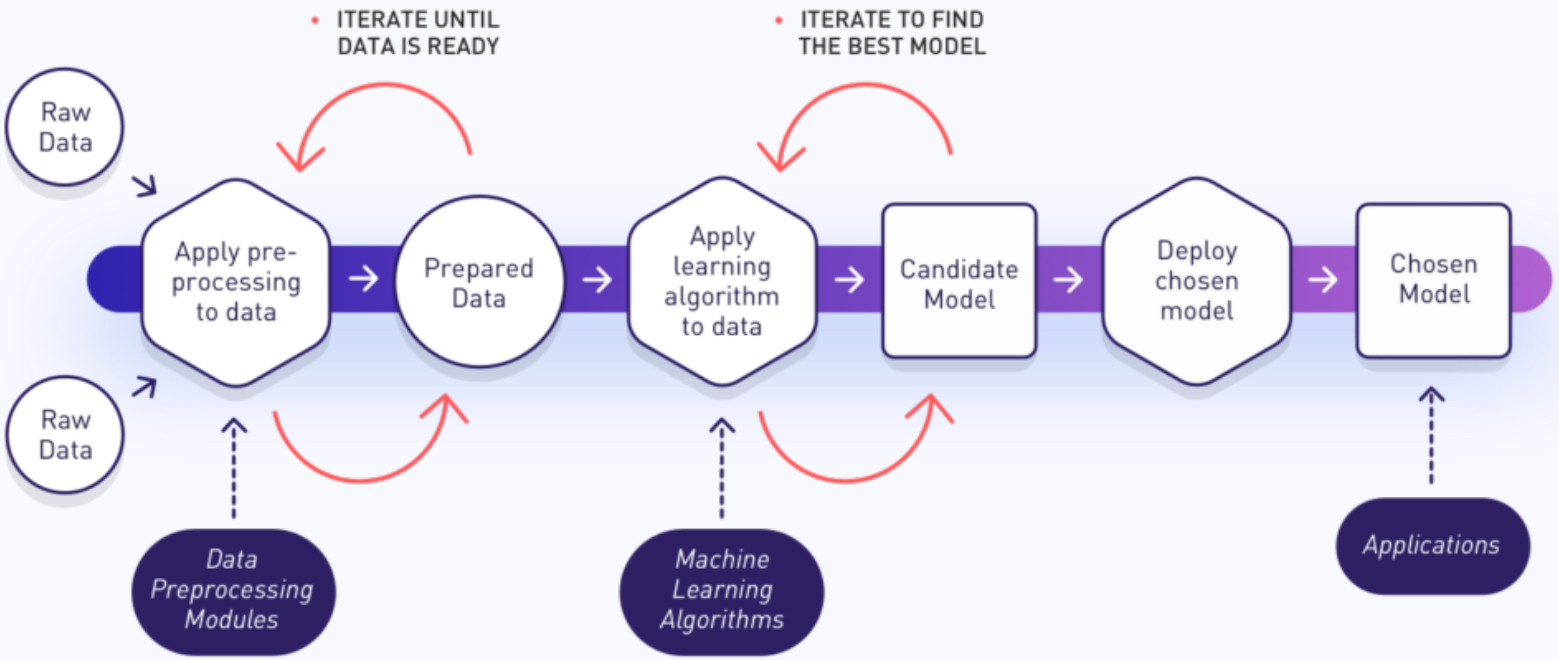


Figure 1: Quy trình cơ bản

# Machine learning trong Python

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