Softech®
Gather to thrive together

Aptech™
COMPUTER EDUCATION
Unleash your potential

BÀI 9

Tổng quan về Data Science

Tóm Tắt Nội Dung

Trong bài học này chúng ta sẽ đi tìm hiểu lần lượt các nội dung

1. Giới thiệu về Data Science
2. Các thư viện phổ biến trong Data Science
3. Cài đặt Môi trường
4. Làm quen với NumPy

9.1 Giới thiệu về Data Science

Data Science là gì?

Data Science (Khoa học dữ liệu) là lĩnh vực kết hợp giữa lập trình, thống kê, và trực quan hóa để trích xuất thông tin và hiểu biết từ dữ liệu.

Các khía cạnh chính

Thu thập và xử lý dữ liệu

1

Thu thập dữ liệu, làm sạch và chuẩn bị dữ liệu để phân tích

Phân tích và mô hình hóa

2

Sử dụng các thuật toán và kỹ thuật, bao gồm học máy (Machine Learning) và trí tuệ nhân tạo (AI), để phân tích dữ liệu và xây dựng các mô hình dự đoán

Trực quan hóa và diễn giải

3

Biến kết quả phân tích thành thông tin dễ hiểu thông qua trực quan hóa dữ liệu (ví dụ: biểu đồ, đồ thị) và trình bày cho các bên liên quan

9.1 Giới thiệu về Data Science

Một số ứng dụng phổ biến của Data Science

Data Science góp một phần không hề nhỏ vào ứng dụng thực tế

Lĩnh vực	Ứng dụng cụ thể	Mô tả
Thương mại điện tử	Gợi ý sản phẩm	Phân tích lịch sử mua hàng, hành vi người dùng để gợi ý sản phẩm phù hợp.
Tài chính – Ngân hàng	Phát hiện gian lận	Dùng mô hình học máy để phát hiện giao dịch bất thường.
Giao thông – Logistics	Dự đoán thời gian giao hàng	Dựa trên dữ liệu tuyến đường, thời tiết, lịch sử giao hàng.
Y tế	Dự đoán bệnh, chẩn đoán hình ảnh	Dùng dữ liệu bệnh nhân, ảnh X-quang, MRI để hỗ trợ bác sĩ chẩn đoán.
Giải trí	Gợi ý phim, nhạc	Netflix, Spotify, YouTube dùng Data Science để cá nhân hóa nội dung.
Khí tượng – Môi trường	Dự báo thời tiết, ô nhiễm	Phân tích dữ liệu cảm biến và vệ tinh.

9.1 Giới thiệu về Data Science

Quy trình cơ bản trong Data Science

1. Xác định vấn đề
2. Thu thập dữ liệu
3. Khám phá dữ liệu
4. Xử lý & làm sạch dữ liệu
5. Phân tích & Mô hình hóa dữ liệu
6. Trình bày kết quả **(Báo cáo)**

9.1 Giới thiệu về Data Science

1 **Business Question (Xác định vấn đề)**

- **Mục tiêu:** Hiểu rõ *vấn đề thực tế* mà doanh nghiệp hoặc tổ chức cần giải quyết.

- **Ví dụ:**

  - Làm sao để dự đoán khách hàng nào có khả năng rời bỏ dịch vụ?
  - Làm sao để tối ưu giá bán sản phẩm?
  - Doanh số bán hàng tháng nào cao nhất năm và tại sao ?

Thông thường đây là một tập hợp các câu hỏi hoặc vấn đề mà cấp trên cần làm rõ trong bản báo cáo

9.1 Giới thiệu về Data Science

2 Get Data (Thu thập dữ liệu)

- **Mục tiêu:** Thu thập dữ liệu liên quan đến vấn đề.

- **Nguồn dữ liệu có thể đến từ:**
  - Cơ sở dữ liệu nội bộ (SQL, CRM, Excel, v.v.)
  - API (Twitter, Google, OpenWeatherMap, v.v.)
  - Web scraping (trích xuất dữ liệu từ web)
  - Sensor / IoT hoặc dữ liệu mở (open data)

- **Kết quả:** Một hoặc nhiều bộ dữ liệu thô (raw data)

9.1 Giới thiệu về Data Science

3 Explore Data (Khám phá dữ liệu)

- **Mục tiêu:** Hiểu cấu trúc, đặc điểm và chất lượng của dữ liệu

- **Công việc chính:**
  - Kiểm tra kiểu dữ liệu, số lượng bản ghi, giá trị bị thiếu.
  - Thống kê mô tả (mean, median, std, count).
  - Trực quan hoá dữ liệu bằng biểu đồ (histogram, boxplot, scatter plot,…)

- **Kết quả:** Cái nhìn sơ bộ về dữ liệu — biết được điều gì đáng tin cậy, điều gì cần làm sạch

9.1 Giới thiệu về Data Science

4 Prepare Data (Xử lý & làm sạch dữ liệu)

- **Mục tiêu:** Chuẩn bị dữ liệu ở dạng sạch và sẵn sàng phân tích
- **Công việc chính:**
  - Loại bỏ hoặc thay thế giá trị bị thiếu (NaN).
  - Chuẩn hoá định dạng (ngày giờ, số, chữ).
  - Mã hoá dữ liệu dạng chữ (label encoding, one-hot encoding).
  - Tạo các đặc trưng mới (feature engineering).
- **Kết quả:** Dữ liệu **"sạch"** (clean dataset) có thể đưa vào mô hình hoặc phân tích.

9.1 Giới thiệu về Data Science

5 Analyze Data (Phân tích & Mô hình hóa dữ liệu)

- **Mục tiêu:** Tìm hiểu mối quan hệ, xu hướng, và đưa ra dự đoán
- **Cách thực hiện**
  - Phân tích thống kê (correlation, regression, hypothesis testing).
  - Xây dựng mô hình học máy (machine learning) như:
    - Hồi quy tuyến tính (Linear Regression)
    - Phân loại (Classification)
    - Gom cụm (Clustering)
  - Đánh giá độ chính xác của mô hình (accuracy, precision, recall, v.v.)
- **Kết quả:** Kết luận, mô hình hoặc insight (thông tin hữu ích) rút ra từ dữ liệu

9.1 Giới thiệu về Data Science

6 **Present Findings (Trình bày kết quả)**

- **Mục tiêu:** Truyền đạt kết quả cho người ra quyết định (không chuyên về dữ liệu).
- **Cách thể hiện**
  - Báo cáo (report, slide, dashboard)
  - Biểu đồ, trực quan hóa dữ liệu.
  - Kể câu chuyện bằng dữ liệu (data storytelling).
- **Kết quả:** Ra **quyết định kinh doanh thực tế** dựa trên dữ liệu phân tích

9.2 Công cụ phổ biến trong Data Science

NumPy là một thư viện mã nguồn mở, miễn phí cho ngôn ngữ lập trình Python, dùng để xử lý hiệu quả các mảng số đa chiều lớn. Tên của nó là viết tắt của "Numerical Python" (Python số học). NumPy cung cấp các cấu trúc dữ liệu mảng và một thư viện lớn các hàm toán học cấp cao, bao gồm đại số tuyến tính, phép biến đổi Fourier và thống kê.

Đặc điểm và chức năng chính:

- Cấu trúc dữ liệu mạnh mẽ
- Hàm toán học cấp cao
- Tốc độ xử lý cao
- Thư viện cốt lõi

9.2 Công cụ phổ biến trong Data Science

Pandas là một thư viện mã nguồn mở và miễn phí cho ngôn ngữ lập trình Python, được sử dụng chủ yếu cho việc xử lý và phân tích dữ liệu có cấu trúc. Nó cung cấp các cấu trúc dữ liệu hiệu suất cao như DataFrame (bảng dữ liệu 2 chiều) và các công cụ mạnh mẽ để thao tác, làm sạch, nhóm và phân tích dữ liệu một cách nhanh chóng và dễ dàng.

Đặc điểm và chức năng chính:

- Cấu trúc dữ liệu: Series, DataFrame
- Tốc độ xử lý cao
- Xử lý và làm sạch dữ liệu
- Thao tác linh hoạt

9.2 Công cụ phổ biến trong Data Science

Matplotlib là một thư viện lập trình trong Python được sử dụng để tạo biểu đồ và đồ thị một cách linh hoạt, giúp trực quan hóa dữ liệu. Thư viện này hỗ trợ tạo nhiều loại biểu đồ khác nhau như biểu đồ đường, cột, hình tròn, và cho phép người dùng tùy chỉnh chi tiết như tiêu đề, trục, màu sắc, cũng như lưu biểu đồ dưới dạng tệp hình ảnh

Đặc điểm và chức năng chính:

- Trực quan hóa dữ liệu
- Tùy biến biểu đồ
- Lưu trữ biểu đồ
- Linh hoạt và tích hợp

9.2 Công cụ phổ biến trong Data Science

Seaborn là một thư viện trực quan hóa dữ liệu mạnh mẽ trong Python, được xây dựng dựa trên nền tảng của Matplotlib, cho phép tạo ra các biểu đồ thống kê đẹp mắt và phức tạp chỉ với vài dòng mã

Đặc điểm và chức năng chính:

- Xây dựng trên Matplotlib
- Tích hợp với Pandas
- Trực quan hóa dữ liệu thống kê
- Linh hoạt và mạnh mẽ

9.3 Cài đặt Môi trường học Data Science

1 **Anaconda Navigator**

Download tại link: https://www.anaconda.com/download

2 **Jupyter Extention cho VS Code**

Menu View → Extentions → Search “Jupyter”

3 **Google Colab**

Link: https://colab.research.google.com

9.3 Thực hành trên công cụ